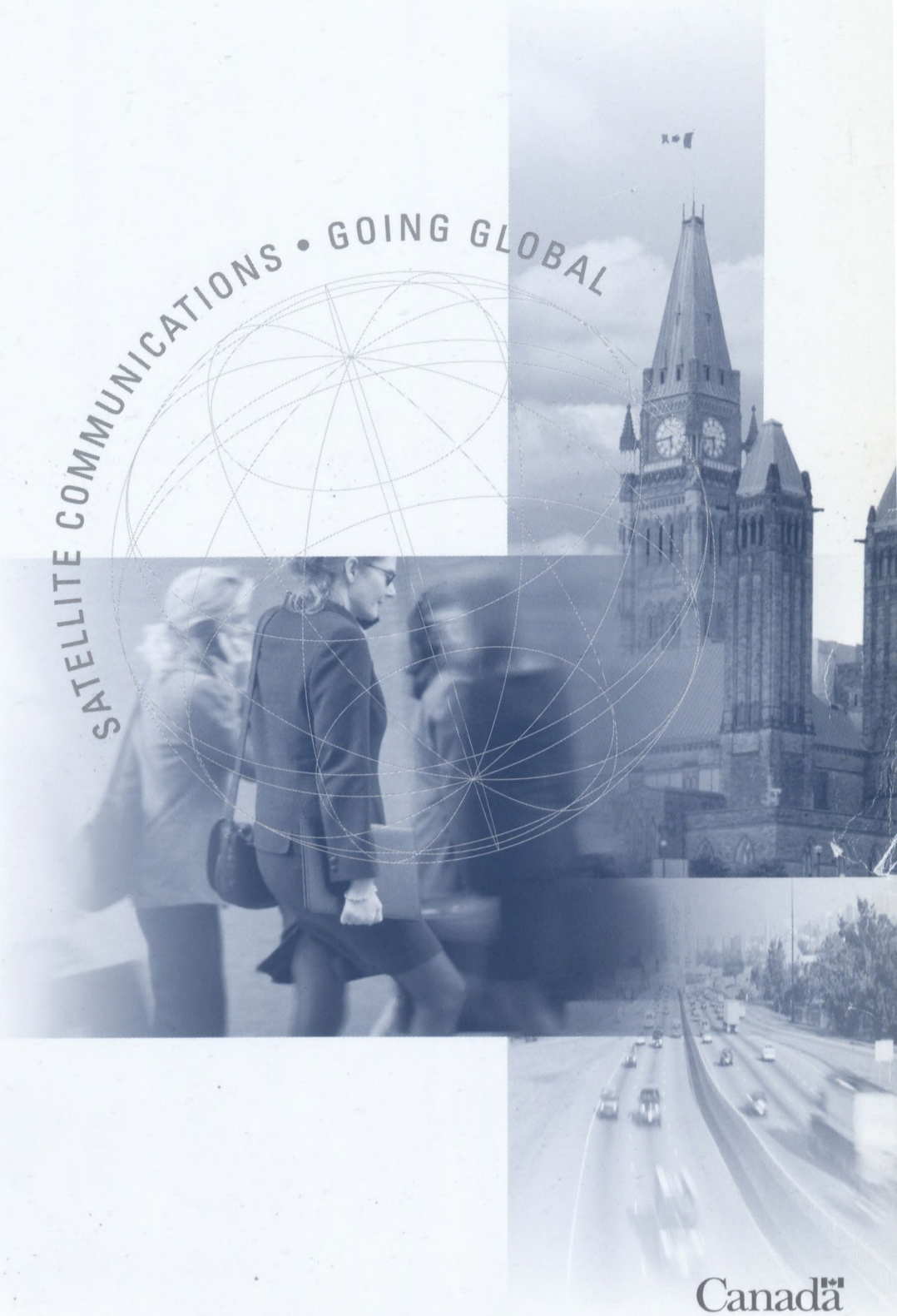
IMSC '99
INTERNATIONAL MOBILE SATELLITE CONFERENCE

June 16-18, 1999
OTTAWA, ONTARIO

Proceedings



CRC Communications Research Centre
Centre de recherches sur les communications

JPL
Jet Propulsion Laboratory
California Institute of Technology

NASA
National Aeronautics and Space Administration

CSA ASC

Industry Canada Industrie Canada

Canada

IMSC'99
INTERNATIONAL
MOBILE SATELLITE
CONFERENCE
June 16-18, 1999

Communications Research Centre
Centre de recherche sur les communications

JPL
Jet Propulsion Laboratory
California Institute of Technology

National Aeronautics and Space Administration

CSA ASC

Industry Canada / Industrie Canada

Proceedings of the sixth

International Mobile Satellite Conference 1999

Ottawa, Canada
June 16 — 18, 1999

Cataloguing Data

International Mobile Satellite Conference (6th : 1999 : Ottawa, Canada)

Proceedings of the Sixth International Mobile Satellite Conference, IMSC '99, Ottawa, Canada, June 16 – 18, 1999
Co-sponsored by the Communications Research Centre, Canada
and the Jet Propulsion Laboratory, USA.

ISBN 0-662-27869-0
Catalogue no C2-264/1999E

Additional copies of this book can be obtained, subject to availability, at no charge, as follows:

In Canada send requests to:
Communications Research Centre
VPSAT
3701 Carling Avenue
P.O. Box 11490, Station H
Ottawa, Ontario, Canada
K2H 8S2

Tel: (613) 998-2862
Fax: (613) 990-6339

In the US, send requests to:
Space Communications Program
Jet Propulsion Laboratory
Re IMSC '99
4800 Oak Grove Drive, MS 238-420
Pasadena, California, USA 91109

Tel: (818) 354-0454
Fax: (818) 354-6825

Foreword

"Going Global" is the theme of IMSC '99. Mobile SATCOM now provides very diverse applications ranging from mobile multimedia to inexpensive low-rate data-gathering systems. The design, development and analysis of these applications are explored at the Sixth IMSC. The directions in which research is heading and the attendant policy and regulatory environment also form part of the Conference program.

The IMSC '99 program contains 15 parallel technical sessions where authors will present their papers. The work of these authors, representing 14 countries and 5 continents, is gathered in these Proceedings. We would like to express our appreciation for their efforts and the sharing of their expertise. We would also like to thank the Session Organizers who helped make the sessions a reality and ensured the papers were collected for publication.

It is our belief that the papers making up these Proceedings will prove to be a valuable reference for all mobile Satcom professionals.

Jack Rigley, Co-chair,
Communications Research Centre

Tsun Yee Yan, Co-chair,
Jet Propulsion Laboratory

Table of Contents

S-UMTS and Multimedia Services I

Special Services and Applications I

Mobile Terminal Technology

Performance of Non-GEO Systems

Multiple Access, Detection and Synchronization I

Propagation

System Performance Enhancement I

Market, Standards and Regulatory Issues

S-UMTS and Multimedia Services II

Space Segment Technology

Modulation and Coding

Networking and Protocols

Special Services and Applications II

Mobile Terminal Antenna Technology

System Performance Enhancement II

Multiple Access, Detection and Synchronization II

ADDENDUM (Papers received after original Proceedings layout completed)

Market, Standards and Regulatory Issues

Space Segment Technology

Multimedia Services for Aeronautical Mobile Satellite Applications

John Broughton, Johnny Nemes
Inmarsat
99 City Road
London EC1Y 1AX, UK
E-mail: john_broughton@inmarsat.org

ABSTRACT

This paper describes how existing satellite services for aeronautical applications are evolving to allow multimedia applications to reach users on aircraft. Using new technologies that allow more efficient use of satellite power and spectrum, cost effective multimedia services are now possible for aeronautical mobile applications.

Higher bit rate services can be realised without having to increase the gain of L-band satcom antennas and without having to increase the HPA power of existing aircraft satcom systems. This can be achieved while retaining the multi-channel capability of current Inmarsat satcom systems, enabling simultaneous use of telephone and high rate data services channels. This paper also describes multimedia applications which are being considered for airlines and corporate aircraft users.

1. INTRODUCTION

Inmarsat's new high speed data (HSD) technology is being extended to serve the aeronautical community. The provision of these services to fast moving mobiles with particular types of fixed infrastructure presents both opportunities and difficulties that are different from the land mobile market. This paper describes the new services being considered and provides some examples of the applications they will support.

As discussed below, these services offer not only increased throughput but significantly lower satellite resource (EIRP and bandwidth) utilization per bit. This is as important a factor as data rate in making new applications viable.

Finally the paper describes some of the ways this new capability can be delivered to different types of users in the corporate and commercial air transport communities.

2. CONSTRAINTS IN THE AERO ENVIRONMENT

There are currently over 2000 aircraft with Inmarsat Aero-H systems installed and operating. The Aero-H system uses a high gain (12 dBic) antenna and offers multiple channels of voice, circuit data and packet data. This population of aircraft, which continues to grow at a rate of around 30 a month, is the main market for Aero-HSD services.

A key feature of the new HSD services is the ability to reuse the existing Aero-H antennas and high power amplifiers. This allows current users of Aero-H systems to benefit from minimised avionics upgrade costs and product introduction time. Safety related data and voice services which are provided by Aero-H will be operated simultaneously with HSD.

There are a number of factors that had to be considered in the development of aeronautical HSD. These include the impact of antenna phase switching on the 16QAM signal, antenna gain variation, and the nature of aeronautical fading channels which are characterized by C/M of 10 to 15 dB, fading bandwidths of 20 to 100 Hz, and differential delays of 10 to 15 microseconds. It is worth noting that the high speed data channel was designed from the outset to operate in both land and aeronautical fading environments.

While not a constraint as such, one of the considerations in this market is the impact on applications performance of equipment such as file servers and cabin PBXs which tend not part of a typical land mobile configuration.

3. HIGH SPEED DATA CHANNELS

Two different types of channels have been developed for the aeronautical market. The first is a circuit mode service based on the land mobile M4 development. The second is a packet data channel optimized for the carriage of Internet type traffic.

At least initially, these channels will be operated only in spot beams. In addition, neither of these channel types will be used for aeronautical safety applications. They do however have applications for aircraft operational and aircraft administrative services. Application of the HSD channels for safety purposes in the future is not ruled out but for the present the link budgets are not set for the 99.9% availability which safety applications normally require.

Circuit Mode Service

The circuit mode channel uses 16QAM modulation and turbocoding to deliver a user bit rate of 64 kb/s. For a more complete description of this technology to reader is referred to "Research Elements leading to the Development of Inmarsat's New Mobile Multimedia Services" by E. Trachtman [1].

Users will access the circuit HSD channel via either dial up modems or by an ISDN interface. Using the former technique, which is most likely for an airline passenger with a laptop computer, the air and ground based modems will train and connect at some rate determined by the nature and quality of the entire communications path. While it is possible that V.90 (56.6 kb/s) connectivity can be achieved, it is more likely that rates will be in the range of 33.4 to 14.4 kb/s. As Figure 1. shows, these rates enable significant improvements in file transfer time with respect to the 2.4 kb/s data rate currently available from Inmarsat or emerging LEO and MEO systems.

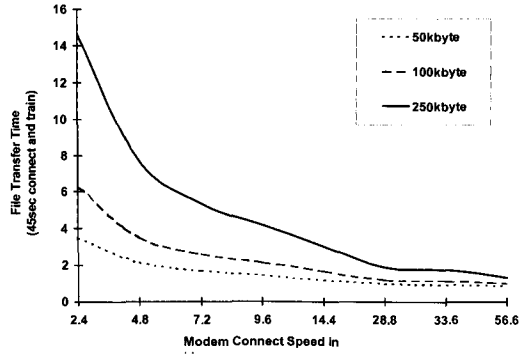


Figure 1: Transfer Times for Various File Sizes and Bit Rates.

It is obvious that the channel must offer attractive economies while being used for data rates below its full capacity. Figure 2 is a graphical representation of the efficiency of the HSD circuit channel. This Figure shows the satellite resource requirement relative to the existing 2.4 kb/s circuit mode service. As can be seen, provided the user can connect at a rate above 9.6 kb/s, the 64 kp/s HSD channel is a more cost effective option than the current service.

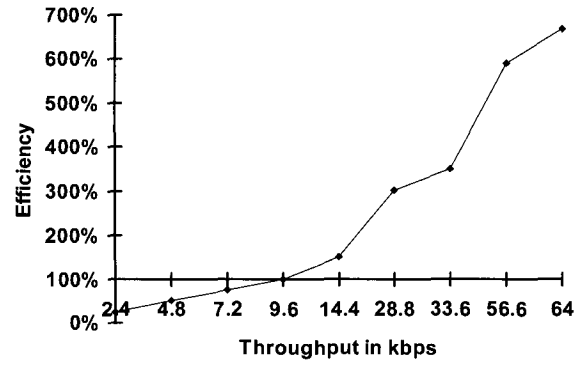


Figure 2: Relative Satellite Resource Usage

When the ISDN interface is used to access the full 64 kb/s capability of the channel, the efficiency in satellite power and bandwidth use rises to 700% that of the current QPSK ½ FEC circuit channel. The ISDN interface options are described in more detail in "An Introduction to Inmarsat's New Mobile Multimedia Service" [2]. It is expected that ISDN (either the S or U interface) will find its primary application in situations where the infrastructure is fixed such as in corporate aircraft and commercial air transport installations for crew and cockpit use.

The circuit channel will allow more effective use of existing applications such as facsimile and PC data and will enable new applications such as video conferencing using the H.320 standard. Other than perhaps video conferencing, which was simply not possible at existing rates, the other applications that will use the circuit channel have been identified for some time. They include but are not limited to:

Weather map broadcast

In Flight Entertainment audio content
Large file transfer
E-mail

Notwithstanding the economies of the HSD circuit channel, there are some applications, such as cost effective access to the Internet, that still require a highly efficient packet data channel.

Packet Mode Service

The second of the two new HSD services is packet data. The packet data channel will operate at a user bit rate of 32 kb/s and, using the same technologies as the circuit channel, achieve a per kilobit resource efficiency 260% times better than the current aeronautical packet data service. The packet channel will support both Internet traffic and connections to private networks.

Applications for the packet channel include:

Shared use by large numbers of users
Internet and Intranet access
Real time credit card validation
E-mail
Engine data
Packet based interactive tele-services

Combinations of Services

Implementation of both new HSD channel types provides the full range of capabilities. The packet channel permits multiple users in an aircraft to access HSD services on a shared basis and by charging only for data transmitted or received, makes Internet browsing and similar server applications economically viable. The circuit HSD channel can then be used when there are large file transfers to be made and for real time applications.

4. AERONAUTICAL USER GROUPS

The current installed base of Aero-H systems fall generally into two distinct groups. These two groups are distinguished by the type of applications they use, the equipment configuration which supports the applications, and by whether they are an aircraft owner and operator as is the case for most of the corporate users or simply a passenger on a commercial airliner.

Commercial Airliners

The commercial air transport or airline market consists of an itinerant user base and the seat installation can best be described as a voice or data phone booth in the sky. Access to the services will be via the passenger's laptop PC connecting to the RJ11 jack which is now standard in in-seat telephone handsets, or through the inflight entertainment system. Inmarsat's objective for this type of customer is to provide the same facility and ease of connection they would experience connecting their laptop computer in a hotel room.

Given that the laptop computer is the primary mode of access, and since the dial-up modem appears set to remain the main mode of communication for some time to come, it is expected that whether the passenger is using the circuit channel or the packet channel, he or she will be doing so using a dial up connection. It is not necessarily the case however that the connection would be to a termination point on the ground. With the inevitable progression towards installation of file servers on aircraft, it is possible that passengers could connect to a proxy server hosting an Internet point of presence (PoP) and access a wealth of locally stored content. They would then occasionally reach beyond the local PoP to the Internet for additional information, most probably using the packet data channel. Figure 3. below shows an example of this type of configuration.

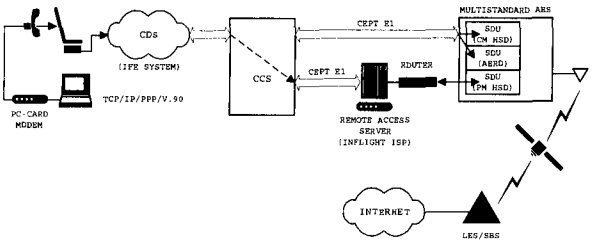


Figure 3: Configuration for commercial air transport aircraft showing on board server for local Internet point of presence.

Corporate Aircraft

Corporate aircraft, in particular the high end of the market offer more opportunity for taking advantage of fixed infrastructure such as fax machines, desktop PCs and in the future video phone and video conference facilities. The voice and data phone booth analogy is less applicable here, the oft touted "office in the sky" is a better description.

Figure 4. below shows an example of a corporate aircraft equipment configuration making use of all of the Aero-H capability including both types of HSD channel as well as regular voice services. In this type of installation an ethernet LAN could be used instead of local dial-up connections.

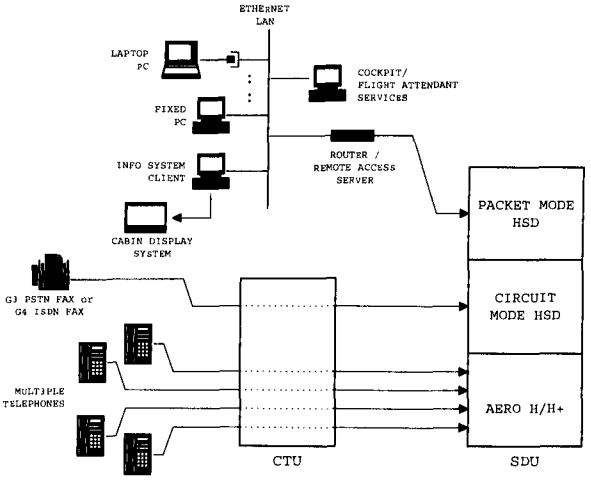


Figure 4: Example of a configuration for corporate aircraft using both circuit and packet services.

5. SUMMARY

Inmarsat is now poised to deploy its unique capability to provide mobile HSD services into the aeronautical market. British Telecom, one of the premiere providers of telecommunications services to the aviation community, has already announced it's intention to support this new capability and other major service providers are expected to follow suit. Once agreements are in place for development of avionics upgrades, plans and schedules for the service introduction will be finalized.

Since the introduction of Aero-H voice and data services to the aeronautical community in 1990, Inmarsat has recognized the need for higher speed data capability. The growth of the Internet and multimedia applications over the past several years has crystallized and solidified the market for this service and Inmarsat's development of 16QAM and turbocode technologies has now made the service technically and financially viable.

REFERENCES

[1] E. Trachtman and T. Hart, Research Elements leading to the Development of Inmarsat's New Mobile Multimedia Services; in Proceedings of the International Mobile Satellite Conference, 1999.

[2] H. Feldman and D.V. Ramana, An Introduction to Inmarsat's New Mobile Multimedia Service, in Proceedings of the International Mobile Satellite Conference, 1999.

Aeronautical Multimedia Service Demonstration at K/Ka Band

Matthias Holzbock [1], Erich Lutz [1], Michael J. Connally [2], Giacinto Losquadro [3]
[1] DLR, German Aerospace Center, Institute for Communications Technology
P.O.B. 1116, 82330 Oberpfaffenhofen, Germany
E-mail: Matthias.Holzbock@dlr.de

[2] JPL, Jet Propulsion Laboratories, Satellite Communications Group
4800 Oak Grove Drive, 91109 Pasadena California, USA

[3] ALS, Alenia Aerospazio, Space Division
Via Bona 85, 00156 Roma, Italy

ABSTRACT

Future satellite services will offer multimedia applications requiring high data rate links. In order to satisfy the expanded bandwidth allocations and avoid the restricted capacity of lower frequency bands, these systems will operate at Ka-band (20/30 GHz) and EHF-band (40/50 GHz).

An aeronautical multimedia service demonstration campaign at K/Ka Band will be presented. This demonstration was conducted within the framework of the project ABATE (ACTS Broadband Aeronautical Terminal Experiment) [1] sponsored by the European ACTS program. A group of 13 international partners developed a satellite system for mobile multimedia throughout Europe. A constellation of up to five satellites operating at K/Ka and EHF bands will provide high data rate services for all kind of mobile terminals. Several field trials involving aeronautical and landmobile channel measurement [2,3] and service demonstration campaigns at K/Ka and EHF band were performed during the project. In this paper the multimedia service demonstration will be highlighted.

MULTIMEDIA SERVICE DEMONSTRATION OVERVIEW

The goal of the multimedia demonstration campaign was to validate the aeronautical terminal prototype and the corresponding earth station equipment developed during the project. High capacity bi-directional links were used to verify the behavior of aeronautical broadband services during operations. An in-flight demonstration of a wide range of multimedia services (e.g. Internet access, video conferencing, video-broadcast, ISDN telephony) and telemedicine applications (video conferencing, EKG, blood pressure and oxygen saturation, ultra-sonic scan) was accomplished.

Fig. 1 shows the overall set-up for the demonstration campaign. The demonstration set-up consisted of an aeronautical terminal, the ITALSAT F1 satellite, and a TDS-6 ground earth station (GES) in Rome.

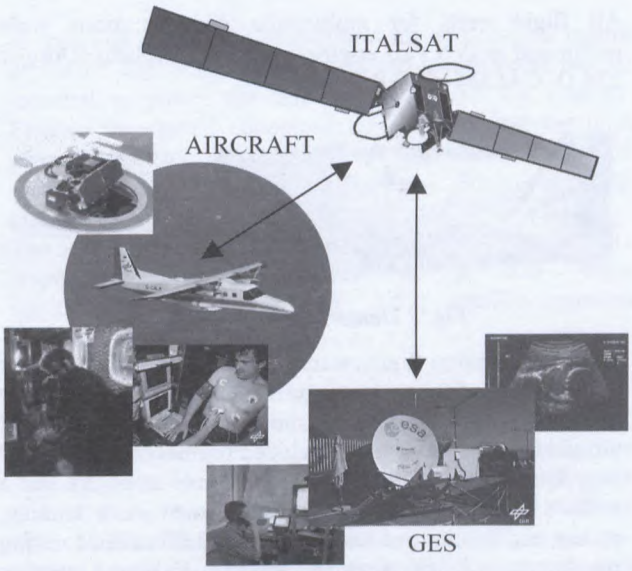


Fig. 1 Demonstration Overview

SATELLITE: ITALSAT F1

ITALSAT-F1's transparent K/Ka band transponders in channel 1 at 29.517 GHz up and 19.720 GHz downlink (mobile to fixed) and channel 2 at 29.737 GHz up and 19.940 GHz (fixed to mobile link) were used. Additionally a pilot tone was transmitted at 19.961 GHz in the forward link, to optimize the mobile antenna tracking.

GROUND EARTH STATION: TDS-6

The GES, located in Rome (elevation angle 41.48°), consisted of a trailer-mounted Cassegrain antenna with corresponding RF section and an additional container housing the IF, baseband and controlling components (Fig. 2). The antenna's diameter was 2,25 m, equal to more than 53 dBi gain at K/Ka band. The output power of the high power amplifier was about 54 dBm.

Fig. 2 Ground Earth Station TDS-6

AERONAUTICAL TESTBED AND DEMONSTRATION SCENARIOS

All flight trials for multimedia demonstrations were performed with a two engine turboprop airplane (Dornier 228 D- CALM) of DLR (Fig. 3).

Fig. 3 Demonstration Testbed

All demonstration flights were performed in a region about 80 km north of Rome at a cruising height between 12.000 ft and 14.000 ft. In order to simulate flight maneuvers of airliners, the flight patterns included normal cruise and 180 deg. U-turns of standard holding patterns arranged like a wide 8. The demonstration also included start, landing, ascent and descent, to test equipment performance during the intense vibrations on the runway. Different weather conditions allowed flights under and above clouds and during heavy rain. A comprehensive investigation of the link performance during all normal fight maneuvers as well as forced antenna shadowing by the aircraft structure was performed in a aeronautical channel measurement campaign described in [2].

AERONAUTICAL TERMINAL

The design and development of mobile aeronautical terminals for different application requirements as well as the production, implementation and testing of an aeronautical multimedia terminal prototype was one of the main initiatives of the project . Because of this the aeronautical terminal will be described in detail.
The starting point is a view of the overall demonstration set-up in Fig. 4. The aircraft's terminal baseband section consisted of a multiplexer providing a variety of data interfaces: serial and parallel data ports, Ethernet, ISDN and analog phone lines. A ultrasonic scan monitor and different cameras could be switched into the video data stream through the user application PC. A monitor for vital parameters of a test subject was supplied as well as a ISDN phone. Two identical modems were used in the aircraft and the GES to feed the IF interface. The GES was laid out with a RX/TX and signal mixing unit, but not as a service provider station. For this reason also a multiplexer was necessary to support the required data stream. Optionally two inverse multiplexers (bonding three ISDN lines with 128 kbps) made the digital data also outside the GES available, but taking the drawback of a reduced data rate. Fig. 4 shows the set-up during a telemedicine demonstration to a medical expert team at the university clinic of Tübingen.

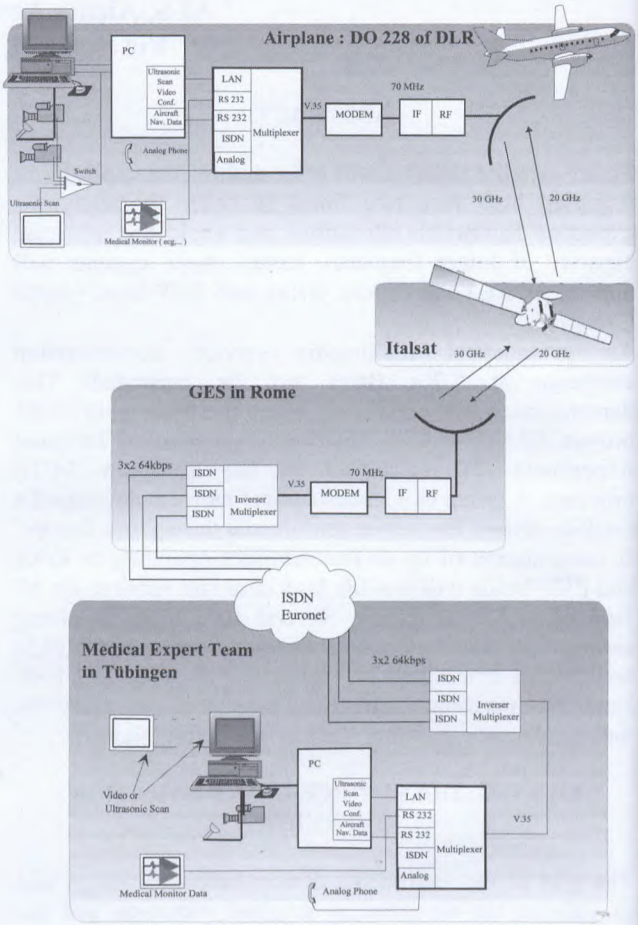


Fig. 4 Telemedice Demonstration Set-Up

Fig. 5 shows a detailed diagram of the aeronautical terminal components and their interaction. In addition to the essential equipment necessary for the demonstration such as the antenna and the user application PC, measurement, monitoring and recording modules were implemented into the aircraft. A more detailed description of the main components of the terminal is given below.

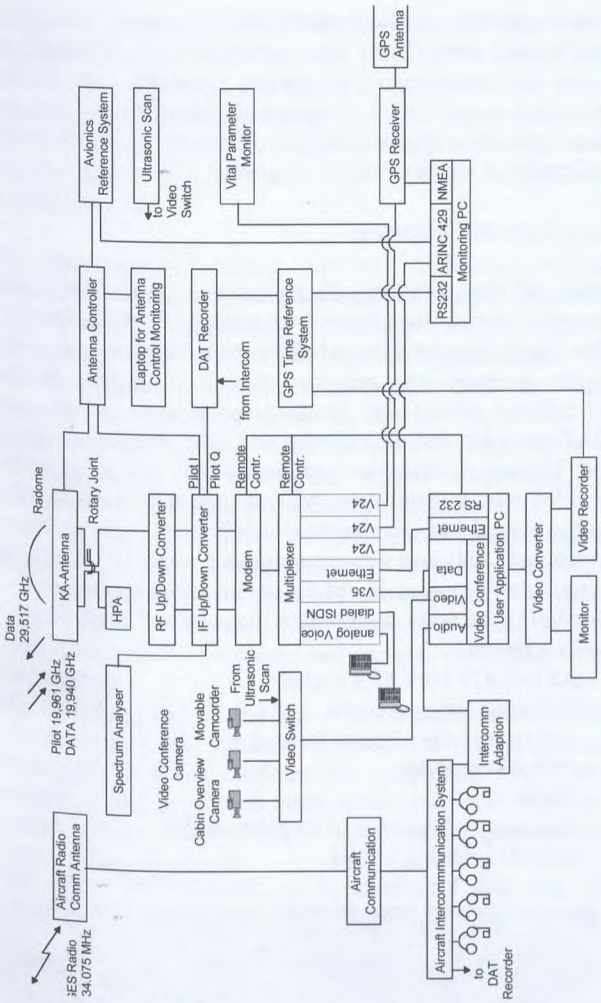


Fig. 5 Aeronautical terminal hardware structure

Mobile Antenna, antenna tracking, RF/IF section

A mechanically steered slotted array antenna (Fig. 6) bulit by EMS Technologies and provided by the Jet Propulsion Laboratory (JPL) was used for the aeronautical terminal. Mounted on top of the aircraft between the wings and the tail structure, the antenna consisted of separate arrays for receive and transmit. The sensitivity of the 20 GHz receive array was about 2 dBic/K, the gain of the 30 GHz transmit array was about 30 dBi. It's beamwidth was about 5 deg
The arrays are mounted on a single platform which is fitted to a gimbal in order to provide full steering range. The maximum tracking rate of the system was 30°/sec in elevation and 60°/sec in azimuth.
The antenna tracked the satellite during all flight and on-ground maneuvers with a combined open and closed loop algorithm. The aircraft avionics data (three axis laser gyroscope, compass and position information) provided by a ARINC 429 bus was used to compensate for fast attitude changes. A transmitted pilot tone at 19.961 GHz was used to correct long term drifts of the sensors.
The overall dimension of the antenna was approximately 15 cm (6 inch) in height and 60 cm (25 inch) width. It was covered by a radome.

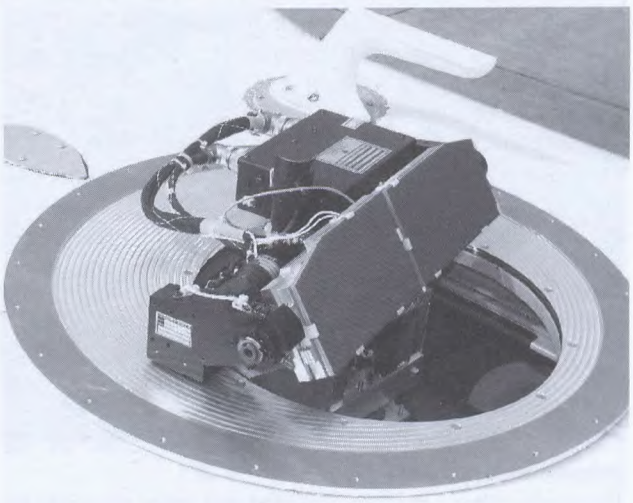


Fig. 6 Aircraft mounted mobile antenna

A traveling wave tube amplifier with a measured output power of 126 Watts was used in the aeronautical mobile terminal to power the transmit section of the antenna. Fexible waveguide connected the antenna and the high power amplifier. The low noise amplifier was fitted directly at the back of the receive array. For all other RF and IF connections coax cables were used.
The uplink and downlink frequencies of 30 GHz, 20 GHz respectively were converted in a two stage up/downconverter to the 70 MHz satellite modem output/input.

Modem and synchronization issues

The modem used BPSK modulation and a Viterbi ½ rate forward error correction code. The automatic frequency control was able to counteract Doppler frequency shifts of up to 30 kHz. This was sufficient during all flight and on-ground maneuvers. Also a Doppler buffer in the modem was implemented to compensate bit delays due to the aircraft's velocity by writing data into this buffer and reading the data out with the transmit clock. A Doppler buffer size of 8192 Bit was used. The frequency and bit synchronization delay of the modem was less than 1 second.
The antenna, RF/IF converters and the modem were capable of handling data rates up to 2Mbps, although lower bit rates were used due to limitations in other equipment.

Multiplexer and synchronization issues

As already mentioned, a multiplexer was used to make a wide range of data interfaces available. The maximum data rate of the mutliplexer was 768 kbps and restricted the demonstration data rate to this value. Modem and multiplexer synchronization over the link was realized by synchronizing all devices to one master and recovering this master-clock in the aircraft terminal, using an other Doppler buffer in the multiplexer to compensate bit and

packet losses. This multiplexer provided a two channel analog voice module, a ISDN module, a two channel high speed data module, two two channel low speed data modules and a 10BaseT Ethernet bridge.
The ISDN module had one ISDN basic rate interface which enables direct connection to an ISDN public network for extending voice and fax access to the aircraft. The aircraft ISDN interface was operated in echo cancellation mode and used a phantom feed for the power supply of the terminal equipment. A commercial ISDN telephone was used in the aircraft to provide and demonstrate dialed ISDN telephony.
The Ethernet module operated at the physical and data link layers of the OSI model and was completely transparent to higher level protocols, such as the demonstrated TCP/IP. The data rate was adapted to the traffic load of the demonstrated scenario.
The analog phone module provided two voice channels for direct connection of a telephone or other equipment. ADPCM digitizing techniques were utilized at a transmission data rate of 16 kBps. An adaptive echo cancellation for handling of near-end echoes was also used. In the aircraft, a commercial analog phone was connected to this interface.
The low speed data module provided synchronous or asynchronous ports at selectable data rates from 300 bps to 384 kbps. These modules were used to interface with the GPS receiver, the avionics ARINC 429 bus, and the vital parameter monitor.
The high speed data module was used to interconnect the user application terminal's videoconferencing system to the muliplexer.
Both modem and multiplexer were remote controlled from a terminal program on the user terminal PC.

User Equipment

The user terminal consisted of an industrial PC with video and audio equipment for the videoconferencing application. The user terminal managed the videoconference system and others applications as files transfer, application sharing, internet access and others. A detailed list of the demonstrated applications is given later on. A F-BAS signal switch was used to select four different video sources for the video input of the video conferencing system. A commercially available Camcorder was used as a movable video source inside the aircraft. The video camera belonging to the videoconferencing system was fixed mounted at the front of the cabin and used as cabin overview camera. A high quality CCD camera was fixed mounted in front of the operators seat and was also used during videoconferencing.

Medical Equipment

For the telemedicine demonstration a ultrasonic scan monitor was implemented and the monitor output signal could also be switched to the videoconferencing system. A vital parameter monitor provided features such as EKG, blood pressure measurement, blood oxygen saturation display and more. These vital parameter were displayed on board and transmitted for viewing simultaneously on the ground. A resuscitation dummy was used for supervising a re-animation performed by the operators in the aircraft assisted by a medical expert on ground.

Measurement Equipment

Most of the data regarding the link availability was recorded during the channel measurement campaign [2]. But also during the demonstration trials the received power in the gateway was recorded during all flights. In the aircraft the pilot I and Q components were stored on a DAT recorder and all avionics and GPS data were noted, too. The screen of the user application PC was recorded on a video tape and all operator conversation including the audio signal of the videoconferencing system were monitored. All data was synchronized via IRIG signal to allow an post-processing of the data in the laboratory.
The following table summarizes the recorded data:

- link power
- ARINC429 (aircraft's attitude)
- GPS (aircraft's position)
- user terminal's monitor (video)
- channel error rate
- Eb/No
- frequency offset (due to Doppler shifts)
- antenna pointing angels

All data was synchronized via IRIG-B GPS based time reference.

DEMONSTRATION APPLICATIONS AND RESULTS

Many test flights were performed, including approval, test and calibration flights for the equipment components (i.e., such as radome and equipment racks approval, antenna pointing and avionics system tests). Highlights were three demonstration flights performed in September 1998 for a interested group of selected experts at Alenia premises and on a second day for telemedicine conference at University clinic in Tübingen. A second demonstration campaign was flown during the 4th Ka Band Utilization Conference in Venice at the beginning of November 1998. Both trials were performed with a precise date and time schedule and to a wide public. This indicates that the systems components operated very reliably and had reached a high standard of operability.

During the first demonstration for a interested group of selected experts at the Alenia premises the following in-flight office and in-flight entertainment applications were demonstrated:

in-flight office and in-flight entertainment: videoconferencing at 384 kbps, ISDN telephone and fax, analog telephone, shared applications, Internet access, e-mail, data transfer, and on board TV.

During all demonstrations the GPS and avionics attitude data of the aircraft were transmitted for display on the ground.

A main link data speed of 768 kBps and 512 kBps was chosen during this demonstration. The Eb/No was above 7dB for the 768 kBps and above 9 dB for the 512 kBps link rate in the aircraft during all flight phase including take off and landing. The bit and frame error rate with Viterbi coders was less than 10^{-9} and not measurable during the trials.

The demonstration during the 9th Conference of the 'Deutschen Gesellschaft für Katastrofenmedizin' (German catastrophe medicine society) in Tübingen was focused on the transmission of medical support applications. IN addition to the ultrasonic scan monitor and the vital parameter monitor, a resuscitation dummy was part of the demonstration. A person with no prior medical training was supervised by a medical expert on-ground via videoconference during a heard attack scenario.

- Telemedicine: a video conference to a medical expert team on ground was established. Special microscope cameras and ultrasonic scan were switched into the video data stream. In parallel vital parameters of the patient like EKG, blood pressure, and oxygen saturation were transmitted.

The following Fig. 7 and Fig. 8 show the screens of the user application PC monitor on-board and the on-ground display of the video conferencing system. In Fig. 7, the window in the upper left edge shows the local on-board camera. In the other window the medial expert on the remote side is displayed.

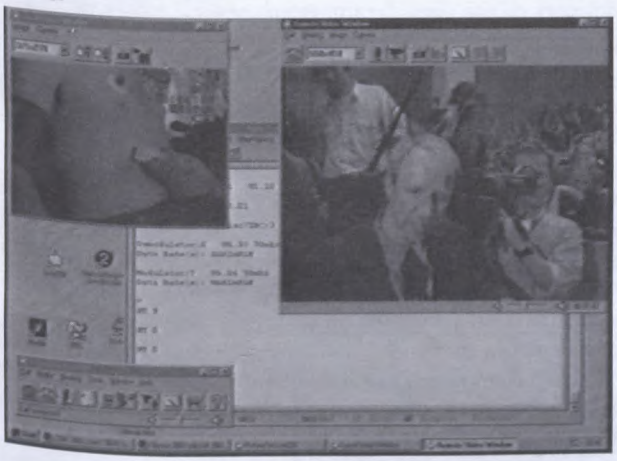


Fig. 7 On-board display

In Fig. 8 the upper left window shows the monitor output of the remote ultrasonic scan.

Because of using the inverse multiplexers the link rata rate was reduced to 386 kBps during this demonstration.

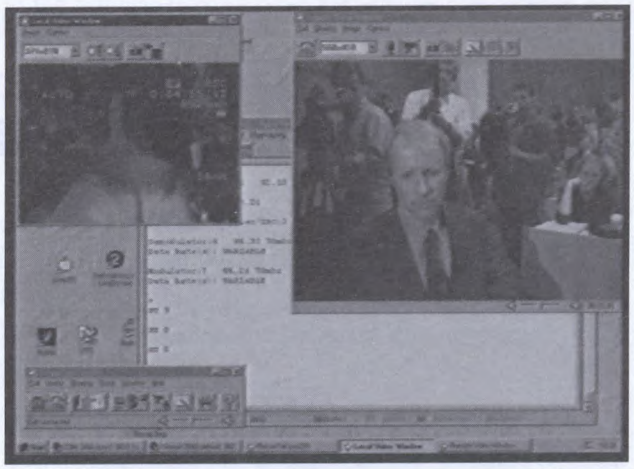


Fig. 8 On-ground display

The topics of the demonstration trials flown during the 4th Ka Band Utilization Conference in Venice at the beginning of November 1998 were similar to the ones described before.

As an example of the recorded data some link quality measurements are shown, performed during the aircraft ferry to Italy. The following table summarizes a Eb/No measurement while the aircraft was flying form the edge of the spotbeam coverage (Munich) of ITALSAT into the center (Rome) at a data rate of 512 kbps.

UTC	Latitude	Longitude	Eb/No
	Munich		6.4
9:48:20	45.53	11.60	9.0
9:52:00	45.33	11.60	9.3
9:57:50	45.01	11.61	9.5
10:02:39	44.76	11.53	9.6
10:08:05	44.51	11.43	9.6
10:18:10	43.99	11.61	9.3
10:23:15	43.69	11.70	9.6
10:48:20	Rome		9.7

REFERENCES

[1] G. Losquadro, M. Luglio F. Vatalaro: A geostationary satellite system for mobile multimedia applications using portable, aeronautical and mobile terminals; in Proceedings 5th International Mobile Satellite Conference (IMSC 97), pp. 427--432, 1997.

[2] M. Holzbock, A. Jahn, E. Lutz: Aeronautical Channel Measurement Trials at K Band; in Proceedings International Mobile Satellite Conference (IMSC 99), 1999.

[3] A. Jahn, M. Holzbock: EHF-Band Channel Characterisation for Mobile Multimedia Satellite Services; in Proceedings 48th IEEE Vehicular Technology Conference (VTC'98), pp. 209-212, 1998.

H.263 Codec with Improved Synchronization for Video Transmission over a Land Mobile Satellite System with Turbo Coding Operating in the Ka and L Bands

Durhan Guerrrero, Dimitrios Makrakis
Advanced Communications Engineering Center
Dept. of Electrical and Computer Engineering
The University of Western Ontario
London, Ont., CANADA, N6A 5B9
E-mail: dguerrer@julian.uwo.ca, dimitris@engga.uwo.ca

Abstract - This paper investigates the performance of a modified H.263 video encoder/decoder (Codec) over land mobile satellite communication systems that use Turbo Codes, and operate over the L or Ka frequency bands. The receiver makes use of a channel estimator; its output is used by the turbo decoder. Two different channel estimators are tested; one uses an IIR Filter, and the other a Kalman Filter. The Soft Output Viterbi Algorithm is used as the base-decoding algorithm, along with channel interleaving. Different types of channel fading are investigated, and their effect on the performance of the considered system is analyzed. The performance addresses the issues of delay and bit error rate, along with video quality.

INTRODUCTION

Today, satellite communications is growing rapidly as new LEO and MEO satellites networks are being lunched to provide PCS services anywhere in the world [1]. Multimedia services is expected to become available in the future, however, in its present state, satellites are not able to provide these services due to technology limitation and spectrum shortage. Existing systems serve (primarily) voice and low speed data. To address the problem of bandwidth scarceness, it is planed that satellites services are offered in the Ka band (18 GHz to 30 GHz). However, in order to provide reliable video based applications through the mobile satellite service, it is important that the technology id developed to provide very low Bit Error Rates (BER) through channels severely impaired by shadowing and multipath fading [2]. Also, the implications associated with the operation of the Video Encoder/Decoder (Codec) using mobile satellite systems has to be addressed, since the existing video technology is based on the use of highly reliable communication systems and networks.

Real-time video services capable of delivering low bit rates video over narrow bandwidth networks is possible today due to advances in video coding technology. A popular low bit rate Codec is the H.263, proposed by the ITU-T, which targets the transmission of video streams over the public switched telephone network (PSTN), at rates less than 64 kbits/s [3]. The H.263 in its present form is not able to deal with the impairments that are present in mobile satellite channels. The H.263 Codec must be adapted to be able to operate in the unreliable mobile satellite channel. One way to meet this objective, is by using suitable formats of forward error correction (FEC) coding in order to meet the Quality of Service (QoS) requirements.

Turbo Codes (TC) were first introduced in 1993 by C. Berrou, A. Glavieux, and Thitimasjshima [4]. They provide excellent error correction capabilities and are known as one of the most powerful FEC schemes available today [5]. The considerable improvements provided by the TC in fading channels generate the potential for the development of new products. However, the heavy processing requirements of turbo decoders and their potential for introducing relatively long delays in the delivery of the data stream, opens several issues that have to be investigated, especially when their use in delay and delay jitter applications is considered. However, to the best of our knowledge, there have been no thorough studies on the performance of Turbo Codes in Land Mobile Satellite Channels (LMSC) using BPSK mapping and how well they protect a video stream under these conditions. This paper investigates the performance of a modified H.263 Codec with Turbo Codes in land mobile satellite communication systems operating in the L and Ka bands using different estimation methods, and the impact of impairments in the original and modified H.263 Codec. The two channel estimation methods proposed here are the Kalman and IIR filters and are used along with the Soft Output Viterbi Algorithm (SOVA) [6]. The implications of using these filters in the fading estimation process are investigated. Also, the performance of the modified H.263 is assessed and is compared to the performance of the original H.263 Codec.

The remainder of this paper is organized as follows. In the next section, the model of the Land Mobile Satellite Channel (LMSC) is described, along with the parameters used in our performance analysis. Section 2 describes the modifications performed on the H.263 Codec. Section 3 provides a description of the communication system with the Turbo Coding scheme we used in this work. In section 4, the performance evaluation results regarding Bit Error Rate (BER), latency, and decoded video frames from both the original and modified H.263 Codec are presented and analyzed. Finally, section 5 concludes this work.

1. CHANNEL MODEL

The channel model used in this paper is the one proposed by C. Loo [7], [8] a well accepted statistical model for the flat-fading LMSC. This model cover both, L (1.5 GHz) and Ka (20 GHz) bands. According to the model, the statistical behavior of the signal envelope follows a Ricean distribution, with its local mean, the line-of-site (LOS) component, following a log-normal statistical distribution. The amount of signal distortion happens to be worst in Canada due to the high northern altitude, which forces the signals transmitted from geostationary orbit satellites to reach the receivers with small elevation angles (usually between 15-20 degrees).

The probability distribution function (PDF) of the received fading signal envelope can be mathematically represented as [9]

$$p(r) = \frac{r}{b_o\sqrt{2\pi d_o}} \int_0^\infty \frac{1}{z} \cdot \exp\left(- \frac{(\ln(z) - \mu_o)^2}{2d_o} - \frac{r^2 + z^2}{2b_o} \right) I_o\left(\frac{rz}{b_o}\right) dz \quad (1)$$

where μ^o is the mean value due to shadowing, d^o is the variance due to shadowing, and b^o is the average scattered power due to multipath. $I^o(\bullet)$ is the modified Bessel function of zero order.

The parameters for the L and Ka band LMSC [7][8] that have to be placed in the above equation are summarized in the Table 1, shown below, and are used in this paper to simulate the mobile satellite channel. Parameters are given for the following cases: light, average, and heavy shadowing.

Table 1: Model Parameters of a Land Mobile Satellite Channel at L and Ka bands

channel	freq.(GHz)	b_o	μ_o	$\sqrt{d_o}$
Light Shadowing	1.5	0.158	0.115	0.115
	20	0.1585	-0.230	0.0115
Average Shadowing	1.5	0.126	-0.115	0.161
	20	0.0398	-1.95	0.46
Heavy Shadowing	1.5	0.0631	-3.91	0.806
	20	0.10	-2.30	0.046

H.263 CODEC

The original H.263 Codec in its present form cannot deal with the wireless environment due to slow macroblock updating and its inability to synchronize at the macroblock layer when an error occurs [10]. Therefore, modifications to the original H.263 Codec are required, in order to deal with this environment.

2.1 Macroblock Synchronization

The original H.263 Codec does not support macroblock synchronization in the event that a Group of Block (GOB) or Macroblock header is modified in the external environment. It only supports synchronization at the GOB layer [10]. This means that a single error causes the loss of a GOB slice. Figure 2 shows clearly how serious this problem can be, when four errors are introduced in the video stream.

Figure 1. Synchronization of H.263 at the GOB Layer

The answer to this problem is to introduce two new headers at the macroblock layer to provide the decoder with synchronization at this layer. The first header is called the Synchronization Macroblock (SMB) Header, which is similar to the Group of Block Synchronization header (GOBSC) header used in the GOB. This allows the decoder to synchronize at the Macroblock level quickly, without increasing complexity at the decoder.
The second header following the synchronization header is the Macroblock Number (MN). The MN tells the decoder where the macroblock is located in the frame, allowing the decoder to set parameters needed for the decoding of each Macroblock. This MN can also be useful for future options, where the position of the macroblock might be needed. The size of the MN header can be varied from 6 to 11 bits, and make better use of the available bandwidth, since the MN header will use only the necessary bits needed for the specified picture format. The reason for the bit range (6 to 11) is that the 6 bits will cover the smallest picture format (sub-QCIF), and the 11 bits will cover the largest picture format (16-CIF) [10].

2.2 I-pictures

The original H.263 codec was designed to update the Intra macroblock every 132 frames. The problem with this is that in the event that multiple errors occur in the I-frame, the H.263 Codec is not able to update the Intra blocks fast enough, leading to poor video quality.
The standard was designed on the assumption that the H.263 Codec will be used in a wired environment, where

the reliability of the channel is high (the probability of errors is low). In a wireless environment, faster updating of the Intra blocks is needed in order to deliver an acceptable video stream. In order to improve the video stream, the option of inserting I-pictures every x number of P-pictures has to be made available at the decoder. This modification improves the quality of the video stream in a wireless environment, making it possible to deliver the necessary video quality needed for video conferencing. The encoder was modified to add I-frames every x P-frames, where x is the distance of P-frames between I-frames.

3. COMMUNICATION SYSTEM MODEL

The models of the transmitter and receiver systems are shown in Figure 5 and 6 respectively. In our work, we have investigated the performance implications of using a Chebichev IIR and a Discrete Kalman Filter instead of a FIR filter. The reason for pursuing this study was inspired from our belief that use of an IIR filter or a Kalman Filter could provide us with improvements over the FIR filter, when estimating the fading. The IIR filter is better than the FIR filter due to better amplitude response IIR filters have when compared to the response of FIR filters of same order and similar complexity level [11]. The Kalman Filter on the other hand performs better than both the IIR and FIR filters, due to its estimation of past, present and future states, and it can do so when even the precise nature of the channel is unknown [12]. This is advantageous, as the Kalman Filter does not need to know the fading (B_dT product) of the channel, while this must be provided for both the FIR and IIR filters in advance. The results proved our expectations correct. In our simulations, when the IIR or FIR filters are used to estimate the fading, the filter is set equal to the Doppler bandwidth [13] of the process. With the Kalman filter, one sets its parameters once, and there is no need to adjust them for different Doppler bandwidths. The noise variance turns out not to be critical as it was found in [14], and can be set to a constant or use the technique described in [14]. In our study, we set the noise variance to a constant value for the IIR and FIR filters, and used the technique described in [14] for the Kalman filter in order to obtain better results.

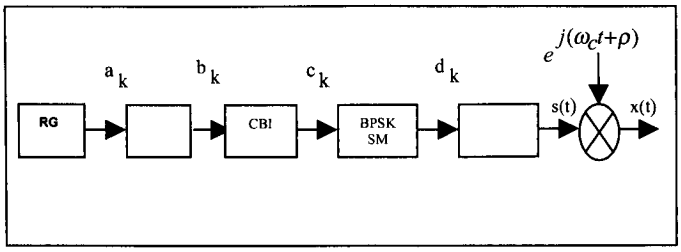


Figure 2. Block diagram of the Tranmitter

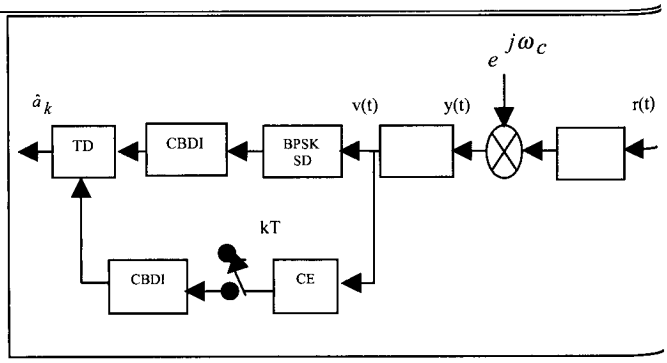


Figure 3. Block diagram of the receiver

4. PERFORMANCE EVALUATION RESULTS

The performance evaluations are based on computer simulations and the use of a software based H.263 Codec. As mentioned earlier, performance evaluations were carried out for both, L and Ka bands, with the modified and non-modified H.263 video stream. The communication system is based on coherent BPSK (we have expanded our work for the cases of QPSK as well as differential and multiple differential detection, however, space limitations do not allow us to include these results in the present paper). The roll-off factor of the pre-modulation and post-detection filters used in our simulator has been set to 0.5. Puncturing was used with our codes, in order to produce ½ rate, and the constraint length K for the RSE is equal to 3. The estimates of the fading signal envelope is obtained from the output of the fading estimation filter (see Figure 3). The Kalman filter that used as fading estimator is a one step predictor with a window size of 30. A 6-tap Chebishev IIR filter is also used for comparison purposes and has a 3 dB cut-off bandwidth set equal to the B_dT product.

A set of evaluations was conducted for a BPSK system, with the objective to assess the accuracy of our simulator. Figures 4 (L band) and 11 (Ka band) display theoretical and simulation results of BER versus Eb/No, for light, average and heavy shadowing. As can be seen, there is a very good agreement between theoretical curves [15] and simulations results, which allows us to conclude that our simulation environment is highly reliable.

The BER performance results for the L band are presented in Figures 5 to 10 and correspond to light and heavy shadowing using a Kalman filter. Each case (light, heavy) has been evaluated for $B_d T = 0.005$ and 0.05. Also, the CBI depth has been set to 1024 (32x32 matrix) in all cases. In each figure, we are providing BER curves of the turbo coded system, when the decoder uses 1, 3, and 5 iterations. We have included in each figure the BER curve of the uncoded system as well, for comparison purposes. From the results, it is evident that use of turbo codes offers improvements in excess of 22 dB under light shadowing conditions, 25 dB under heavy shadowing (these numbers correspond to a BER value of 10^{-4} and are produced by comparing the curve corresponding to the uncoded system and the one corresponding to the turbo coded system that uses only one iteration at the receiver). Observing these figures, we realize that use of more iterations at the

decoder results always to some improvement, however, three iterations seem to be enough to get almost all the possible additional gain. The gain achieved by using a higher number of iterations seems to be in the range of 2 to 3 dB. Comparing the curves obtained for B_d T = 0.005, with those obtained for B_d T = 0.05, we realize that they seem to be quite close with each other. The reason for this is the following. Higher values of B_d T tend to "shorten" the duration of the fading periods. However, since the coded system already introduces enough randomization through the inherent interleaving properties of the turbo codes and of the block interleaver, the additional "randomization" introduced by the higher value of B_d T, does not have any noticeable effect on the performance. Figures 9 and 10 show how well Kalman Filter performs when compare to the IIR and FIR filters, and to the maximum performance that can be achieved by Turbo Codes with exact fading knowledge. The Kalman filter performance is better than that of the IIR and FIR filters, due to its accurate fading estimation at higher Eb/No, for both light and heavy shadowing.

Similar evaluations with those reported above for the L band were performed for the Ka band as well. The BER vs Eb/No curves are displayed in Figures 12 to 17, and correspond to light and heavy shadowing, with B_d T equal to 0.005 and 0.05. The CBI depth is again equal to 1024. The obtained results demonstrate that use of turbo codes provides a gain of at least 25 dB for light shadowing and 28 dB for heavy shadowing (these gains refer to a BER value of 10^{-4}). This allows us to conclude that use of Turbo Codes is more beneficial in systems operating at Ka band. The performance of the coded system under B_d T = 0.005 and B_d T = 0.05 is very similar. The same explanation with the one given above for the L band applies here as well. The improvement offered by the use of a higher number of iterations at the decoder is in the order to 3 to 4 dB. Again, most of the gain comes by using three iterations. Additional iterations increase the complexity and processing delay, without offering any substantial improvement. Figures 16 and 17 show again the Kalman Filter performance is better than that of the IIR and FIR filters for both light and heavy shadowing.

The purpose of Figure 18 is to help us understand what is the impact of CBI on the performance of the coded system. We assume operation in a L band channel experiencing average shadowing, having B_d T = 0.05. The displayed curves correspond to a turbo coded system whose decoder performs 5 iterations and the system: i) does not have channel block interleaver; ii) it has a channel block interleaver of depth 512 bits (corresponding to a interleaving matrix of 16x32); iii) it has a channel block interleaver of depth 1024 bits (corresponding to an interleaving matrix of 32x32); iv) it has a channel block interleaver of depth 2064 bits (corresponding to an interleaving matrix of 32x64). From the curves, we can conclude that use of channel block interleaving can offer an additional improvement in the range of 1.5 to 2 dB (at BER of 10^{-4}). Also, most of the improvement to be gained by the use of block interleaving can be achieved with a CBI depth equal to 512. Use of interleavers with larger interleaving depths increases the delays associates with the decoding of information, without any substantial performance improvement. Consequently for delay sensitive applications such as video, it is recommended that no block interleaver, or a short block interleaver, is used.

Figure 19 to 20 provides us with the processing complexity of the decoder, which increases with; i) an increase in the number of iterations performed by the decoder; ii) an increase in the size of the CBI depth. Figure 19 displays the floating point operations that have to be performed by the decoder with an IIR filter versus the number of iterations (performed by the decoder), and Figure 20 displays the results when a Kalman filter is used in place of the IIR filter. The following cases are investigated: i) when the CBI depth is equal to 512, ii) when the CBI depth is equal to 1024, and iii) when the CBI depth is equal to 2064. As we can see, the additional processing load contributed by an increase in the size of the CBI depth is small. At the same time, the processing load increases linearly with the number of iterations performed at the decoder. The Kalman filter needs twice the processing power at one iteration when compared to the processing power needed for an IIR filter. This indicates that judgement must be used when using Kalman filters, as additional processing is linked to higher hardware/software cost and processing delays. Attention has to be paid to these results when dealing with latency sensitive applications (e.g. video, interactive applications etc.).

Performance evaluations were carried out for the Modified and non-modified H.263 Codec stream consisting of 6 QCIF frames in a L band LMSC with B_d T = 0.05 and average shadowing. The Codec inserted a single I-Picture in the 5th frame, thus allowing us to see the effect the new I-Picture. The CBI is varied along with the power for the Turbo Codes. The CE uses a Kalman filter to estimate the fading envelope. The results for the simulations with different parameters are given in Table 2. The Figures 21 to 22 represent six frames without any errors, and are given here as reference.

The Figures 23 to 24 show the first and sixth frames encoded and decoded with the original H.263 Codec. The simulation parameters are set to an Eb/No value equal to 0.5 dB, CEI and TEI block size of 1024 with 3 iterations. The results show that a BER of 6.25 x 10^{-4} for the regular H.263 Codec produces a corrupted video stream that is useless in teleconferencing.

Figures 25 to 26 represents the first and sixth frames encoded and decoded by the modified H.263 Codec. Here, the parameters are set again to Eb/No of 0.5 dB, CEI and TEI block size of 1024 with 3 iterations. The BER is 6.45 x 10^{-4}, which is similar to the BER obtained for the unmodified H.263 Codec with the same transceiver and channel parameters. The decoded video stream quality has improved vastly over the video stream decoded by the non-modified H.263 decoder video stream (see Figures 23 to 26). The modified decoder adds 5 to 10 % overheads produced by the SMB and MN headers, which are added to both, the I-frame and P-frame. This is not a significant

price to pay when the results show a large gain in picture quality. The gains in video quality obtained by using an additional I-frame are easily observed. However, this increases the amount of data that are produced by the encoder. The result is that higher processing is required by the turbo encoder and decoder to process these data, which results to higher processing delays (they associated with the operation of the turbo decoder). The increase in processing load is significant as shown in Table 2 (see the column indicating the number of floating point operations). This suggests that the use of I-frames should be avoided if possible in short video sessions, or have I-frames be inserted with large number of P-frames in-between.

2. CONCLUSIONS

We have conducted a thorough investigation of turbo coded systems operating in L or Ka band mobile satellite channels using a Kalman filter as channel estimator, and we provided suggestions on how they can be used in video over mobile satellite systems. In addition, we have introduced modifications in the existing H.263 Codec structure, making it suitable for the unreliable mobile satellite channel. Our work has addressed a number of system design issues, such as complexity versus gain, increase in processing load, processing delay versus reduction of BER, and video quality. The content of our research work can assist the design engineer to make the right decisions in terms of performance versus complexity, especially when dealing with design of systems servicing delay and delay jitter sensitive applications such as video or interactive multimedia applications.

Table 2: Simulation results for the video streams in LMSC using Turbo Codes

Eb/No (dB)	CBI & TEI Block Size	Iterations	BER	Floating Points Operations (Original H.263)	Floating Points Operations (Modified H.263)
0.5	512	3	8.66E-04	5.25E+08	1.00E+09
1.5	512	3	3.13E-04	5.25E+08	1.00E+09
0.5	2048	3	3.98E-04	5.69E+08	1.04E+09
0.5	1024	3	6.45E-04	5.44E+08	1.02E+09
0.5	no interleaver (TEI = 1024)	3	0.0041	4.79E+08	9.79E+08

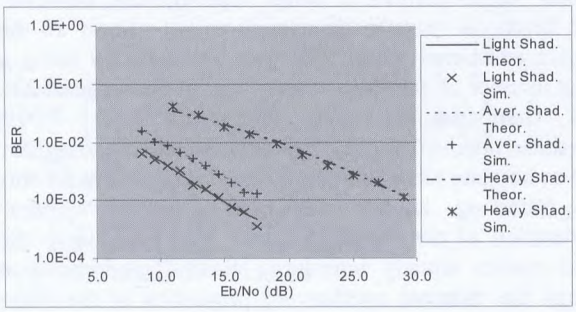


Figure 4. LMSC Bit Error Rate Comparison of Theoretical vs. Simulation Results for Light, Average, and Heavy Shadowing at the L-band.

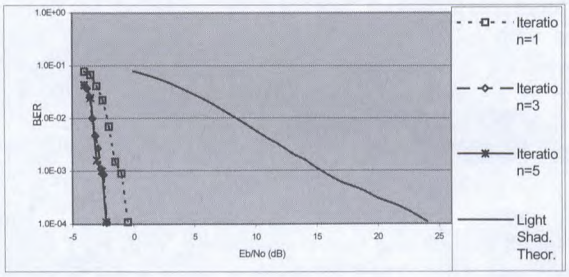


Figure 5. Turbo Codes Results for Light Shadowing (L-Band) with Block Size = 1032 bits and B_dT =0.005.

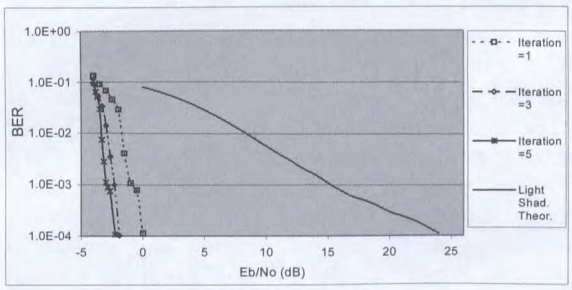


Figure 6. Turbo Codes Results for Light Shadowing (L-Band) with Block Size = 1032 bits and B_dT =0.05

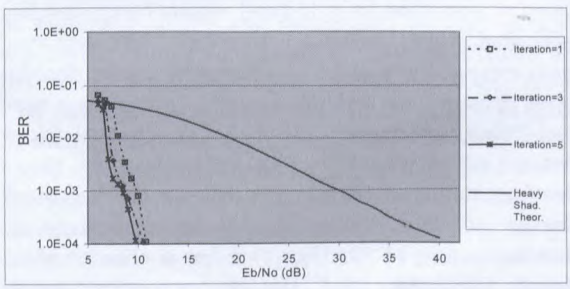


Figure 7. Turbo Codes Results for Heavy Shadowing (L-Band) with Block Size = 1032 bits and B_dT =0.005.

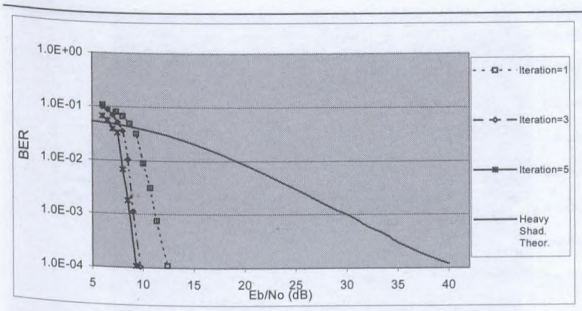


Figure 8. Turbo Codes Results for Heavy Shadowing (L-Band) with Block Size = 1032 bits and B_dT =0.05.

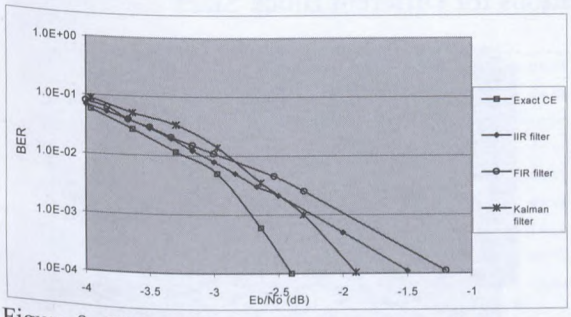


Figure 9. BER performance of a turbo coded system (L-band) experiencing light shadowing (B_dT = 0.05). The curves correspond to the following cases: i) when a Kalman filter is used in the fading estimation unit; ii) IIR filter; iii) FIR filter; iv) exact fading is known.

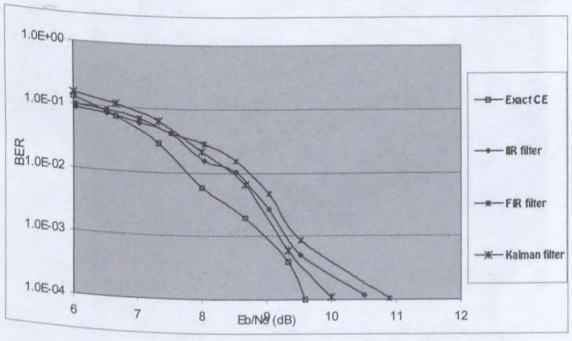


Figure 10. BER performance of a turbo coded system (L-band) experiencing heavy shadowing (B_dT = 0.05). The curves correspond to the following cases: i) when a Kalman filter is used in the fading estimation unit; ii) IIR filter; iii) FIR filter; iv) exact fading is known.

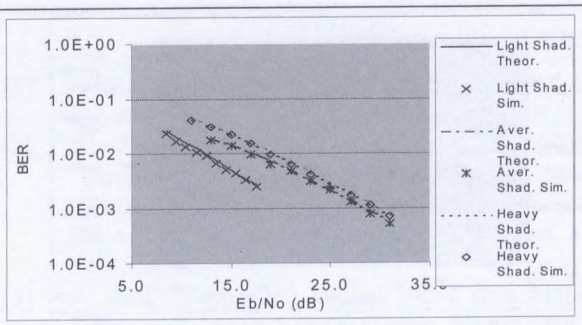


Figure 11. LMSC Bit Error Rate Comparison of Theoretical vs. Simulation Results for Light, Average, and Heavy Shadowing at the Ka-band.

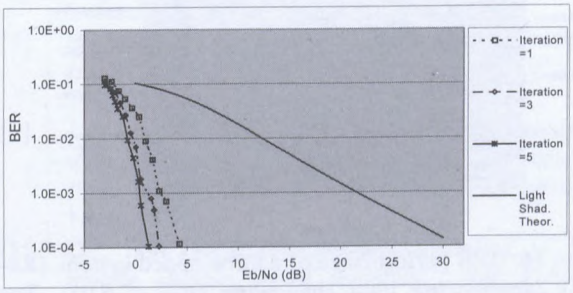


Figure 12. Turbo Codes Results for Light Shadowing (Ka-Band) with Block Size = 1032 bits and B_dT =0.005

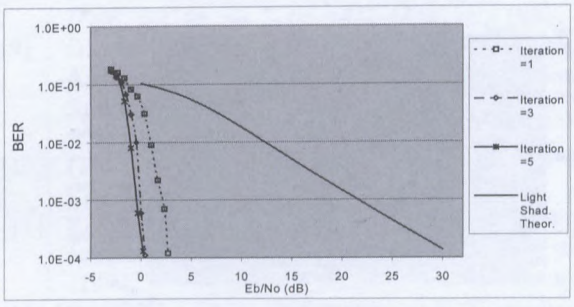


Figure 13. Turbo Codes Results for Light Shadowing (Ka-Band) with Block Size = 1032 bits and B_dT =0.05

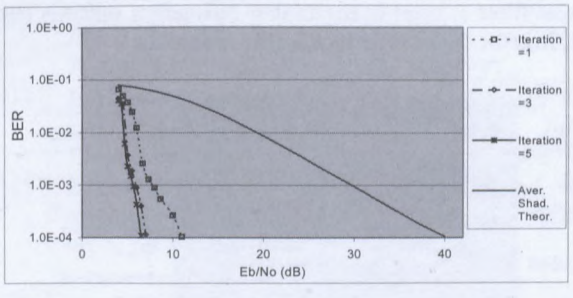


Figure 14. Turbo Codes Results for Heavy Shadowing (Ka-Band) with Block Size = 1032 bits and B_dT =0.005.

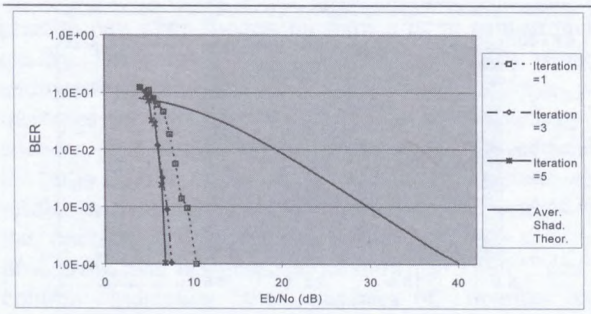


Figure 15. Turbo Codes Results for Heavy Shadowing (Ka-Band) with Block Size = 1032 bits and B_dT =0.05.

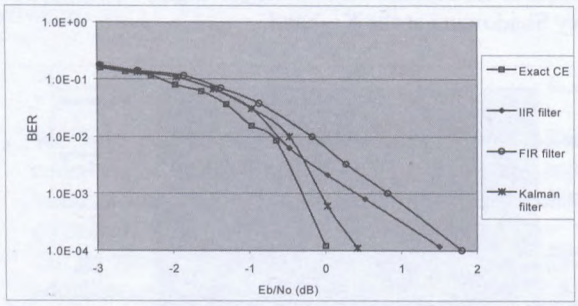


Figure 16. BER performance of a turbo coded system (Ka-band) experiencing light shadowing (B_dT = 0.05). The curves correspond to the following cases: i) when a Kalman filter is used in the fading estimation unit; ii) IIR filter; iii) FIR filter; iv) exact fading is known.

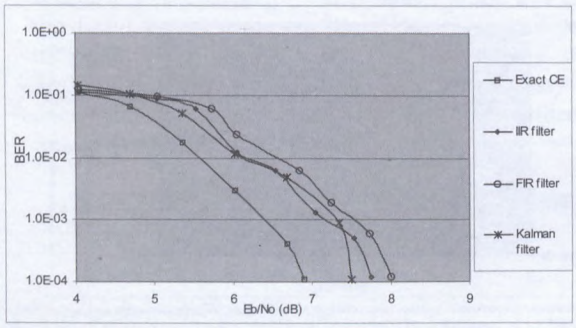


Figure 17. BER performance of a turbo coded system (Ka-band) experiencing heavy shadowing (B_dT = 0.05). The curves correspond to the following cases: i) when a Kalman filter is used in the fading estimation unit; ii) IIR filter; iii) FIR filter; iv) exact fading is known.

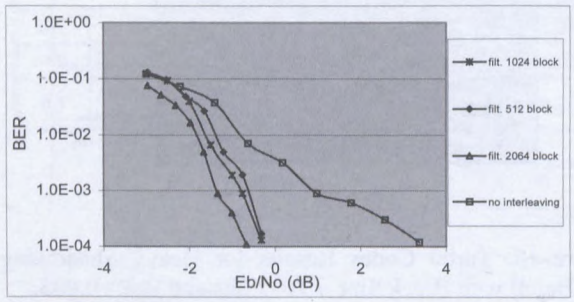


Figure 18. Different Block Sizes are Compared to Each Other for Average Shad. in the L-band and B_dT Product of 0.05 and no Interleaving.

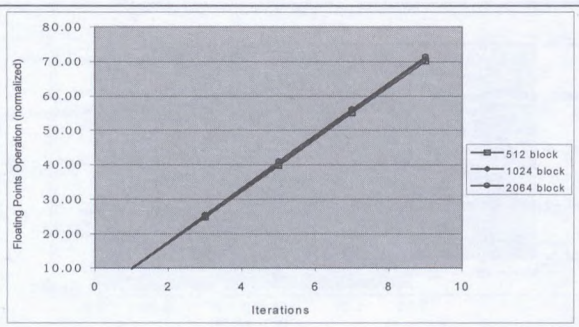


Figure 19. Floating Point Operations vs. Iterations for Different Block Sizes

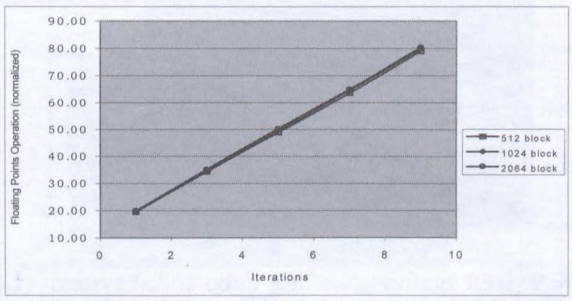


Figure 20. Floating Point Operations vs. Iterations for Different Block Sizes

Figure 21. Frame # 1 with no errors.

Figure 22. Frame # 6 with no errors.

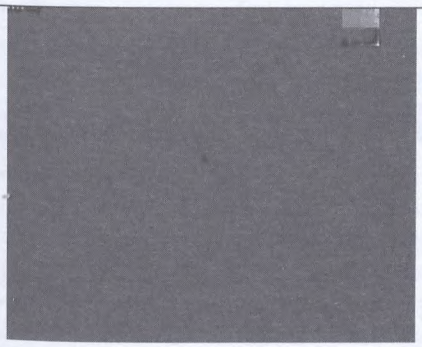


Figure 23. Original H.263 Frame # 1 with errors

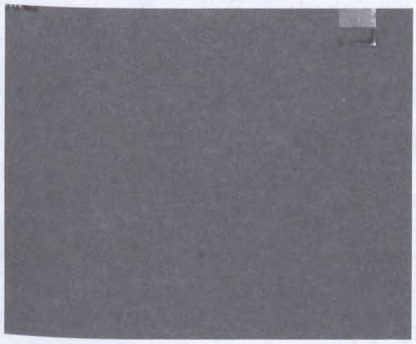


Figure 24. Original H.263 Frame # 6 with errors.

Figure 25. Modified H.263 Frame # 1 with errors.

Figure 26. Modified H.263 Frame # 6 with errors.

REFERENCES

[1]. B. Miller, "Satellites free the mobile phone," *IEEE Spectrum,* vol 35, no. 3, pp. 26-35, Mar. 1998.

[2]. W. Zhuang, J. Y. Chouinard, and D. Makrakis, "Dual-Space Diversity over Land Mobile Satellite Channels Operating in the L and K Frequency Bands," *Wireless Personal Comm.,* vol. 4, no. 3, pp. 277-298, May 1997.

[3]. P. Bahl, "Supporting Digital Video in a Managed Wireless Network," *IEEE Comm. Magaz.,* vol. 36, no. 6, pp. 94-102, June 1998.

[4]. C. Berrou, A. Glavieux, and Thitimasjshima, "Near Shannon limit error-correcting coding and decoding: turbo-codes," *ICC 1993,* May 1993, Geneva, Switzerland, pp. 1064-1070.

[5]. B. Sklar, "A primer on Turbo Code Concepts," *IEEE Comm. Magaz.,* vol. 35, pp. 94-102, Dec. 1997.

[6]. J. Hagenauer and P. Hoeher, "A Viterbi algorithm with soft-decision outputs and its applications," *Proc. IEEE Glovecom Con* . (Dallas,TX), pp. 1680-1686, 1989.

[7]. C. Loo, Measurements and modeling of land-mobile satellite signal statistics," *1986 Vehicular Tech. Conf.,* Vol. 34, pp. 262-267, Dallas, TX, May 1986.

[8]. C. Loo and J. S. Butterworth, "Land Mobile Satellite Channel Measurements and Modeling," *Proc. of the IEEE,* vol. 86, no. 7, pp. 1442-1462, July 1998.

[9]. D. Guerrero and D. Makrakis, "Performance Analysis of Turbo Codes for Land Mobile Satellite Systems Operating in the L and Ka Bands," ICT'99, 1999.

[10]. ITU-T Recommendation H.263 (1995): "Video Coding for Low Bitrate Communication".

[11]. R. M. Mersereauand J.T.Smith, Digital filtering : a computer laboratory textbook. New York : J. Wiley, c1994.

[12]. G. Welch and G. Bishop, "An Introduction to the Kalman Filter," *UNC Chapel Hill,* TR 95-041, Oct. 1998.

[13]. M. C. Valentind B. D. Woerner, "Variable latency turbo codes for wireless multimedia applications," *Proc., Inter. Symp. On Turbo Codes and Related Topics,*(Brest, France), pp. 216-219, 1997.

[14]. J. Hagenauer, "Iterative decoding of binary block and convolutional codes," *IEEE Trans. Inform. Theory,* vol. 42, no. 2, pp. 429-445, Mar. 1996.

[15]. C. Loo, "Digital transmission through a land mobile satellite channel," *IEEE Trans. on Comm.,* vol. 38, no. 5, pp. 693-697

A Study of Next-Generation LEO System for Global Multimedia Mobile Satellite Communications

Ryutaro Suzuki, Keiichi Sakurai, Shinichi Ishikawa, Iwao Nishiyama and Yasuhiko Yasuda*
Next-generation LEO System Research Center,
Telecommunications Advancement Organization of Japan, * WASEDA University
2-11, Higashida, Kawasaki-Ku, Kawasaki 210-0005, Kanagawa, Japan
Email: ryutaro@nels.tao.go.jp

ABSTRACT

Next-generation LEO System (NeLS) Research Center started the development of Next-generation LEO System which would provide a global multimedia mobile satellite service by means of a group of LEO satellites with the user data rate up to 2 Mbps for handy terminals. Firstly, the satellite constellation of the NeLS was designed which could avoid the Van Allen radiation belts and reduce the delay of multimedia communications. Then, NeLS Research Center is focusing on the development of key technologies such as the satellite antenna, ISL system, on-board ATM switch and modulation methods.

INTRODUCTION

In Japan, the Ministry of Posts and Telecommunications (MPT) conducted a preliminary study in 1996 on a Next-generation LEO System (NeLS). The target year for commercial implementation is 2010. In order to develop the feasibility study, the NeLS Research Center was formed by the Telecommunications Advancement Organization of Japan (TAO), in cooperation with the telecommunications operators, manufacturers, universities and governmental research organization in the end of 1997.

In the Research Center, constellation of the NeLS was designed for realizing the multimedia mobile satellite communication system.

Assuming that the direct-radiating, active phased array antenna technology is applied, the satellite antenna will need several thousand radiating elements. Development of such a large-scale satellite antenna would be one of the most important breakthroughs. As a means of relaxing the requirement of the satellite antenna design, the realization of higher antenna gain in the user terminal using the phased array technology should be also investigated.

The inter-satellite link (ISL) technology is also important, because the inter-satellite network is essential to realize the low delay network connection for multimedia services. ATM technology can be applicable to the satellite on-board switch. However, new ATM control algorithm should be developed for the LEO system having a dynamic topology.

This paper describes the service image, constellation design, link budget calculations, and key technologies to be developed in the NeLS Research Center.

FUTURE PROSPECTS FOR MOBILE SATELLITE COMMUNICATIONS

Service Image of the NeLS

Multimedia Communications: In 2010, high-speed terrestrial networks using asynchronous transfer mode (ATM) technologies will have already been established all over the world. Through these high-speed backbones, the present voice communications, computer communications, and broadcasting will be merged into multimedia communications.

Global Mobile Communications: Currently, first-generation LEO/MEO satellite systems (e.g. Iridium, Globalstar and ICO systems) are in service or in preparation, which provide mainly voice communications through handy terminals. NeLS will provide global multimedia service. Fig. 1 shows the position of the NeLS.

SYSTEM DESIGN OF THE NeLS

LEO Constellation Design

The LEO constellation parameters such as orbital altitude, orbit inclination, number of satellites, number of orbital planes are discussed from the following points [1].

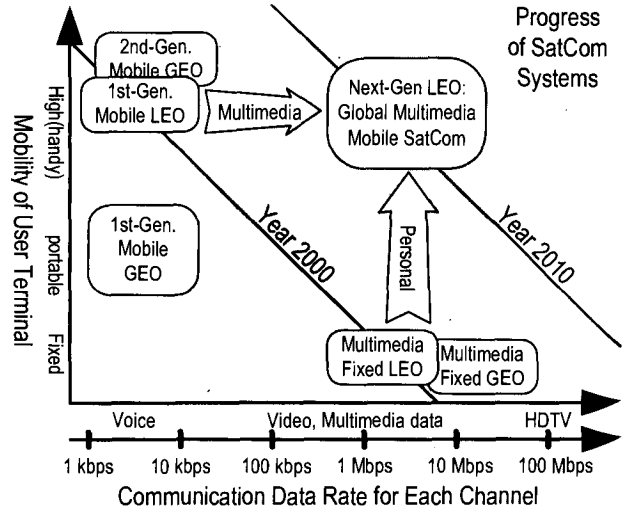


Fig.1 Position of the NeLS

Van Allen Belts Radiation Effect: Since radiation levels depend on the orbital altitude and inclination, the selection of the orbital parameters is made from the point of minimizing the radiation effects such as Single-Event Upsets (SEU), Single-Event Latch-up (SEL) and the defect of semiconductors caused by total accumulated dose. To avoid the effects of South Atlantic Anomaly of Van Allen belts, the altitude above 1000 km is desired . The relation among total accumulated dose (using an aluminum shield of 100 mil), altitude (from 700 to 1,300 km) and inclination (from 35 to 75 deg.) is shown in Fig. 2.

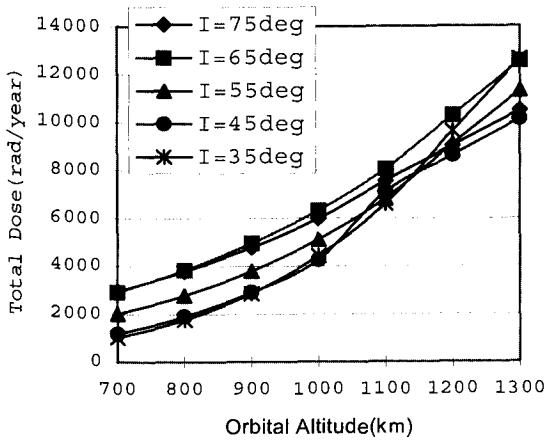


Fig. 2 Altitude/Inclination versus total accumulated dose (Aluminum 100mil)

The mission lifetime of the satellite is considered about 5 years. Radiation hardness level for commercial semiconductor devices is considered 5 years × 10^4 rad/year. Therefore, the altitude under 1,200 km and the inclination of 75, 55, 45 and 35 deg. (i.e. excluding 65 deg.) are preferred.

Minimum Elevation Angle: The necessary number of satellites and that of orbits are calculated for a specified value of minimum elevation angle. Here, we assume the minimum elevation to be 20 deg.

Service Area Coverage: According to the world population density by latitude [2], it can be deduced that the area from –60 deg. to +70 deg. in latitude should be covered.

Inter Satellite Links: The operation of both the intra- plane and inter-plane ISL are assumed. If a circular orbit is employed, the link distance and pointing angle between neighbor satellites is constant for the intra-plane ISL. As for the inter-plane ISL, the relation between a satellite and its neighbor on the adjacent plane is time-variant but simple because these satellites on different planes move synchronously. In this regard, the circular orbit is preferable. The assumed ISL requirements are as follows;

Intra-plane ISLs per satellite : 2
Inter-plane ISLs per satellite : 2
Link Distance : < 5,000 km
Angular Coverage : - 90 to + 90 deg. Azimuth
30 deg. Elevation

Selected Constellation: The candidate constellations obtained from the above discussions are as follows;

Number of orbital planes : 10
Number of satellites per orbital plane : 12
Orbital altitude : 1,200 km
Eccentricity : 0
Inclination : 75, 55, 45, 35 deg.
Difference angle of ascending node between adjacent orbital planes : 36 deg.
Phase angle : 3 deg.

The computer simulations were carried out to check whether the candidate constellations meet our requirements. Fig. 3 shows the minimum elevation angle versus latitude. The figure shows clearly that the constellations of I=35, 45 and 75 deg. do not meet to the minimum elevation angle requirement of 20 deg.

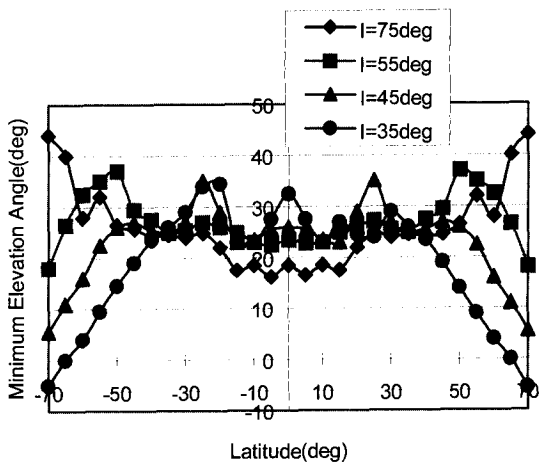


Fig. 3 Minimum elevation angle versus latitude

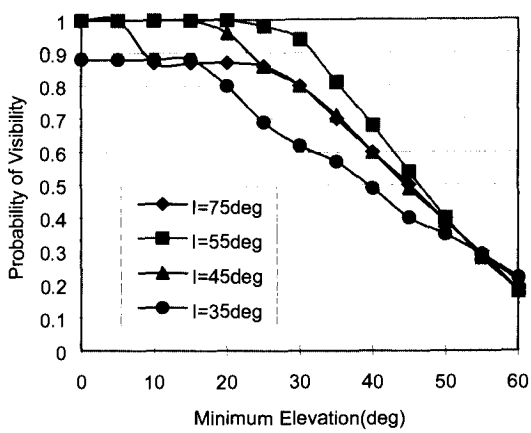


Fig. 4 Probability of visibility versus minimum elevation angle

Fig. 4 shows probability of visibility versus minimum elevation over the region from –70 to 70 deg. in latitude. From the simulation results of Fig. 3 and Fig. 4, we selected the inclination of 55 deg. Then, ISL configuration was evaluated for I=55 deg. As the results, the pointing angle variation for the inter-plane ISL can be kept within -55 to

+55 deg. in azimuth and -11 to -19 deg. in elevation. Distance variation for the inter-plane ISL can be kept within 3000 to 5000 km.

Link Budgets of the NeLS

Table 1. Link Budgets of the NeLS

Up-Link	C-band	Ku-band	Ka-band	Ka-band
User terminal				
TX Power (W)	3.0	1.0	1.0	1.0
Feeder Loss (dB)	4.8	0.5	0.5	0.5
Antenna Gain	3.0	13.0	16.0	20.0
EIRP (dBW)	6.8	12.5	15.5	19.5
Propagation				
Frequency (GHz)	5.2	13.0	18.0	30.0
Sat. Altitude (km)	1200	1200	1200	1200
Path (km; 20deg)	2456.0	2456.0	2456.0	2456.0
Free space Loss	174.6	182.5	185.4	189.8
Atmosph. Loss	2.0	2.0	2.0	2.0
Satellite RX				
Antenna Size (m)	3.5	3.0	2.6	1.7
Antenna Gain*	41.5	48.1	49.7	50.5
Feeder Loss (dB)	0.5	0.5	0.5	0.5
System Noise (K)	600.0	700.0	700.0	800.0
G/T (dB/K)	13.2	19.2	20.8	20.9
Sat C/No (dBHz)	72.0	75.8	77.5	77.2
Data Rate	2.0	2.0	2.0	2.0
Eb/No (dB)	9.0	12.8	14.5	14.2
Required Eb/No	6.2	6.2	6.2	6.2
Margin (dB)	2.8	6.6	8.3	8.0

Down-Link	C-band	Ku-band	Ka-band	V-band
Satellite TX				
TX Power/1ch (W)	3.0	1.0	1.0	1.0
Feeder Loss (dB)	0.5	0.5	0.5	0.5
Antenna Size (m)	3.5	3.0	2.6	1.3
Antenna Gain* (dBi)	44.0	46.7	49.7	50.6
EIRP (dBW)	48.3	46.2	49.2	50.1
Propagation				
Frequency (GHz)	6.9	11.0	18.0	40.0
Sat. Altitude (km)	1200	1200	1200	1200
Path (km; 20deg)	2456.0	2456.0	2456.0	2456.0
Free space Loss (dB)	177.0	181.1	185.4	192.3
Atmosph. Loss (dB)	2.0	2.0	2.0	2.0
User terminal				
Antenna Gain (dBi)	3.0	12.0	16.0	23.0
Feeder Loss (dB)	1.0	0.5	0.5	0.5
System Noise (K)	480.9	566.3	521.75	621.8
G/T (dB/K)	-24.8	-16.0	-11.7	-5.4
Receive C/No (dBHz)	73.0	75.7	78.8	79.0
Data Rate (Mbps)	2.0	2.0	2.0	2.0
Eb/No (dB)	10.0	12.7	15.8	16.0
Required Eb/No (dB)	6.2	6.2	6.2	6.2
Margin (dB)	3.8	6.5	9.6	9.8

* Antenna gain includes the steering Loss of the phased array antenna for 52.3 deg. beam offset.

Link Parameters: Link budget calculations of the NeLS, focusing on the provision of 2 Mbps user links, were carried out for the C-, Ku-, Ka- and V-band systems.

Terminal Selection: In the link budget calculations, two types of user terminals were assumed;
i) C-band handy terminal with 3 dBi non-tracking (omni-directional) antenna.
ii) Ku- to V-band portable terminals with the electronically steering antenna having 5 cm aperture.

Transmission power of user terminal and satellite is 3 W/1ch for C-band system, and 1 W/1ch for Ku- to V-band system. In each band system, the link budget was calculated for the user data rate of 2 Mbps.

Table 1 shows the link budgets of the NeLS.

Evaluation of the Link Budgets

The proposed LEO system could support the data rate of 2Mbps only on good link conditions, and could reduce the data rate by using the variable rate modulation technique (e. g. variable rate CDMA), if the satellite link is degraded by fading or shadowing.

3 dBi omni-directional antenna of the user terminal can be used for C-band system when the transmission power of user terminal and satellite is 3W/1ch instead of 1W/1ch. The active phased array antenna with digital beam forming technology for the user terminal is assumed to be used for Ku-, Ka- and V-band system, because the size of the phased array antenna is small enough to be installed in the handy terminal. In this case, antenna pointing control technology for the handy terminal is a key technology to be developed.

The link margin for 2 Mbps data transmission can be kept between 2.8 and 8.3 dB for up-link, and between 3.8 and 9.8 dB for down-link, respectively, in these band systems. This margin should include rain margin, fading margin and hardware implementation losses. Especially in the case of handy terminal similar to cellular phone, fade margin should be kept higher than the vehicular terminal. We assume that the user terminal can operate with variable rate condition, where, using the variable rate modulation, we can get a fading margin of about 30 dB at least for the voice only or low data rate communications.

DEVELOPMENT OF KEY TECHNOLOGIES

Satellite Antenna

According to the above link budget calculations, a high gain satellite antenna, whose diameter is approximately 4 m in the C-band system, should be designed if the user data rate of 2 Mbps is to be provided in the LEO system. A promising one is a direct radiating active phased antenna with very narrow multi-beams. Assuming that this kind of antenna is applied to the NeLS, it will need a few thousand radiating elements, each element being connected with a

very small-sized module consisting of filters, amplifiers, and phase shifters. Development of such a large-scale direct-radiating, active phased array antenna would be one of the most important breakthroughs.

Design of Satellite Antenna: The development of key devices for the active array antenna is, at the first step, planned at C band. Since the antenna element size is of the same order as a wavelength, C band frequency is tractable for the antenna design where the MMIC (including filters, amplifiers and digital phase shifters) and antenna element could be easily combined in one IC package.

When we design Ka band antenna at the next step, the size of devices could be scaled down to about 1/3 to 1/5 compared with the C band system. If Time division duplexing (TDD) scheme between the up and down links is employed, the antenna design would be easier from the point that the TX/RX diplexer in the active antenna element could be replaced as the small sized switch.

Deployable Antenna: A method to mechanically deploy such a large phased array antenna on the spacecraft shall also be investigated. One of the candidates is a two dimensional deployable phased array antenna [3]. Fig.5 shows the deployment sequence of the antenna structure.

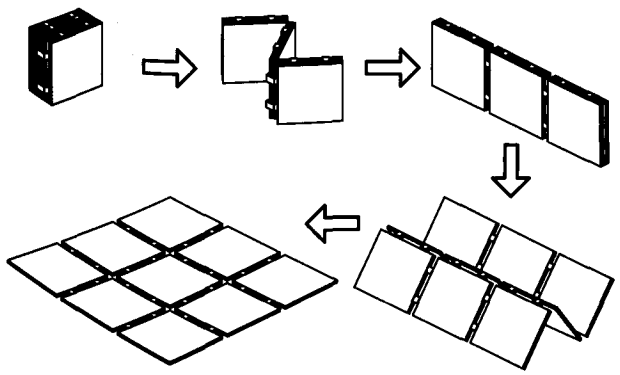


Fig. 5 Deployment sequence of the antenna structure

Beam Scanning: As stated above, the satellite antenna for the NeLS should have very narrow multi-beams. If the beam footprint on the earth moves according to the satellite's orbital motion like the Iridium system, the hand-over between two adjacent multi-beams occurs too frequently for each user link. Then, we assume to employ the ground fixed beam method that each satellite antenna beam is controlled to fix the beam footprint to a certain area on the earth. Fig. 6 shows the concept of the satellite antenna control.

Beam Forming Network: The multi-beams are controlled by the beam forming network (BFN). There are two kinds of BFNs conceivable, i.e., analog BFN and digital BFN. The analog BFN, which uses digital phase shifters in RF band, could be realized with the current technology. The

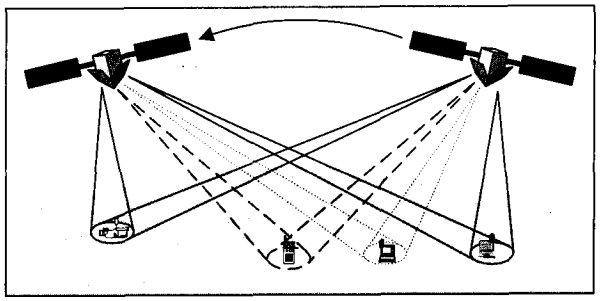


Fig. 6 Ground fixed beam control of the satellite antenna

digital BFN, which uses digital signal processing technology instead of RF processing and is suitable for the manipulation of a large number of multi-beams, could be realized if the progress of IC technology enables the lower power consumption and faster processing speed of devices. Then, NeLS Research Center started the development of the digital BFN as a sub array unit of the deployable antenna shown in Fig. 5.

Satellite Network Design

In the NeLS, ISL might be essential to flexibly provide global multimedia communication services. We assume that the optical space communication technology is employed for the ISL, because the smaller sized antenna and higher data rate could be realized in the optical ISL than other alternative, i.e., Ka or V band ISL. The conceptual block diagram of on-board satellite communication subsystem is shown in Fig. 7.

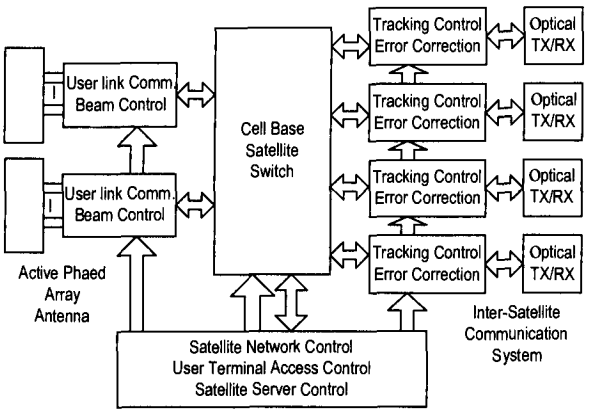


Fig. 7 Concept of on-board satellite communication subsystem

In Fig. 7, optical TX/RX units are prepared for ISL connections. User link communication units with an active phased array antenna are used for the communications with user terminals and with the gateway stations having high gain antenna. On-board cell base satellite switch works like the ATM switch. Satellite server control unit works as the proxy server for non-realtime data services. If required, this server can be used for IP router, multicast server, etc.

Optical ISL

Since the provision of high speed multimedia communications service is targeted in the NeLS, it is

assumed that the optical ISL system shall have meet the link performances:

- Transmission speed: up to 10Gbps/link total
- Bit error rate: less than 10^{-10} for ATM connection

The followings are some of key issues to be developed for the optical ISL system.

- High power/ high efficiency optical amplifier
- Modulation/detection method:
  (e.g.) Intensity Modulation /direct detection (IM-DD), PSK/coherent detection, or DPSK/differential detection
- WDM technology
- Error Correction method for optical ISL system

Intra-plane Optical Network: Optical ISLs among the satellites in the same orbital plane can be designed with ring network as shown in Fig. 8. In this example, circular orbit is assumed and 7 (instead of 12, in convenience) satellites are placed with equal spacing in one orbital plane.

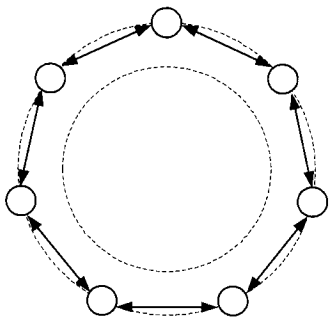


Fig. 8 Ring topology for intra-plane connection

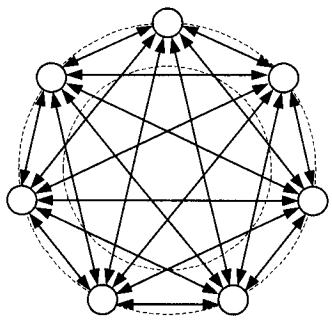


Fig. 9a Full mesh topology for intra-plane connection

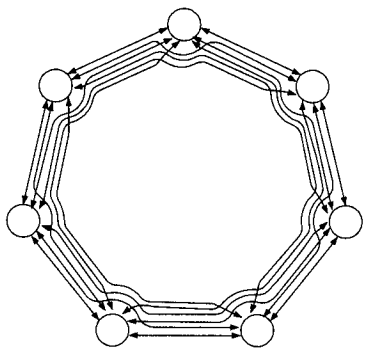


Fig. 9b WDM Full mesh connection for intra-plane ISL

For constructing this optical ISL ring network, the utilization of wavelength division multiplexing (WDM) technology is envisaged. WDM is a multiplexing scheme which combines optical signals of different wavelengths into one fiber, and it has emerged as a means of increasing the point-to-point link capacity in the terrestrial and undersea fiber optic systems. In addition, WDM has got the enhancement in network capabilities due to the emergence of wavelength add-drop multiplexers (WADMs) which enable arbitrary wavelength(s) to be added or dropped at a single site without demultiplexing the entire wavelength bundle [6].

If this WDM technology is applied to the above optical intra-orbital plane ISL, logical full mesh connections can be achieved by assigning the appropriate wavelength for all links between N(=7) satellites as shown in Fig. 9a and 9b. The number of wavelengths necessary for such full mesh connections is $(N^2 - 1)/8$ [7].

In the case that both the originating and destination user terminals are located in the areas covered by the satellites of the same orbital plane, in such full mesh connections, the traffic signal visits only two on-board cell switches of the source and destination satellites and bypasses the on-board cell switches of the satellites located between them. This is the different point from Teledesic, in which the traffic signal visits the on-board switches of all the satellites located between the source and destination satellites [4].

Fig.10 shows the concept of ISL subsystem with WDM technology.

Inter-plane Optical Network: Optical ISLs among the satellites in different orbital planes are shown in Fig. 11. As discussed in LEO constellation design, inter-plane ISL parameters are as follows;

Inter-plane ISLs per satellite :	2
Number of orbital planes :	10
Number of satellites per orbital plane :	12
Orbital altitude :	1,200 km
Eccentricity :	0
Inclination :	55 deg.
Difference angle of ascending node between adjacent orbital planes :	36 deg.
Phase angle :	3 deg.

ISLs are assumed to operate when the directions of two neighboring satellites in adjacent orbits are in parallel. As shown in Fig. 11, inter-plane ISLs form dual spiral connection. In this case, all satellites are connected by single ring topology.

Traffic Routing Algorithm

Study of traffic routing algorithm in a moving satellite node environment is one of key issues for the project. It includes a comparison between the datagram [4] and virtual channel approaches [8] with regard to the ATM cell transmission. Computer simulation is planned for this study.

On-board Switch

In order to develop the on-board cell switch, the NeLS Research Center has a plan to prepare the software simulation model and the hardware model based on conventional ATM switches.

Device technology for the cell switch which is endurable in the space radiation environment is also a subject for further study.

User Terminal/Multimedia Terminal

There is no fixed image for the mobile multimedia terminal. Some ideas for the multimedia terminal are proposed in the IMT-2000 system. User terminal for the NeLS should be compatible with the IMT-2000 terminal.

CONCLUSION

The Next-generation LEO System (NeLS) Research Center has started to develop the key technologies for the global multimedia mobile satellite communications system, and has a plan to launch experimental demonstration satellites.

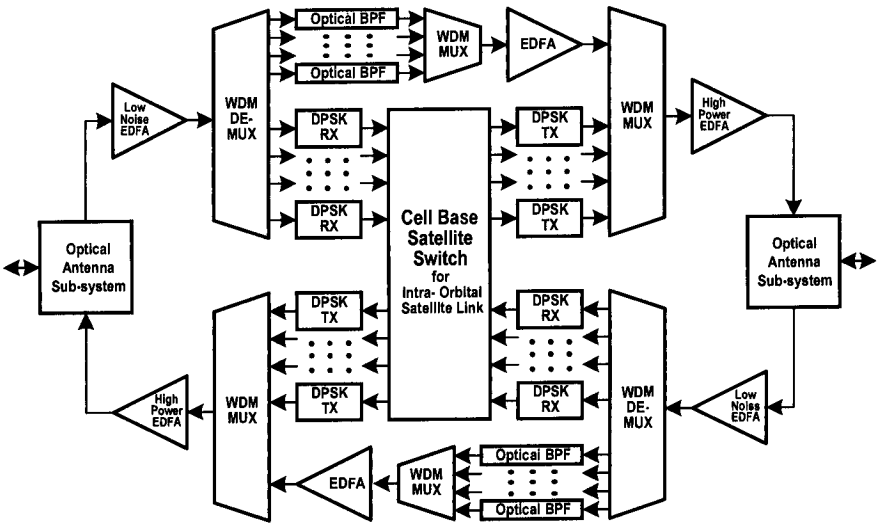


Fig. 10 Concept of ISL subsystem with WDM technology

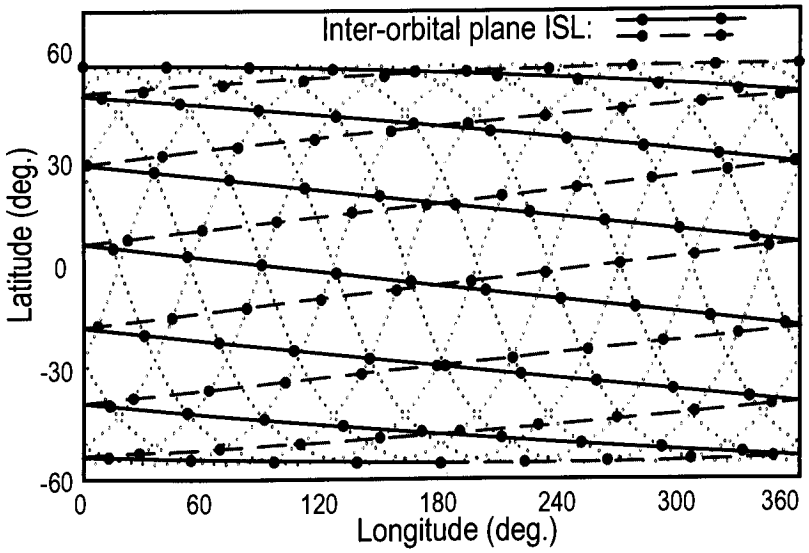


Fig. 11 Inter-orbital plane optical ISL connection

REFERENCES

[1] S. Ishikawa et al., "Conceptual LEO Satellite Constellation Design", 2nd International Symposium on Spacecraft Ground Control and Data Systems, Brazil, February 1999.

[2] D. Diekelman, "Mission Design and Implementation of Satellite Constellation, Design Guidelines for POST-2000 Constellations", Proceeding at an international workshop, Toulouse, France, November 1997.

[3] M.Tabata, "Conceptual Study of a 2-D Deployable Active Phased Array Antenna," ICAST'97.

[4] M.A.Sturza, "Architecture of the TELEDESIC satellite system," IMSC'95, pp.212-218.

[5] C. Argagnon et al., "From Stentor to Skybridge, different problems one product line for active antennas," IAF'97, M 4.07.

[6] J.P.Ryan, "WDM: North American Deployment Trends," IEEE Commun. Mag., Feb. 1998, pp.40-44.

[7] A.F.Elrefaie, "Multiwavelength Survivable Ring Network Architectures," ICC'93, pp. 1245-1251.

[8] M.Werner,"A Dynamic Routing Concept for ATM-Based Satellite Personal Communication Networks," IEEE JSAC, Oct. 1997, pp.1636-1648.

New Signal Structures for Future GNSS

Robert Schweikert, Thomas Woerz
Audens ACT Consulting GmbH, P.O. Box 1110, D-82224 Seefeld, Germany
E-mail: Robert.Schweikert@audens.com, Thomas.Woerz@audens.com

Riccardo De Gaudenzi
European Space Agency, Keplerlaan 1, P.O. Box 299, NL-2200 AG Noordwijk ZH, The Netherlands,
E-mail: rdegaude@estec.esa.nl

Alexander Steingass, Armin Dammann
Institute for Communications Technology, DLR, P.O. Box 1116, D-822230 Wessling, Germany
E-mail: Alexander.Steingass@dlr.de, Armin.Dammann@dlr.de

ABSTRACT

This paper presents results of an ESA study on a desirable signal structure for a second generation satellite navigation wrt. to the synchronisation performance of spreading code and carrier phase measurements. Two channel types are investigated. Results for the AWGN channel are based on theoretical findings; results for multipath/fading channels are gained by simulations.

INTRODUCTION

The Global Positioning System (GPS) has been demonstrating the huge potentiality of a first generation satellite based navigation system. Among others, GPS civil signal performance is limited by the so-called Selective Availability (SA) inserted on purpose to limit civil users positioning accuracy and by the non availability of a second carrier frequency for civil applications. The two above limitations, although affecting the final achievable accuracy, have been partly counteracted by elaborated post-processing or augmentation techniques. The latter approach is being pursued by the so-called (European) GPS navigation overlay system (EGNOS) complementing the GPS constellation with Geostationary satellites and providing integrity information for critical applications such as aircraft landing. Furthermore, recent announcement by USA Authorities indicates their intention to provide a second civil frequency for the second generation GPS together with SA removal.
In parallel to this GPS evolutionary approach, the European Space Agency, in co-ordination with the European Commission, is investigating the main aspects of a truly innovative second generation civil navigation system dubbed now *Galileo* whose main signal design drivers are

- Compatibility with the frequencies bands for satellite navigation that are likely to become available in the near future.
- Target UERE accuracy in the order of 1-2 m with single carrier ranging and 1-2 cm with three carriers differential phase positioning techniques.
- Reduced time-to-first-fix.
- Provision of integrity information and extended navigation message to users requiring enhanced performance.
- Enhanced signal robustness to the satellite fading channel.
- Adoption of state-of-the-art digital communication techniques.

The above challenging targets have been achieved in a recent ESA study activity adopting innovative signal design concepts finding their roots in modern digital wireless CDMA communication techniques. The complete results can be found in [DLR98b], refer also to [DSP98] [GNSS98]. In this contribution, extracted from [DLR98b], we focus on the synchronisation performance (spreading code and carrier phase measurements) of the proposed signal structure in several mobile environments.
The contribution is organised as follows: Firstly, the frequency plan is illustrated and the proposed signal structure are introduced. Then the synchronisation performance is elaborated based on theoretical findings and simulation results.

FREQUENCY PLAN

The signal baseline design has been performed to fit the L-band frequency slots that are likely to become available for a European GNSS2 system (refer to Figure 1). In the following frequency plan, the baseline bands are indicated with the acronyms E1, E2, E3, E4. As a matter of fact, E1, E2 and E3 bands are part of the ESA ENSS-1 frequency filing. However, during the signal study and related TCAR investigations [TCAR98], it became evident that the E3 0.6 MHz bandwidth is too narrow for resolving carrier phase ambiguity with TCAR in mobile applications. Therefore the use of the E4 band whose utilisation has been formally requested in a recent ESA ENSS-1 filing extension, is considered mandatory. The use of E1, E2, E4 bands allows to have 4 MHz available in all frequency bands (baseline). Larger bandwidth can be achieved sharing part the GLONASS bandwidth (named G1 and G2) in addition to E2. In this case, almost 20 MHz of spectrum becomes available in G1, G2 that jointly with E2 constitutes the so-called option 1. A second option is constituted by the use of a potential 30 MHz frequency slot at C-band (5000-5030 MHz) in alternative to E2. Option 2, is less likely from the regulatory standpoint and also requires a more complex user terminal RF front-end.

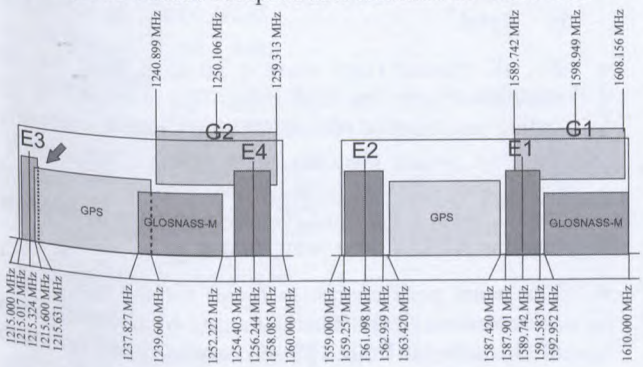


Figure 1 – Frequency plan for L-band

GNSS2 SIGNAL DESIGN

In the following section the results of the signal design [DLR98b] are presented. For the motivation and the reasoning of the chosen signal structure the interested reader is also referred to [DSP98] and [GNSS98]. The signal parameters are summarised in Table 2.

As indicated in the last section three signals in three different frequency bands form one option. Also, three different message types are foreseen for transmission: E-NAV, E-NAV', and EA-NAV. E-NAV and E-NAV' are intended to contain all the system data necessary for fulfilling the navigation task in a receiver such as e.g. satellite ephemeris and almanac, system time, ionospheric data, and integrity information. Due to the high data rate EA-NAV is foreseen to be used for the delivery of user group specific data (e.g. aeronautical users). It is intended to transmit partially the same data at the same time from several or all satellites to allow diversity reception for improved data detection performance.

The generation of each signal can be described with the block diagram shown in Figure 2. The incoming data stream is FEC encoded, interleaved and structured into frames with a length of one second. Furthermore, a preamble is added. The coded bits are then spread by a complex spreading sequence $c_p + jc_Q$. The two satellite specific sequences are taken from the set of Gold sequences with length 1023. The choice of the length is preliminary and subject to further investigations. The generated signal scatter diagram is QPSK like (refer to Figure 3).

The resulting chip stream is fed into a chip shaping filter with a square-root-raised-cosine (SRC) characteristic with roll-off $\beta = 0.2$. Assuming a chip rate of $R_c = 3.069$ Mcps each signal occupies a bandwidth of 3.68 MHz. The signal structure is referred as quadrature spreading with square-root-raised-cosine chip pulse shaping (SRC QPN).

For the multiple access from different satellites CDMA is chosen.

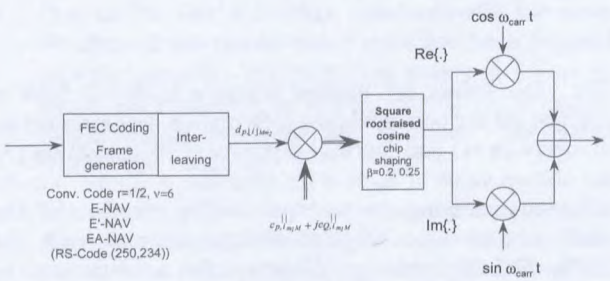


Figure 2 – Block diagram of signal generation

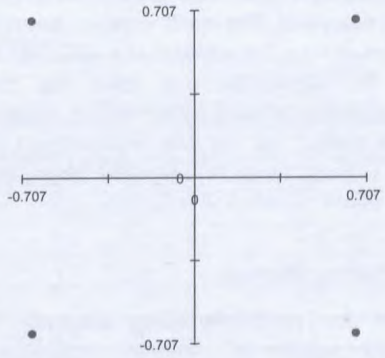


Figure 3 – Signal scatter diagram for simplified QPN

CODE / CARRIER PHASE PERFORMANCE

For the results shown in the following a simple receiver structure comprising a non-coherent DLL for spreading code phase tracking and a subsequent Costas loop for

carrier phase tracking are assumed. The adopted parameters are:

- NC-DLL: pre-detection bandwidth W=3000 Hz, early-late spacing 1 T_c, loop bandwidth B_{DLL}=2 Hz
- Costas loop: symbol rate T_s=3000 Hz, loop bandwidth B_{cl}=20 Hz

AWGN channel

In Figure 4 the performance of the different signals with respect to the standard deviation of the code phase jitter is shown for an AWGN channel. A scenario of K=12 satellites in view is considered where the interference of 11 satellites is modelled by an increased noise power (To take into account correlation losses due to Doppler shifts, the interferers are assumed to have twice the power of the desired signal). The code phase jitter is calculated according to the following relationships

$$\left(\frac{C}{N_0 + I_0}\right) = \frac{C/N_0}{1 + 2T_c(K-1)C/N_0} \quad (1)$$

$$\sigma = cT_c\sqrt{\frac{B_n}{2[C/(N_0 + I_0)]}\left(1 + \frac{2W}{[C/(N_0 + I_0)]}\right)} \quad (2)$$

The C/N_0 values are marked where a E_b/N_0=12.5 dB is reached for the corresponding information data rates for no diversity (L=1) and fourfold diversity (L=4) reception (At the chosen value of E_b/N_0 a bit error rate of $P_b=10^{-6}$ can be achieved assuming the multipath/fading channels of the next section). At $C/N_0 = 45$ dBHz, one can reach the E_b/N_0=12.5 dB with no diversity for a data rate of 1500 bps (E-NAV') and provide at the same time a code phase variance of 0.5 m (0.1 m) with a chip rate of 3.069 Mcps (15.345 Mcps). A reduced data rate of 750 bps (E-NAV) leads to further 3 dB increased margin (15.5 dB) for the data detection. The application of fourfold diversity increases the margin by additional 6 dB. Due to the high data rate of 12000 bps one need for EA-NAV a $C/N_0 \mp 47$ dBHz combined with fourfold diversity in order to reach the margin for the data transmission (As a result of the high C/N_0 and chip rate (24.552 Mcps) one achieves a low code phase variance of 0.02 m).

Multipath/Fading channel

The results for multipath/fading channels have been obtained by means of software simulations. The simulations results are referring to a QPN SRC signal with a chip rate of R_c = 3.069 Mcps (e.g. E1)
For the realisation of the multipath/fading channel model a tap-delay approach is used. Each tap represents a group of received echoes that can not further distinguished [PRO89].

The first tap is characterised by a Rician amplitude distribution (includes the direct path and a diffuse component) with a Rice factor C/M and a rectangular shaped Doppler power spectrum with a maximum Doppler frequency $f_{DMax} = (v/c) \cdot f_{carr}$ (v = speed of mobile, c = speed of light, f_{carr} = carrier frequency). The other taps are characterised by a Rayleigh amplitude distribution (includes only a diffuse component), a rectangular shaped Doppler power spectrum with a maximum Doppler frequency f_{DMax}, and a certain delay wrt. to the first tap. The parameters of the channel model are a function of the environment to be modelled. Table 3 summarises the parameters for the considered cases.

Results for code phase tracking (Figure 5):

- Performance of code phase jitter versus C/N_0 improves monotonically moving from channels 1) to 5):
  1) En-route aeronautical
  2) Urban pedestrian
  3) Final approach aeronautical
  4) Urban car
  5) Rural
- For all channel types there is an error floor for the code phase jitter for large C/N_0. This is caused by the fading processes of the echoes which result, especially for slow fading channels, in a time-varying loop S-curve zero crossing point. Depending on the channel model, from a certain C/N_0 this variation is stronger than the AWGN jitter contribution.
- The worst performance is found for the type " En-route aeronautical". This is caused by a combination of two effects. Both diffuse components (direct path and echo) have a very low Doppler bandwidth (1 Hz) and a relatively high power in the echo. With a loop bandwidth of 2 Hz the NC-DLL is able to track the combination of direct path and the diffuse components. Due to the fact that the presence of an echo causes a slow varying bias for the loop-S curve zero-crossing point, we get a large variation for the resulting measured code phase.
- For the type " Final approach aeronautical " the jitter is improved mainly due to the large Doppler bandwidth (420 Hz) of the echo. In this case the NC-DLL can not track the variations caused by the echo.
- The "Urban pedestrian" is affected by the number of echoes and their relative large delay (up to 250 ns = 75 m). According to the multipath envelope considerations, e.g. considered in [DLR98b], see also [PAR96], the maximum impact of an echo occurs for

delays between 50 and 75 m for the SRC QPN signal. Again the effect is magnified by the small Doppler bandwidth of 4 Hz.

- The code phase jitter performance is further improved for "Urban car" and "Rural" due to the reduced multipath power and the large Doppler bandwidth of 70 and 140 Hz, respectively.

Results for carrier phase tracking (Figure 6):
It should be noted that the carrier phase error is here measured with respect to the phase of the direct path (which is set to zero). For other applications (e.g. in the communication area) the phase error is measured with respect to the phase corresponding to the sum of the direct path and its diffuse component. This reflects the fact that for symbol or bit detection one tries to gather as much energy as possible and also the diffuse component is helpful. For measuring the carrier phase of the direct path, the diffuse component disturbs the measurement and is therefore included in the error.
The parameters of the multipath/fading channels are the same as given in Table 3.
Before looking on the results for the carrier phase jitter the following difference between the DLL and the Costas loop behaviour should be noted.

- Influence of echo delay:
  - The DLL is mostly disturbed by an echo with a certain delay greater than zero.
  - The Costas loop is mostly disturbed by an echo with delay zero. The influence of echoes further apart is weighted by the auto-correlation function.
- Influence of echo phase:
  - The DLL is maximal disturbed by an echo phase of 0 or 180 degree (at 90 or 270 degree the influence is minimum).
  - The Costas loop is maximal disturbed by an echo phase of 90 or 270 degree (at 0 or 180 degree the influence is minimum).

Both loops have in common that a higher Doppler bandwidth of multipath/fading components lead to an improved tracking result. Fast variations of these components relative to the loop bandwidth are filtered out and do not disturb the estimated quantity (code or carrier phase jitter).

The results for the carrier phase jitter can be summarised as follows (see Figure 6):

- Performance of carrier phase jitter is increasing for C/N_0 in the same order as the code phase jitter:
  1) En-route aeronautical: The worst performance is caused by a small delayed strong echo combined with a very low Doppler bandwidth.
  2) Urban pedestrian: The bad performance is caused by the large power of the diffuse component affecting the direct path (zero delay!).
  3) Final approach aeronautical: Slightly better performance compared to the "Urban pedestrian" channel due to a smaller power of the diffuse component. The influence of the echo is very small due to the high Doppler bandwidth.
  4) Urban car: Improved performance due to relative high Doppler bandwidth. Main disturbance is caused by diffuse component at delay zero.
  5) Rural: refer to "Urban car" channel.
- For all channel types there is an error floor for the carrier phase jitter for increasing C/N_0 caused by the echoes with a delay greater zero, which leads to a time-varying zero-crossing point for the Costas loop S-curve .
- Due to the fact that cycle slips occurred for some simulations the carrier phase error has been mapped into the interval $[-\pi/2, \pi/2]$. The points in Figure 6, which are connected by a line, include the mapping. Single points represent the resulting jitter value without this mapping.
- Table 1 summarises the number of carrier phase cycle slips observed for a simulation duration of 30 s. They occur with a relative high frequency for the "En-route aeronautical" and "Urban (pedestrian)" channel. This coincides with the observation that these channels also experience the highest carrier phase jitter (see Figure 6).

	Channel				
C/N_0 [dB]	En-route	Final approach	Urban (car)	Urban (ped.)	Rural
30	16	8	5	8	5
35	3	-	-	-	-
40	3	-	-	3	-
45	2	-	-	-	-
50	1	-	-	-	-
55	1	-	-	-	-
60	2	-	-	3	-

Table 1 – Number of carrier phase cycle slips observed for a 30 s simulation period (in multipath/fading environment)

Summarising the results for the carrier phase jitter the following can be stated under the assumption that a direct

component is available and a possible frequency deviation (e.g. caused by satellite and/or user movement) is corrected e.g. by a frequency estimation loop:

- The reduction of the loop bandwidth offers the advantage that the carrier phase jitter becomes smaller. This is due to the fact that the tracking loop can not follow the carrier phase changes caused by the echoes. Only the phase of the direct component is tracked. As a consequence, also the cycle slip rate is decreasing.
- On the other hand, if a loss of lock occurs the time for re-acquisition is increasing with decreasing loop bandwidth.

SUMMARY AND CONCLUSIONS

It has been demonstrated that using the proposed signal design and exploiting modern digital signal processing techniques the required ranging accuracy can be achieved with a modest L-band bandwidth occupation in AWGN (code phase jitter $\sigma = 0.5$ m at $C/N_0 = 45$ dBHz)
In mobile environments the code (carrier) phase jitter varies from 0.5 m to 5 m (0.5 – 2 cm) for a $C/N_0 = 45$ dBHz depending on the type of channel. Poor performance can be expected in channels, where the diffuse component has a delay smaller than one chip length combined with a strong power and/or a Doppler bandwidth in the order of or smaller than the DLL (PLL) bandwidth

ACKNOWLEDGEMENTS

The authors thank the European Space Agency ESA/ESTEC for sponsoring the study for which the work has been carried out.

BIBLIOGRAPHY

[DLR98b] R. Schweikert and T. Woerz: „Final Report", Signal Design and Transmission Performance Study for GNSS2, ESA Ref. AO/1-3156/NL/JSC, 1998.

[DSP98] R. Schweikert, T. Woerz and R. de Gaudenzi: „On Signal Design for a Second Generation Satellite Navigation System", Proc. DSP'98, 6th International Workshop on Digital Signal Processing Techniques for Space Applications, session 6.2, Sept. 1998.

[GNSS98] R. Schweikert, T. Woerz and R. de Gaudenzi: „On New Signal Structures for Future GNSS", Proc. GNSS'98, Second European Symposium on GNSS, pp. IX-O01 1-7, Oct. 1998.

[PAR96]. B.W. Parkinson and J.J. Spilker: „GPS: Theory and Applications Volume I", Progress in Astronautics and Aeronautics, Vol. 163, 1996

[PRO89] J. Proakis: „Digital Communications", McGraw-Hill, 1989

[TCAR98] Spectra Precision TERRASAT and Socratec GmbH, Study on Precise Relative Position Using GNSS-2 TCAR, ESA Ref. 12406/97/NL/DS,

	Target C/N_0 [dBHz]	Carrier Freq. [MHz]	Chip rate [Mcps]	Data stream	Info. Bit rate [bps]	FEC code rate (coded data rate [bps])	Code length (Gold seq.) [chips]	Code duration [μ s] (code periods/cod. bit)
Baseline								
E1	45	1589.742	3.069	E-NAV'	1500	1/2 (3000)	1023	333,3 (1)
E2	45	1561.098	3.069	E-NAV'	1500	1/2 (3000)	1023	333,3 (1)
E4	45	1256.244	3.069	E-NAV	750	1/2 (1500)	1023	333,3 (2)
Option 1								
G1	45	1598.949	15.345	E-NAV'	1500	1/2 (3000)	1023	66,7 (5)
E2	45	1561.098	3.069	E-NAV	750	1/2 (1500)	1023	333,3 (10)
G2	45	1250.106	15.345	E-NAV'	1500	1/2 (3000)	1023	66,7 (5)
Option 2								
E1	45	1589.742	3.069	E-NAV'	1500	1/2 (3000)	1023	333,3 (1)
E4	45	1256.244	3.069	E-NAV	750	1/2 (1500)	1023	333,3 (2)
C	47	5014.746	24.552	EA-NAV	12000	1/2 (24000)	1023	41,7 (1)
	(38		15.345	E-NAV'	1500	1/2 (3000)	1023	66,7 (5))

Table 2 - Signal parameters for different options

	C/M (direct path) [dB]	Doppler BW [Hz]	Number of echoes	Delay [ns]	Relative Power [dB]	Doppler BW [Hz]	C/M (total) [dB]
En-route	15	1	1	50	-3	1	3.0
Final Approach	10	1	1	44	-6	420	4.7
Urban car	7	70	4	60 100 130 250	-27 -27 -27 -27	70 70 70 70	6.8
Urban ped.	7	4	4	60 100 130 250	-27 -27 -27 -27	4 4 4 4	6.8
Rural	6	140	2	100 250	-28 -31	140 140	5.9

Table 3 – Channel parameters for use in L-band

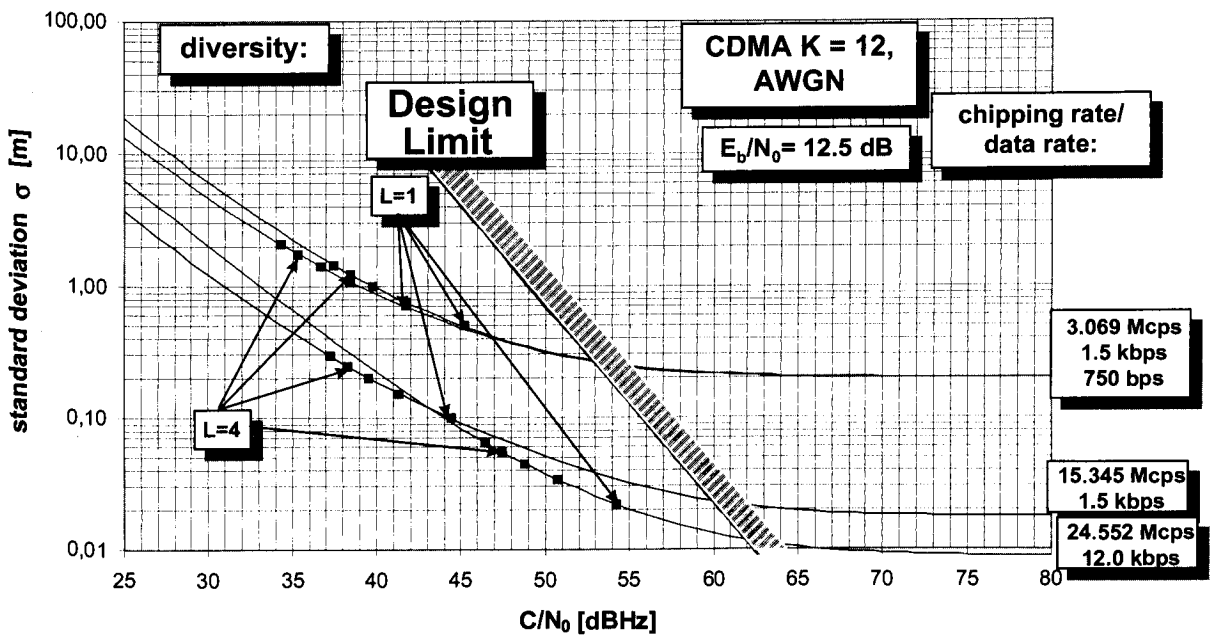


Figure 4 – Code phase jitter vs. C/N_0 for SDS Options

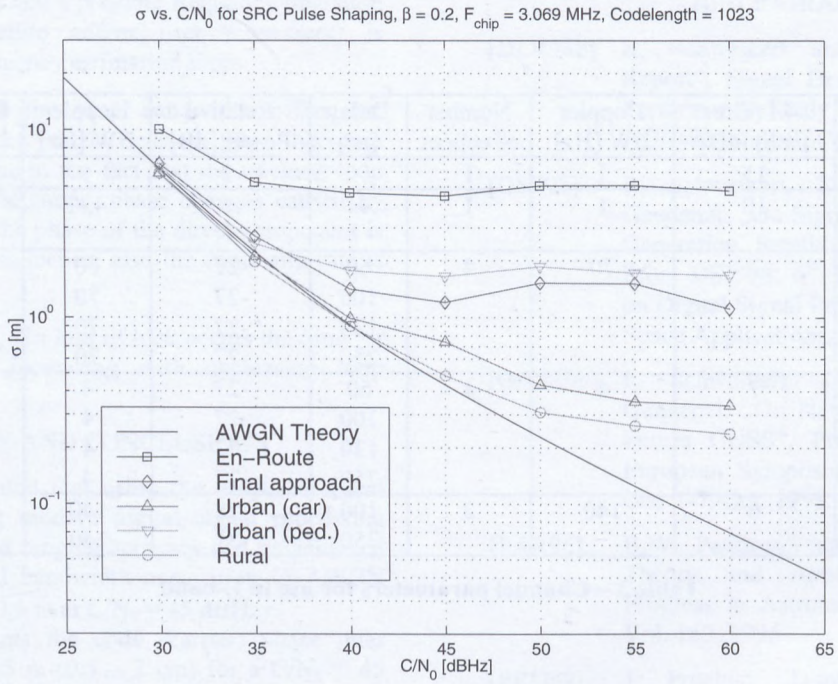


Figure 5 – Code phase jitter for considered QPN signal in multipath/fading environment

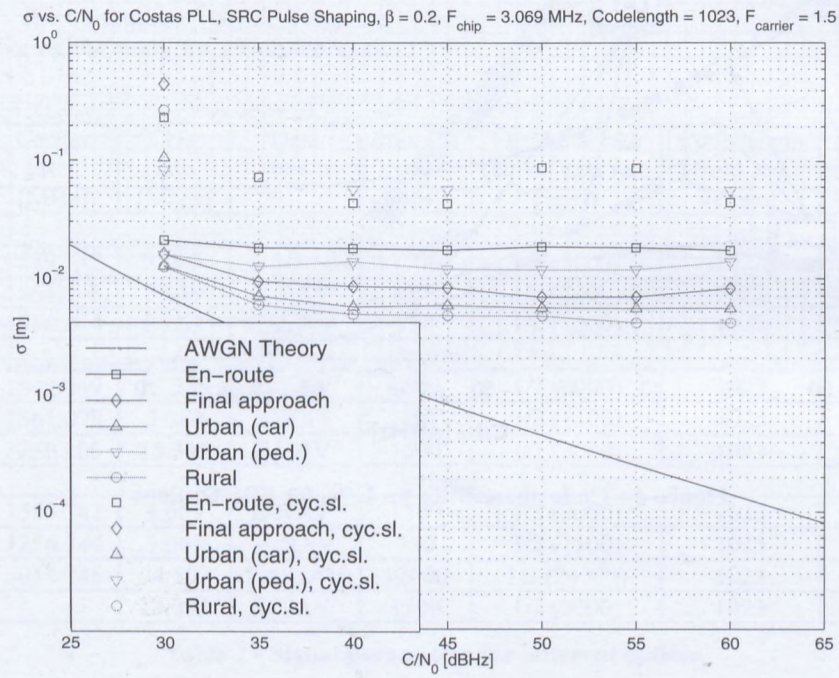


Figure 6 - Carrier phase jitter for considered QPN signal in multipath/fading environment

Overview of the Cospas-Sarsat Satellite System for Search and Rescue

J.V. King
Communications Research Centre,
3701 Carling Avenue, Ottawa, ON, K2H 8S2, Canada
e-mail: jim.king@crc.ca

ABSTRACT

Cospas-Sarsat, an international satellite system for search and rescue, started operating in 1982 and has been credited with saving thousands of lives since then. Hundreds of thousands of aviators, mariners and land users worldwide are equipped with Cospas-Sarsat distress beacons, which could help save their lives in emergency situations anywhere in the world.

This paper outlines the evolution of the system and describes how satellites are constantly circling the globe monitoring for 'SOS signals', while tracking stations on six continents receive the satellite signals, compute the locations of the distress events and forward the calls for help to the appropriate rescue authorities.

This humanitarian system is unique in the way that it is funded and operated, while its use remains free of charge to the end user in distress.

BACKGROUND

In the 1970s, light aircraft were carrying small, battery-operated radio transmitters that could be activated in an emergency distress situation. Such transmitters, called Emergency Locator Transmitters (ELTs), operating at the international distress frequency of 121.5 MHz, emitted a low-power signal that could be picked up by a receiver in another aircraft in the vicinity or in a nearby air traffic control tower. Some types of ELTs could be automatically activated by the impact of a crash to transmit the distress signal without human intervention. Marine vessels also started carrying similar distress beacons, called Emergency Position Indicating Radiobeacons (EPIRBs), which could float off a sinking ship and automatically emit a distress signal.

However, if a plane or ship went down in a remote area or in inclement weather, there might be no aircraft around to detect the distress signal for days or even weeks, which could be long after the distress beacon's batteries were depleted.

By the mid-1970s, more than 250,000 distress beacons were in service in Canada, Europe and the USA. Lives of aviators and mariners were being saved thanks to these transmitters, but there was still room for improvement, particularly as it was now the 'space age'.

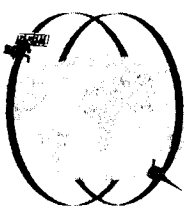
A NEW SATELLITE SYSTEM

To improve the detection of such distress signals, particularly from remote areas, the concept of a satellite receiving system was proposed. In the mid-1970s, experiments were conducted in Canada, by the Communications Research Centre and the Department of National Defence, which used an amateur radio satellite, called OSCAR, to demonstrate the feasibility of using satellites for detecting and locating the source of distress signals. Similar experiments at NASA in the United States and the French Space Agency (CNES) further showed the technical viability of such a satellite system. These agencies agreed to set up a joint experiment for search and rescue satellite-aided tracking (SARSAT).

In 1979, the former USSR (and later Russia) agreed to join the experiment and develop a compatible system called COSPAS (a Russian acronym, Cosmicheskaya Sistyema Poiska Avarinyich Sudov, meaning a space system for the search of vessels in distress), and the Cospas-Sarsat System was born.

DEMONSTRATION AND EVALUATION PHASE

The development of an automated system to detect and locate very weak distress signals presented a formidable challenge. In the late 1970s and early 1980s, the design of the Cospas-Sarsat system was begun, radio frequencies were allocated, host satellites were arranged, search and rescue satellite payloads were designed and built and special ground receiving stations, called local user terminals (LUTs) were developed and installed. The world's first Cospas-Sarsat LUT was located in Ottawa, Canada.



The basic Cospas-Sarsat System [1] utilizes a constellation of four low-Earth-orbit (LEO) satellites in near-polar orbit, as depicted in Figure 1. With this type of orbit, a single satellite eventually scans the entire globe.

Figure 1: Cospas-Sarsat satellites in polar orbit.

However, there is a time delay because the global coverage is not continuous, due to the limited instantaneous field of view of a low-altitude satellite.

When the first satellite was launched in 1982 the 'experiment' was officially underway, and within days a real 121.5 MHz distress signal was detected. The 'experimental system' made headlines when all 3 people onboard a small aircraft were successfully rescued after their plane crashed in the mountains in a remote area of British Columbia, Canada. Their ELT distress signal was picked up by an overflying satellite and relayed to the ground station in Ottawa, some 4000km away, where the location of the distress was automatically computed. This information was sent to the search and rescue authorities and a rescue plane was soon on scene.

Even while the experiments were being conducted to assess the technical performance of the System, real distress signals were routinely being detected and every few days additional lives were being saved, thanks to the Cospas-Sarsat System. The first maritime rescue occurred soon after in the Atlantic Ocean, off the east coast of the USA.

In addition to providing distress alerting and locating services for the hundreds of thousands of existing owners of 121.5 MHz distress beacons, Cospas-Sarsat was also developing a new, more sophisticated, distress beacon operating at 406 MHz. This type of beacon allowed the distress location to be pinpointed more accurately and also transmitted a unique identification code. Search and rescue forces would then know where, as well as what, they were going to search for, making for a more effective mission.

At the outset, the demonstration and evaluation of the experimental Cospas-Sarsat System was scheduled to conclude in the mid-1980s, but the proven success of the system created enough demand for it to be continued, rather than being turned off. Since several other countries had also participated in the experiment and used the system to save lives, finding a way to transform it into an operational system was highly desirable. The four founding countries undertook to set up an official, worldwide System and declared the System operational in 1985.

FORMALIZING THE SYSTEM

In 1987, the Cospas-Sarsat Secretariat was established at the headquarters of the International Maritime Satellite Organization (Inmarsat) in London. In 1988, a formal intergovernmental agreement was signed, thus assuring the long-term continuity of the System, in which three United Nations agencies were also involved:

- the International Maritime Organization (IMO) for worldwide shipping,
- the International Civil Aviation Organization (ICAO) for worldwide aviation, and
- the International Telecommunication Union (ITU) for radio frequency allocations.

Participation by various other government bodies and industry was also initiated in order to get equipment standards adopted, new distress beacons type approved, manufactured and distributed to consumers, and more ground receiving stations installed around the world.

The System continued to expand with more countries sharing in its operation and use, more ground stations coming on line and more distress beacons being installed on ships and aircraft, resulting in more lives being saved every year thanks to the Cospas-Sarsat System.

In 1985, Cospas-Sarsat also started evaluating the use of geostationary-Earth-orbit (GEO) satellites as an enhancement to the polar-orbiting system to provide almost immediate alerts, with identification, for 406 MHz beacons. This topic is presented in greater detail in a related paper at IMSC '99 (Demonstration and Evaluation of 406 MHz Geostationary Search and Rescue Systems), so is not described further in this paper.

PRINCIPLE OF OPERATION

System Concept

The basic concept of the Cospas-Sarsat System [1] is illustrated in Figure 2.

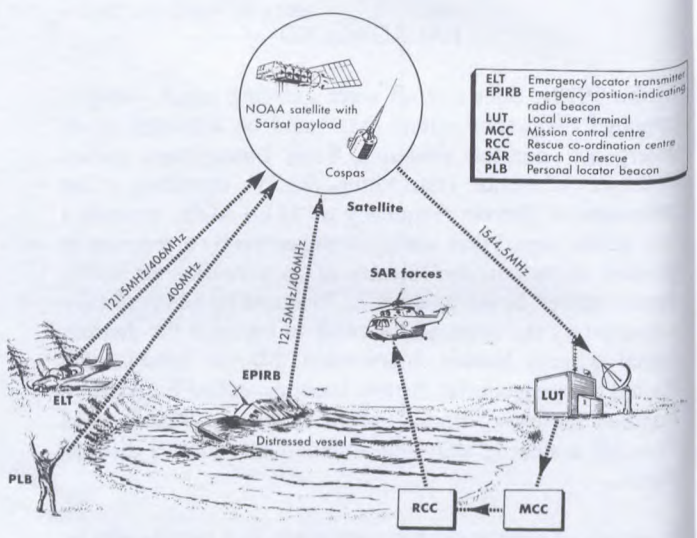


Figure 2: Basic Concept of the Cospas-Sarsat System

There are three types of radiobeacons: aviation ELTs, maritime EPIRBs and Personal Locator Beacons (PLBs). These beacons transmit signals that are detected by Cospas-Sarsat satellites equipped with suitable receivers. The signals are relayed to Cospas-Sarsat LUTs, which process the signals to determine the beacon location. Alerts are then relayed, together with location data, via a Mission

Control Centre (MCC), either to another MCC or to the appropriate Rescue Coordination Centre (RCC) or a Search and Rescue point of contact (SPOC) in that region.

Satellite Configuration

Figure 3 shows the path, or "orbital plane", of a satellite circling the earth around the poles. The satellite travels in this plane while the earth rotates underneath it, enabling a single satellite to eventually view the entire Earth's surface. At most, it takes only one-half rotation of the Earth (i.e. 12 hours) for any location to pass under the orbital plane.

Figure 3: Orbital plane of a polar-orbiting satellite

With a second satellite, having an orbital plane at right angles to the first, only one quarter of a rotation is required, or 6 hours maximum. Similarly, as more satellites orbit the Earth in different planes, the waiting time is further reduced. The Cospas-Sarsat System design constellation is four satellites, which provide a typical waiting time of less than one hour at mid-latitudes.

Doppler Effect

Satellites at low altitude must move quickly over the Earth to stay in orbit. This movement causes a shift in the radio frequency called the "Doppler effect", as illustrated in Figure 4. Doppler location, using the relative motion between the satellite and the beacon, is the means used to locate these very simple devices. The resultant "Doppler curve" of frequency versus time has an inflection point when the satellite is at its time of closest approach (TCA) to the beacon.

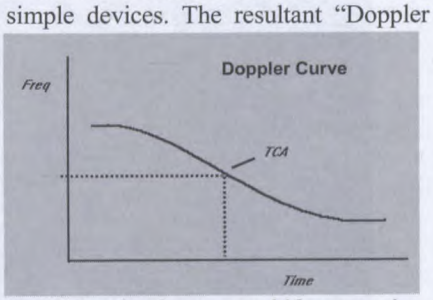


Figure 4: Doppler frequency shift versus time

Cospas-Sarsat selected low-altitude satellites in order to optimize Doppler performance and to be able to detect lower power distress beacons, while the near-polar orbit provides full global coverage, albeit with some time delay.

The Doppler calculation generates two possible positions for each beacon, the true position and its mirror image relative to the satellite ground track, as illustrated in Figure 5. This ambiguity in position can be resolved either by waiting for a second satellite pass or by calculations that take into account the Earth's rotation. On a subsequent satellite pass another pair of positions would be produced, but only one of those would overlap with one from the previous pass, thus establishing which is the true position of the beacon.

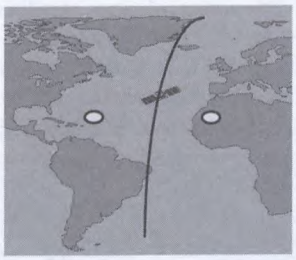


Figure 5: Two possible positions, one on each side of the satellite sub-track, result from Doppler positioning

With appropriate frequency stability, as specified for 406 MHz beacons, the true solution may be determined in a single satellite pass, because of the slight skewing of the Doppler curve due to the Earth's rotation during the 15-minute satellite pass. For 121.5 MHz beacons, normally a second satellite pass is required to resolve the ambiguity.

SYSTEM DESCRIPTION

The Cospas-Sarsat System comprises the space segment and the ground segment up to the point where the alert data leaves the MCC. The subsequent RCCs and SAR response units are existing national or regional entities, which utilize Cospas-Sarsat alert data to facilitate their operations. The distress beacons procured by the users are controlled on a national basis, and the 406 MHz beacons must be type approved by Cospas-Sarsat to ensure they meet the Cospas-Sarsat performance requirements.

Space Segment

The nominal System configuration comprises four satellites, two Cospas and two Sarsat, but often more than four are in operation, since older satellites, even when replaced by newly-launched satellites, continue to be used as long as they can provide some service.

Russia supplies two Cospas satellites, shown in Figure 6, placed in near-polar orbits at an 83° inclination at 1000 km altitude and equipped with SAR instrumentation at 121.5 MHz and 406 MHz.

Figure 6: Cospas satellite

The USA supplies two multi-mission NOAA meteorological satellites, shown in Figure 7, placed in sun-synchronous, near-polar orbits at a 98° inclination at about 850 km altitude, and equipped with SAR instrumentation at 121.5 MHz and 406 MHz supplied by Canada and France.

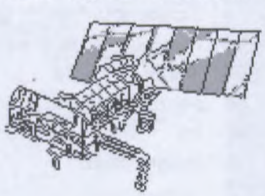


Figure 7: USA's NOAA satellite

A 243 MHz payload, onboard some of the satellites, operates in the same manner as the 121.5 MHz system, so is not further described in this paper. The 121.5 MHz payload is a repeater that retransmits all received signals in real-time, which is called the 'local mode' of operation. The 406 MHz payload repeats the band, as well as partially processes the received beacon signals, and retransmits them in real-time and also stores the digital data in satellite memory, which is continually replayed for several orbits to broadcast the data to all LUTs. This is called the 'global mode' of operation.

Each satellite makes a complete orbit of the earth around the poles in about 100 minutes, travelling at a velocity of 7 km per second. The satellite views a "swath" of the earth more than 4000 km wide as it circles the globe, giving an instantaneous "field of view" about the size of a continent. When viewed from the earth, the satellite crosses the sky in about 10 to 15 minutes, depending on the maximum elevation angle of the particular pass.

Distribution of Alert and Location Data

The alert and location data generated by LUTs are forwarded to appropriate SPOCs through the Cospas-Sarsat MCC network. Since a single distress incident is usually processed by several LUTs, in particular in the 406 MHz global mode of operation, the alert and location data are sorted by MCCs to avoid unnecessary transmission of identical data. The principle of continuous downlink transmission to all LUTs in visibility of a satellite allows simple downlink transmission procedures and provides a high level of redundancy in the ground processing system, both at 121.5 MHz within overlapping LUT coverage areas and at 406 MHz worldwide.

Distress Beacons

121.5 MHz beacons: It is estimated that there are now about 600,000 121.5 MHz beacons in use worldwide, primarily aboard aircraft, and are required to meet national specifications based on ICAO standards, which were not initially developed for a satellite system.

Such beacons, as illustrated in Figure 8, transmit only about a 0.05 to 0.1 Watt signal, having swept tone, amplitude modulation, which produces a warbling 'wow, wow, wow' sound in a nearby receiver. The carrier frequency of the beacon is not very stable and is significantly affected by the ambient temperature.

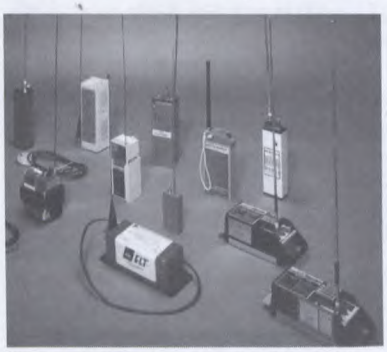


Figure 8: Typical 121.5 MHz ELTs used in aircraft

Therefore, the 121.5 MHz Cospas-Sarsat System was designed to serve the existing type of beacons, even though system performance would be constrained by their characteristics. Parameters such as System capacity (number of simultaneous transmissions in the field of view of the satellite which can be processed by LUTs) and location accuracy would be limited, and little or no information would be provided about the operator's identity. Even with these limitations, the efficiency of 121.5 MHz beacons has been greatly enhanced by the use of satellite detection and Doppler location techniques, and many lives have been saved.

406 MHz beacons: Development of a new generation of beacons transmitting at 406 MHz commenced at the beginning of the Cospas-Sarsat project. The 406 MHz units were designed specifically for satellite detection and Doppler location by having:

- high peak power output and low duty cycle;
- improved radio frequency stability;
- a unique identification code in each beacon;
- digital transmissions that could be stored in a satellite's memory; and
- spectrum dedicated by ITU solely for distress beacons

These parameters make the 406 MHz system superior to the older 121.5 MHz system by providing the following features:

- increased system capacity;
- improved ambiguity resolution and location accuracy (typically within 2 km versus 20 km for 121.5);
- identification of the user in distress;
- global coverage; and
- no interference from aircraft voice transmissions

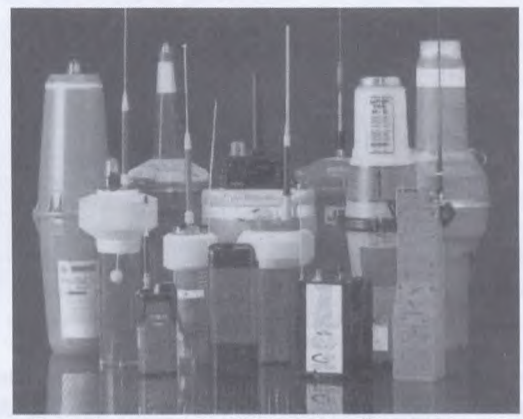


Figure 9: Various models of 406 MHz distress beacons

406 MHz beacons, shown in Figure 9, transmit a 5-watt, half-second burst approximately every 50 seconds. The carrier frequency is phase-modulated with a digital message. The low duty cycle provides a multiple-access capability of more than 90 beacons operating simultaneously in view of a polar orbiting satellite, versus only about 10 for 121.5 MHz beacons.

An important feature of 406 MHz emergency beacons is the addition of a digitally encoded message, which provides such information as the country of origin and the identification of the vessel or aircraft in distress, and optionally, position data from onboard navigation equipment. An auxiliary homing transmitter is usually included in the 406 MHz beacon to enable SAR forces to home on the distress beacon.

There are now more than 20 manufacturers of 406 MHz beacons in 12 countries, and many more distributors around the world, with over 100 different models type-approved by Cospas-Sarsat [2]. The number of 406 MHz beacons in use has increased dramatically from virtually zero in 1985 to about 20,000 in 1990 and to more than 156,000 by 1998.

Ground Segment

Local User Terminals (LUTs): Cospas-Sarsat LUTs are ground stations which track the satellites, receive the distress beacon signals via the satellites, compute the locations of the distress signals and forward the alert data to Mission Control Centres. Most LUTs are fully automated; some of which are unmanned and installed in remote areas and can be operated remotely from the MCC. The LUT steers a tracking antenna to follow the satellite across the sky, so each LUT needs to know the satellite orbit data on an ongoing basis.

All commissioned LUTs meet the Cospas-Sarsat specifications, but the configuration and capabilities of some LUTs may vary to meet the specific requirements of the participating countries. The Cospas and Sarsat spacecraft downlink signal formats ensure interoperability between LUTs and the various spacecraft.

For the 121.5 MHz signals, each transmission is detected and the Doppler information calculated. A beacon position is then determined using these data.

Processing of 2400 bits per second digital data from the satellite (i.e. those generated from 406 MHz transmissions) is relatively straightforward since the Doppler frequency is measured and time-tagged on-board the satellite. All 406 MHz data received from the satellite memory on each pass is processed within a few minutes of pass completion.

At the end of 1997, 38 LUTs were operating on six continents, as shown in Figure 10.

Figure 10: Locations of 38 Cospas-Sarsat LUTs in 1997

Mission Control Centres (MCCs): MCCs have been set up in most of those countries operating one or more LUTs. As depicted in Figure 11, an MCC is where distress alert data from the Cospas-Sarsat System is normally first viewed by a human, since most of the LUTs' receiving, processing and data forwarding systems are automated.

Figure 11: One of the Cospas-Sarsat MCCs (Canada)

The main functions of an MCC are to:

- collect, store and sort the data from LUTs and other MCCs;
- provide data exchange within the Cospas-Sarsat System; and
- distribute alert and location data to associated RCCs or SPOCs.

All MCCs in the System are interconnected through appropriate networks for the distribution of System information and alert data.

STATUS OF SYSTEM

Currently, some 30 countries are participating in Cospas-Sarsat [2], as shown in Figure 12.

Figure 12: Cospas-Sarsat Participants (shaded areas)

The number of ground segment elements in the Cospas-Sarsat System [2] continues to grow, as shown in Table 1. (figures to 1999 were not yet available)

Table 1:Number of System Elements by 1998

Element	Number
121.5 MHz Beacons	600,000
406 MHz Beacons	156,000
MCCs	19
LUTs	38

In 1997, the Cospas-Sarsat System provided assistance in rescuing 1,312 persons, as shown in Table 2. (figures for 1998 will not be available until mid-1999)

Table 2: Summary of System Operations in 1997

Type of Distress	No. of SAR Events	No. of Persons Rescued
Aviation	76	186
Maritime	257	1,036
Land	56	90
Total	**389**	**1,312**

The 406 MHz system was used in 218 of these events (817 persons rescued) and the 121.5 MHz system was used in the other 171 SAR events.

Since its inception, the Cospas-Sarsat System provided assistance in rescuing 8,666 persons in 2,636 SAR events, from September 1982 to December 1997.

ADMINISTRATION AND FUNDING

The cost of implementing and operating the Cospas-Sarsat System is shared by the various member governments, while the distress beacons are paid for by the users. However, users do not pay to access the Cospas-Sarsat System. The four founding countries (Canada, France, Russia and USA) provide the space segment and they, as well as several other countries, installed and operate ground receiving stations and Mission Control Centres. The administrative costs of the Secretariat are shared by all member countries.

The estimated investment to date in Cospas-Sarsat equipment is approximately US$ 500,000,000, of which about half is for the system ($200M for satellite payloads and $50M for LUTs & MCCs) and half is for the distress beacons ($250M) purchased by the users

The program is managed by the Cospas-Sarsat Council, shown in Figure 13, comprising representatives of member governments, and is supported by a Secretariat, based in London, and a group of technical and operational experts (Joint Committee) which meets periodically.

Figure 13: A Cospas-Sarsat Council manages the Program

In 1990, Cospas-Sarsat was awarded the Seatrade Annual Award for Achievement, shown in Figure 14, for its contribution to safety at sea. The Cospas-Sarsat System is an integral part of the IMO's Global Maritime Distress and Safety System (GMDSS) for worldwide shipping, and is also recommended by ICAO for aviation.

Figure 14: Seatrade Award presented to Cospas-Sarsat

FUTURE

Even after more than 16 years of operation, the Cospas-Sarsat System continues to provide a valuable global service and is still expanding with new enhancements and more and more users.

The Cospas-Sarsat System is expected to continue providing service for many years to come, especially as thousands of users are mandated to carry distress beacons. The use of older 121.5 MHz beacons might diminish in the future as the more sophisticated 406 MHz beacons become commonplace, particularly with the addition of 406 MHz geostationary satellites and the inclusion of global positioning system (GPS) chips in some new beacons.

Inmarsat also provides a maritime EPIRB service for beacons operating at L-band. Other satellite systems coming on line in the near future will offer two way communications for persons in emergency situations, which will be very beneficial in many cases. However, these new systems will not likely have user terminals that would activate automatically in a distress situation and trigger a response by search and rescue authorities on a worldwide basis, and yet have no user charges. These new systems will be able to assist persons in distress, but none are foreseen to replace the Cospas-Sarsat System.

ACKNOWLEDGEMENTS

The author expresses appreciation to the Cospas-Sarsat Secretariat in London for the collection and compilation of the system data presented in this paper.

Additional information about Cospas-Sarsat can be obtained from the Secretariat's web site at:

www.cospas-sarsat.org/cospas-sarsat

and from national administrations in participating countries.

REFERENCES

[1] Introduction to Cospas-Sarsat, December, 1994
[2] Cospas-Sarsat System Data, February 1999

DEMONSTRATION AND EVALUATION OF 406 MHz GEOSTATIONARY SEARCH AND RESCUE SYSTEMS

AMANDA MCDONALD

National Search and Rescue Secretariat
275 Slater Street, Ottawa, Ontario, Canada K1A 0K2
E-Mail: mcdonald@nss.gc.ca

BACKGROUND

The geostationary search and rescue (GEOSAR) satellite system consists of geostationary Earth-orbiting satellites and their associated ground processing facilities, that have the capability to detect transmissions from Cospas-Sarsat type approved 406 MHz distress beacons. These satellites orbit at altitudes of 36,000 km at approximately 0° latitude and fixed longitudes, appropriate to the requirements of the satellite provider. Because of the high altitude and fixed orbit position of the geostationary satellites, the geostationary system has the potential to offer several complementary advantages to the Cospas-Sarsat System of low altitude Earth-orbiting satellites (LEOSAR). These advantages include near-instantaneous beacon detection and alerting, near-instantaneous beacon locating for beacons capable of calculating and transmitting their location, and continuous monitoring of the 406 MHz frequency band within the satellite footprint.

At the request of the Cospas-Sarsat Council, a Demonstration and Evaluation (D&E) was conducted to confirm the expected benefits of adopting the GEOSAR satellite system as an enhancement to the Cospas-Sarsat LEOSAR System. The Council further directed that the D&E establish GEOSAR technical and operational performance characteristics.

GOALS

The goals of the D&E were to:

- characterise the technical performance of the GEOSAR components;
- characterise the operational performance of the GEOSAR system;
- evaluate the operational effectiveness of the GEOSAR system and determine the benefits to search and rescue of combined LEOSAR/GEOSAR operations; and
- provide the basis for recommendations to the Cospas-Sarsat Council.

METHODOLOGY

A D&E Plan "Cospas-Sarsat Demonstration and Evaluation Plan for 406 MHz GEOSAR Systems" was developed to provide the framework for the D&E.

The Plan outlined ten technical and eight operational objectives for which tests were to be undertaken, with guidelines for data collection, reporting and analysis. The technical objectives were developed to address the technical compatibility of the various GEOSAR components and to establish the baseline characteristics of the GEOSAR system. The operational objectives were developed to evaluate the contribution to SAR operations of alert data provided by 406 MHz GEOSAR systems and to provide operational experience in the use of GEOSAR alerts. Data collection and evaluation for the technical and operational objectives generally took place between July 1996 and February 1998.

Participation in the D&E was open to all Cospas-Sarsat Participants. Australia, Canada, Chile, France, Spain, the United Kingdom and the United States contributed data in support of technical and/or operational objectives and contributed to the drafting of the D&E study report. Other Cospas-Sarsat Participants (Algeria and Japan) also provided input to the drafting to the report.

RESULTS AND DISCUSSION

The successful completion of the ten technical objectives was hindered by radio frequency interference from a strong signal emanating from Peru. Because of the interfering signal, several of the technical objectives could not be accomplished or completed and are recommended as actions to be completed at the earliest opportunity. Despite this, the technical objectives confirmed that:

- the GEOSAR system could detect transmitting beacons that met Cospas-Sarsat technical specifications and which were visible;

- the GEOSAR system provided near-instantaneous detection and alerting of 406 MHz beacon transmissions;

- calibrated frequency measurements performed by the GEOLUTs were reliable and accurate enough to permit use of GEOSAR derived frequency data in improving LEOSAR positioning estimates;

- 4° is a conservative value to use as the published minimum elevation angle required to assure reliable GEOSAR reception of 406 MHz beacon transmissions; and

- sufficient technical data had been collected to recommend the incorporation of the GEOSAR system as a complement to the Cospas-Sarsat System.

Results from the eight operational objectives documented the performance and confirmed the effectiveness and benefits of the GEOSAR system. In particular, the D&E indicated, in the area of performance, that:

- the GEOSAR system provided a useful time advantage over the LEOSAR System. On average, the GEOSAR alert was received 46 minutes before the first corresponding LEOSAR alert; the median time advantage was 21 minutes;

- the coverage area of the GEOSAR system was a good complement to LEOSAR System coverage. More than 85% of 406 MHz alerts within the 0° elevation angle GEOSAR satellite footprints were detected by the GEOSAR system and valid explanations existed for cases which were not detected;

- there were many GEOSAR alerts, not detected by the LEOSAR System, that were single burst messages or of very short transmission duration which could be indicative of inadvertent activation or testing. However, a short duration transmission could also be the only indication of a catastrophic event; and

- the mean time of beacon transmission durations could be useful in computing 406 MHz system traffic estimates and in determination of system capacity.

The time advantage gained by the GEOSAR system's near-instantaneous alerting was clearly shown to benefit the outcome of SAR events. In particular:

- additional lives and property were determined to have been saved in specific distress cases where a GEOSAR alert was received and used by SAR forces; and

- benefits of the earlier GEOSAR alert and the use of encoded position data were shown in an intercontinental road race, where personal locator beacons, some encoded with one of the new Location Protocols, were used as emergency equipment.

The completeness, accuracy, availability and 24-hour access capability of beacon registration databases were shown to be essential for the above benefits of GEOSAR system to be fully realised. The database information was used to:

- distinguish between real and false alerts and could be useful in preventing the launch of resources on false alerts; and

- obtain rough location information, enabling SAR personnel to take advantage of the earlier notification provided by GEOSAR, even when encoded position information was not available in the beacon message.

Analysis of the impact of the GEOSAR system on the Ground Segment of the Cospas-Sarsat System, notably the system-generated false alert rate, workload, and resolution of LEOSAR location ambiguity, indicated that:

- the number of undetected system-generated false alerts was low. All, but three, system-generated false alerts were eliminated through the use of effective screening measures at the MCCs, before transmission to SAR services. However, some concern exists that the system-generated false alert rate could be higher than that found by the D&E and this element needs monitoring. There is also the possibility of a higher frequency of system-generated false alerts as beacons using Location Protocols become more widespread;

- the volume of GEOSAR alert processing varied among the MCCs but was considered to have minimal impact on workload and was justified by the benefits; and

- GEOSAR alert data were effective in resolving LEOSAR location ambiguity, even though manual processes had been used.

BENEFITS OF THE GEOSTATIONARY SYSTEM IN ACTUAL CASES

The saving of additional lives is a critical factor used to assess the benefit of including a geostationary satellite component within the Cospas-Sarsat System. Below are two case reports which clearly highlight this benefit.

Hunters at Hall Beach, Northwest Territories, Canada

On 15 January 1996, two native hunters from Igloolik, set out overland by snowmobile for Repulse Bay with a 406 MHz personal locator beacon (PLB) on loan from the Northwest Territories government. The weather

deteriorated the next day, 16 January to a point where their snowmobile froze and they were stranded in a "whiteout". At approximately 15:00 they activated their beacon. At 15:05, the GOES-8 geostationary satellite relayed the signal and at 15:25, it was processed at the Canadian Mission Control Centre (CMCC). The PLB registry was accessed to identify the owner of the beacon as the Northwest Territories government. Within 10 minutes this information was faxed to the RCMP Operational Communication Centre in Yellowknife, which advised the Igloolik RCMP officer (at 15:55) of the PLB alert in the area. The encoded serial number of the PLB identified it as the one requisitioned by the hunters en route to Repulse Bay. At 16:07 the CMCC provided location co-ordinates, which indicated that the hunters were in the Hall Beach area. The Igloolik SAR Committee transferred SAR responsibility to Hall Beach, and by 21:00, the hunters had been found and rescued.

Fishing Vessel Incident At Sea off the California/Oregon Border, United States

On 1 December 1996, the 40 foot fishing vessel Beach King capsized when hit by a large wave. The vessel was fishing in 20 - 25 foot seas near the California/Oregon border. The crew of three was not able to transmit a mayday call or signal a distress in any way; however, the vessel's 406 MHz EPIRB activated automatically when the vessel capsized. The transmission from the 406 MHz EPIRB was relayed via the GOES-9 satellite to the USMCC and then on to RCC Seattle. RCC Seattle forwarded the information to the Coast Guard Operations Centres in North Bend, Oregon and Humboldt Bay, California. The Operations Centres issued an Urgent Marine Information Broadcast while the RCC called the emergency point of contact listed in the beacon registration database. Upon learning that the Beach King was fishing off the Klamath River, the RCC dispatched a rescue helicopter to search the area. The helicopter located the capsized vessel and hoisted two crew members from the cold (11° Celsius) water. The third crew member was never found. The helicopter transported the crew members to a local hospital where one of the crew required intensive care treatment for hypothermia. Post analysis of the case confirmed that the crew members were not wearing survival suits or flotation devices and would have perished had they remained in the water much longer.

The first notification of distress was an unlocated alert from a geostationary satellite, which arrived at the USMCC 47 minutes before a located alert from a low altitude Earth-orbiting satellite. The Coast Guard was able to investigate the alert (i.e., call the point of contact, check the harbour, make radio call-outs, and launch a helicopter) so that rescue forces could proceed to the vicinity of the distress immediately, even without the located alert data. This resulted in saving the lives of two people.

STUDY CONCLUSION

The D&E confirmed the complementary nature of the GEOSAR system to the Cospas-Sarsat LEOSAR System. It indicated that incorporation of the GEOSAR system as a complement to the Cospas-Sarsat System would generate significant benefits and would save additional lives and property. It was also concluded that the benefits would increase when beacons with encoded Location Protocols come into widespread use.

The D&E highlighted the requirement that the 406 MHz beacon user community needs to be informed of GEOSAR system performance in respect of the Cospas-Sarsat System. Greater care in the use of 406 MHz beacons will be required as inadvertent activations (even of short-term duration) would probably be detected by the GEOSAR system and could set off an unnecessary chain of events with a corresponding use of SAR resources.

RECOMMENDATIONS

As the D&E confirmed the expected benefits of the GEOSAR satellite system as a complement to the Cospas-Sarsat LEOSAR System, it is recommended that:

1. the Cospas-Sarsat Council consider adopting the GEOSAR satellite system as an enhancement and complement to the Cospas-Sarsat System; and

2. the Cospas-Sarsat Council should initiate all actions necessary for this enhancement including the commissioning, as soon as possible, of the existing experimental GEOLUTs.

In order to ensure the optimal technical and operational performance of the enhanced Cospas-Sarsat System, it is recommended that:

3. the Cospas-Sarsat Secretariat should publish and widely distribute the results of the D&E;

4. all manufacturers, administrations and others who develop educational programs and materials should stress the importance of proper handling, shipping, storage and testing of 406 MHz beacons in view of avoiding false alerts;

5. all administrations should establish and maintain complete, accurate, and up-to-date beacon registration databases that are available for SAR agency access on a 24-hour a day basis;

6. Ground Segment Operators should consider the use of GEOSAR detections to resolve LEOSAR location ambiguity, agree on a standardised procedure for inclusion in document C/S A.001 (DDP), and implement software in the MCCs to automate the agreed procedure; and

7. all administrations should review beacon test policies and procedures and revise them as necessary to reflect the incorporation of GEOSAR alerts in the Cospas-Sarsat System.

CONCLUSION

As a result of the D&E, the Cospas-Sarsat Council decided to formally integrate the GEOSAR system into the Cospas-Sarsat System, as soon as possible.

Nextsat—The Challenge of the Next Generation of Aeronautical Satellites

Peter Wood
PETER WOOD ASSOCIATES
9508 Mandolin Court, Vienna, VA. 22182-1610 U.S.A.
e-mail: peterwood@ieee.org

ABSTRACT

For the past nine years satellites in geostationary orbits have provided aeronautical communications services. The technical requirements have been documented in ICAO Standards and Recommended Practices, ARINC Characteristics, and RTCA Minimum Operating Performance Standards. Recently the aeronautical services provided through these satellites have been enhanced through the use of lower data rate voice codecs and spot beam satellites A new generation of aeronautical satellite services is now being offered through constellations of satellites operating in low- and medium-earth orbits, and documentation is being prepared for these services.

INTRODUCTION

Once, the provision of satellite communications for air traffic services was as simple as using a municipal water supply. An international treaty organization (Inmarsat) providing near-global service and operated four geostationary satellites and a number of spares. Support to CAAs was provided by the local Signatory to that organization. A set of ICAO Standards and recommended Practices had been adopted for aeronautical services. Use of the spectrum that had been allocated for aeronautical safety services was controlled by an international organization (ICAO), which permitted its use for non-safety aeronautical communications until the spectrum was required exclusively for the provision of safety communications. Avionics and antennas were available from several organizations. Documents defining the form fit and function of the system and the tests necessary for certification had been approved.

Now the position has become much more complicated. The ITU allows that part of the spectrum originally dedicated to AMS(R) S to be used for all mobile services (with safeguards for AMS(R) S traffic). The original operator of the satellites has become a private company. Enhancements to the Inmarsat system have been introduced, in some cases using proprietary technology. And competing services are now being introduced based upon satellites operating in low- and medium-earth orbits.

A draft report to Working Group F of ICAO's Aeronautical Mobile Communication Panel states: The long development time that has led to the introduction of the current AMS(R)S system also makes it unpractical (*sic*) to envisage any major changes in the current technology before the year 2010.

Is this true? Is the aeronautical community prepared for the new developments that are now taking place?

This paper considers these developments in two different areas: enhancements to the existing system, and completely new systems. In addition to the technical differences, it also discusses the commercial and regulatory implications. The table at the end of this paper outlines the basic characteristics of the systems that are covered..

THE INMARSAT SYSTEM

To date, aeronautical satellite communications have been based on a constellation of four satellites located in geostationary orbit and operated by Inmarsat. In addition to supporting aeronautical communications, the Inmarsat system also provides maritime services (in particular GMDSS—the Global Maritime Distress and Safety Service) and land mobile services. Aeronautical services operate in the band that was designated for AMS(R)S traffic, and were permitted under ICAO rules to support public correspondence (AMSS) in the same frequency band until it was required for aeronautical safety services.

For data services, communications from the aircraft are based upon slotted Aloha for short messages and reservation TDMA for longer ones. In the ground to air direction, communications are based on a one-to-many TDM approach

For voice services, communications use a FDMA approach, with a voice codec digitization rate of 9.6 kb/s.

Because of the limitation of the geostationary orbit, communications with aircraft can only be guaranteed up to about 75° latitude.

SYSTEM ENHANCEMENTS

Two major enhancements have been made to the original system. The first is the introduction of a (proprietary) voice codec operating at a coding rate of 4.8 kb/s. This has the benefit of both enhancing spectral efficiency, and allowing installations to support more voice channels with

a given amount of aircraft power. Tests have demonstrated that the 4.8 kb/s voice codec has a performance as good as (and sometimes better than) the original 9.6 kb/s. codec. Although apparently a simple technological improvement, this innovation has led to significant institutional problems.

The second enhancement directly results from the launching of Inmarsat's third generation of satellites, which include spot beams to concentrate the satellite power on selected regions of the earth's surface. This allows aircraft to be equipped with a lower-gain antenna, and requires reduced avionics power. In addition, the new system, known as Inmarsat-I, also includes the lower data rate codec, referred to above, and so has been subject to the same institutional problems.

NEW GEO SYSTEMS

Some of the new systems, such as Japan's MTSAT, are intended to work directly with the Inmarsat system, permitting "transparent" usage as an aircraft transitions from one system to another. This has proved to be more difficult in practice than in theory, largely due to the need for transmission of a common system table, showing the frequencies used by the services. The institutional steps that commonality involves have proved to be formidable obstacles, but hopefully these can be resolved.

SECOND GENERATION SYSTEMS.

For years, geostationary systems have been the backbone of mobile satellite communications. More recently, a number of satellite systems are being introduced which are based upon a constellation of satellites operating in low-earth orbit (typically 800km.) or medium-earth orbit (typically 10,000km).

A geostationary satellite system can be based upon a minimum of three satellites in nominally fixed position in the equatorial plane (although the Inmarsat system uses four operational satellites). In contrast, systems operating in low (LEO)- or medium- (MEO) orbit comprise many satellites that move over the earth's surface, with the complete constellation ensuring that any point in the coverage area is always in sight of at least one satellite. To date, three such systems are being introduced (see the table). Of these three system, Iridium plans to support both aeronautical safety, AMS(R)S, and public correspondence (AMSS). ICO has stated that it will support public correspondence, but has not announced plans to support AMS(R)S. Globalstar has no plans at present to support aeronautical communications.

Among the advantages claimed for systems operating in LEO or MEO are reduced power requirements and reduced delay time. Addirionally, if the orbit is appropriately chosen, it is possible for these satellites to provide communications over the whole surface of the earth.

THE IRIDIUM SYSTEM

Iridium entered commercial service in November 1998, and plans to begin to support aeronautical communications during 1999. Both the aircraft and ground equipment to support aeronautical communications is being supplied by AlliedSignal. AlliedSignal intends to provide single channel, as well as 5 and 8 multi-channel systems. Currently the single channel system is being tested.

Iridium operates sixty-six satellites in six global planes. All of the satellites are now in orbit, together with a minimum of six spares, one in each orbital plane. The satellites orbit at a height of 780 km., with an orbital inclination of 86.4 degrees. A satellite completes one orbit of the earth in just over 100 minutes. Each satellite can support up to 48 spot beams, although the number of spot beams is reduced over northern and southern latitudes to reduce redundancy between satellites.

One of the unique features of the Iridium system is the use of intersatellite links, in which a message is handed off from satellite to satellite until it is finally transmitted to the desired ground station (or "gateway").

Iridium operates on a TDMA/FDMA basis in a band of frequencies from 1.616 to 1.626.5 GHz. It uses the same frequency for both transmit and receive, and each frequency can support up to four individual calls. This is achieved by dividing each 90 msec time frame into nine slots, with the first slot being reserved for system management, followed by eight 8.28 msec slots, four being devoted to receiving calls, interleaved with four transmitting calls.

THE GLOBALSTAR SYSTEM

Globalstar will being service at the end of 1999. It will comprise forty eight satellites in eight global planes at an altitude of 1,414 km. Because of the inclination of the satellite orbits it will not cover polar regions. Globalstar transmits to the user terminal in the 2.5 GHz band, and the return signal is in the 1.6 GHz band. CDMA is used in both directions. Globalstar is not considered further because it does not plan to support aeronautical communications at this time.

THE ICO SYSTEM

In contrast to the two previously mentioned systems, ICO will operate in medium earth orbit at an altitude of 10,355 km. The system will use ten satellites in two orbital planes, with two spares in orbit. Each satellite can support up to 163 spot beams. The first launch is scheduled for the end of 1999, with the system becoming fully operational in 2000. The inclination of the satellite orbits will result in the polar regions being covered, and because of the high altitude at which the satellite operate, two satellites will be in view for 90% of the time.

ICO will operate on a conventional TDMA basis and, unlike Iridium, will not use intersatellite links. It transmits to the user at 2.1-2.2 GHz, and the return signal is at 1.9-22.0 GHz.

THIRD GENERATION SYSTEMS

Several other systems have been proposed for initial operation in the early 2000s. Most of these (such is Inmarsat's Horizons and the Boeing/Motorola Teledesic) are intended to provide high speed digital services, primarily for use with the Internet. At this time it is not possible to say whether any of these will support aeronautical services.

BENEFITS OF NEW SERVICES

It has to be recognized that the days of a single system designed and operating to one common standard are over. We are faced now with a number of systems, many of which are incompatible, but each of them possesses certain advantages.

Both the enhancements to existing systems and the features of new systems present advantages to the aeronautical community. For example, at least one system offers truly global communications, an important factor now that over-the-pole routes are becoming more common.

Improved spectral efficiency will allow for better use of the spectrum that is available. Spot beam satellites help to reduce the power required from both satellite and aircraft. In addition, the lower path loss that arises with lower earth orbits means that both less satellite and aircraft power is required. For example, the maximum path loss is approximately 25 dB lower for satellites in low earth orbit than for ones in a geostationary orbit. Typical power requirements for the aircraft avionics vary from 1 watt per voice channel for the Iridium system, to 10 watts per voice channel for the basic (Inmarsat-H) system.

Aircraft that are equipped with avionics for handling signals from satellites in low- or intermediate earth orbit will only require a simple non-directional antenna, and no beam steering mechanism will be necessary.

Interest is now being shown in developing a generic set for "Required Communications Performance" that can be customized for specific applications. Some of the requirements can be satisfied by installing separate equipment that will work with different satellite service providers. For example, it has been suggested that fully redundant systems will be required to achieve the standards of availability and continuity that are now being considered. This can be more economically achieved through the use of two differing systems, each based on different constellations of satellites, than by a single fully redundant system supported by only one service provider.

CONSTRAINTS ON NEW SERVICES

While both enhancements to the existing satellite services and completely new offerings present benefits to the aeronautical community, there are many obstacles to be overcome before they can enter into general service. None of these are strictly technical in nature, but they are still formidable.

Standards

The existing system is well documented, in ICAO Standards and Recommended Practices (SARPS) , RTCA Minimum Operating Performance Standards (MOPS), and ARINC Characteristics that describe form, fit and function.

It is now considered that developments that will occur in the future would be better handled by "generic" documents which describe the requirements in general, supplemented by further technical documents each devoted to one specific system. While theoretically this is an approach that will better support new developments, in practice it can lead to further delays. For example, the Minimum Aviation System Performance Standards for Aeronautical Mobile Satellite Systems that is being prepared by RTCA Inc., has been worked on for several years, and is still not complete.

Additionally, there is a reluctance on the part of the aeronautical community to accept the fact that new systems will increasingly be proprietary in nature. The Minimum Operating Performance Standards that described tests in support of the enhancements to the existing system, have been delayed several months because the voice codec used in the enhanced system is proprietary and available from only one manufacturer.

Institutional Issues.

The ideal would be a completely transparent system, in which an aircraft could move from region to region and from satellite operator to satellite operator without the need to make any changes. In practice, this is proving difficult, if not impossible, to achieve.

Even when different satellite systems use the same standards, interoperability has not been achieved as yet. This is partly due to parochial interests that require service to be provided by the national satellite service provider. Also, the institutional arrangements necessary to support a transparent operation, such as the transmission of a common system table, are difficult to put into place.

A country might impose requirements that must be satisfied before they will license a system for service. One country has approved a license on the basis that services will not be supported in areas that are held by rebel troops. Similarly, restrictions on satellite usage may be imposed to

protect radio astronomy. The impact of these restrictions on aeronautical safety traffic need to be considered before a satellite provider is selected.

Another factor that needs to be taken into consideration is that satellite services are now being provided by public for-profit companies and the financial stability of these organizations is important. Originally, service was provided by an international treaty organization, Inmarsat. Inmarsat will be privatized in mid-April 1999, and will henceforth be subject to its shareholders rather than the governments that originally controlled the organization. A separate body, the International Mobile Satellite Organization has been established to oversee the Public interest services provided by Inmarsat, but this is restricted to maritime (GMDSS) and does not cover aeronautical safety services.

Economic Issues

While competition between service providers should result in reduced costs of service, ultimately the decision on which system to select will be based upon the return on investment involved, taking into account the investments that have already been made. This is affected by analyses which have shown that the major return on investment by airlines will result from improvements in aircraft operating efficiency, rather than the revenue generated from passenger correspondence.

There is little benefit to replacing existing equipment since the savings in the operating costs of the satellite communications system are unlikely to offset the capital costs of replacing the existing systems. Many widebodied aircraft have already been equipped with satellite communication systems. It is likely, therefore, that second generation systems will initially find their primary market in narrow-body aircraft and general (particularly corporate) aviation. The reduced physical size is particularly advantageous in this market segment. It has been mentioned previously, however, that benefits may result from installing dissimilar systems if the highest degree of availability is required.

SPECTRUM

It is impossible to ignore the impact that the decisions that were made at the 1997 International Telecommunications Union World Radio Conference (WARC-97) have made upon aeronautical satellite communications. In effect, as a result of decisions made at that conference, there is no longer part of the spectrum reserved for aeronautical safety communications. That part of the spectrum previously allocated to AMS(R)S services has now been given a generic mobile allocation. Equally importantly, the protection given to AMS(R)S services in terms of preemption and priority only apply to those satellite service providers who are supporting such a service.

It is highly unlikely that real-time frequency management between satellite systems will be introduced in the near term. Satellite systems will continue to operate in dedicated frequency bands, established through bilateral or multilateral negotiations. For this reason it is essential that all aeronautical systems make the most efficient use of the spectrum that is available; for example, by adopting low data rate voice codecs.

CONCLUSIONS

Satellite systems are now being introduced that could result in significant benefits to the aeronautical community, and that would help to offset some of the adverse impacts resulting from the decisions reached at WARC '97. Significant changes to existing procedures will be required if the industry is to take advantage of these new developments.

Much of the support in the development of standards has been provided by the satellite service operators and by equipment manufacturers. Continuation of this support is essential if any standards are to be prepared in a timely manner. At the same time, procedures must be developed that can accommodate the rapid developments in new technology without which many of the claimed advantages cannot be enjoyed

MOBILE SATELLITE SYSTEMS

SYSTEM	COVERAGE	ORBIT	SATELLITES	AMS(R)S	AMSS	STATUS
Inmarsat (global)	Global except polar	GEO	3	YES	YES	Operational
Inmarsat (spot beam)	Land masses and main oceanic routes	GEO	3	YES	YES	Operational
AMSC/TMI	North America	GEO	2	NO	YES	Operational
MTSAT	Japan	GEO	1	YES	YES	Operational 1999
Globalstar	Global except polar	LEO	48	NO	NO	Operational 2000
ICO	Global	MEO	10	NO	YES	Operational 2000
Iridium	Global	LEO	66	YES	YES	Operational 1998 Aeronautical 1999

Note: Inmarsat, Globalstar, ICO and Iridium also have/will have spare satellites in orbit

Doppler Prediction Scheme for User Terminals in LEO Mobile Satellite Communications

Moon Hee You and Soo In Lee
ETRI – Radio & Broadcasting Technology Laboratory
Yusong P.O. Box 106
Taejon, 305-600, Korea
Email: moon@etri.re.kr

ABSTRACT

In low earth orbit (LEO) satellite communication systems, more severe phase distortion due to the Doppler shift is appeared at the received signal than in the cases of geo-stationary earth orbit (GEO) satellite systems or terrestrial mobile systems. Therefore an exact estimation of Doppler shift would be one of the most important factors to enhance performance of LEO satellite communication system. In this paper, we propose a new Doppler prediction scheme by using location information of a user terminal and a satellite. We have simulated the prediction performance of the proposed scheme by using two LEO satellite constellations. The prediction performance compared to Ali's method[1] showed about 5 ~ 20% reductions in average estimation error depending on the satellite orbit characteristics. The proposed scheme needs higher calculation loads but the Doppler prediction error range which has to be covered by frequency synchronization curcuit is smaller.

INTRODUCTION

In low earth orbit (LEO) satellite communication systems, more severe phase distortion is appeared at the received signal than in the case of geo-stationary earth orbit (GEO) satellite systems or terrestrial mobile systems[2][3]. That is the result of the Doppler frequency shift representing on the receiving signal, which comes from the faster movement of a satellite relative to that of the terminal or earth station. Since the relative velocity of the satellite to the surface of the earth is very fast, the Doppler frequency shift affects to carrier frequencies largly in the communication links and the time variation of the Doppler frequency shift also becomes very large. Its effect on communication links makes the performance of the receiver to be degraded, where the amount of performance degradation depends on the transmission schemes used in the communication systems. Usually the Doppler frequency shift is more harmful to the digital communication systems employing coherent demodulators. Therefore there have been many researches preformed to compensate for the Doppler shift[4][5][6]. Recently, a Doppler estimation scheme by using a relative time information is introduced[1], which is available to pre-compensate the Doppler frequency shift before carrier recovery.

The cyclic movement of a LEO satellite is represented by deterministic formula with its period. Therefore, if the movement of the terminal is ignored, the Doppler shift, which is in proportion to the relative velocity between a LEO satellite and a terminal, has also its deterministic function, and can be represented in time domain[7]. Using this in feeder links, the Doppler shift is easily predicted in fixed earth station because the earth station knows the exact information about its own position and the satellite's time-varing location. But in the case of mobile terminals in user links, they may not know their own positions on time when they are required for call.

In this paper, we propose a scheme that is able to estimate a user terminal's position and predict its contineous Doppler frequency shift simultaniously. First of all, the Doppler frequency shift is described in a closed form with geographical parameters. We also briefly describe the Doppler estimation scheme proposed by Ali[1], and the prediction performance will be compared to that of our method. In our prediction method, Location information of a user terminal and a satellite is used. We have simulated the prediction performance of the proposed scheme by using two LEO satellite constellations. The simulation result will demonstrate the enhanced estimation performance of the proposed scheme. Finally, we draw conclusion.

DOPPLER SHIFT CHARACTERISTICS

The Doppler shift, $f_D(t)$ for LEO mobile satellite communication system can be expressed as the following eq. (1) with relative velocity between the satellite and the terminal.

$$f_D(t) = -\frac{f_c}{c}\cdot((\mathbf{v}_s(t)-\mathbf{v}_e(t))\cdot\mathbf{u}) = -\frac{f_c}{c}\cdot\frac{ds(t)}{dt} \qquad (1)$$

where f_c is the transmitted carrier frequency, c is light velocity ($\approx 3*10^5$ km/s), $\mathbf{v}_s$ is a vector velocity of the satellite, $\mathbf{v}_e$ is a vector velocity of the terminal, $\mathbf{u}$ is a unit direction vector from the terminal to the satellite, and $s(t)$

is the distance between the satellite and the terminal as shown in eq. (2).

$$s(t)=\left\{a^2+r^2-2ar\cos\psi(t)\right\}^{1/2} \quad (2)$$

where a is the radius of the satellite orbit and r is the radius of the earth. $\psi(t)$ is the angle between the satellite and the user terminal as shown in Fig.2. Estimation of $\psi(t)$ is a key to calculate the Doppler shift, and it can be expressed in various ways, that is it can be expressed with various parameters such as time and/or location information of the satellite and the user terminal. Therefore it is very important to define $\psi(t)$ with parameters which can be estimated easily or can be assumed with a small amount of error.

One of the classical way of expressing cos $\psi(t)$ can be found in [8], and it is shown in eq. (3). It is characterized in the earth centered inertial coordinate frame by using satellite's orbit angle, $\theta_S(t)$ and the time-varying longitude angle of the user, $\theta_e(t)$. In this frame, the angular velocity of the satellite, ω_s and the angular velocity of the terminal (exactly means of the earth), ω_e are constant, eg. 2π/(orbit period) and 2π /(23 h 56 min), respectively.

$$\cos\psi(t)=\cos T_e\cos\theta_s(t)\cos\theta_e(t) + \cos i\cos T_e\sin\theta_s(t)\sin\theta_e(t)+\sin i\sin T_e\sin\theta_s(t) \quad (3)$$

where T_e is the latitude of the terminal and i is the satellite's orbit inclination angle.

Ali represented cos $\psi(t)$ as eq. (4) using the time, t_m when the satellite makes maximum elevation angle with the user terminal in the earth centered fixed coordinate frame. In this frame, the angular velocity of the satellite, ω varies with latitude due to earth's rotation.

$$\cos\psi(t)=\cos(\theta(t)-\theta(t_m))\cdot\cos\psi(t_m) \quad (4)$$

where $\theta(t)$ is the satellite angle measured on the surface of the earth along the ground trace from the satellite ascending node.

In this paper, we represent cos $\psi(t)$ as eq. (5) using the user terminal's position, eg. the latitude, T_e and the longitude, G_e, which are practically more applicable parameters than $\theta_S(t)$, $\theta_e(t)$ or $\psi(t_m)$, $\theta(t_m)$. It is also expressed in the earth centered fixed coordinate frame.

$$\cos\psi(t)=\cos T_s(t)\cos T_e\cos(G_s(t)-G_e)+\sin T_s(t)\sin T_e \quad (5)$$

where $G_s(t)$ is the satellite's longitude, and we assume that the user terminal can easily obtain the information of satellite ephemerides at time t.

The Doppler frequncy shift can be expressed as eq. (6) by using eq. (5).

DOPPLER PREDICTION SCHEMES

Now, it is very clear that selection of a certain expression for $\psi(t)$ would determin the required parameters for Doppler calculation. In [1], Doppler shift is estimated by using the relative time to the reference time when the satellite makes maximum elevation angle with the user terminal.

The angular velocity of the satellite in the earth centered fixed frame, $\omega(t)$ is approximated to eq. (7), then Doppler shift $f_D(t)$ can be written as eq. (8). The variables t_m and cos $\psi(t_m)$ required in eq. (8) are obtained as shown in eq. (9) and eq. (10) by using the measured Doppler frequency shifts, $f_D(t_0)$ and $f_D(t_1)$, and Doppler shift rates, $f_D(t_0)$ and $f_D(t_1)$ at sampling instants, t_0 and t_1 ($> t_0$), respectively.

$$\omega(t)\approx\omega_s-\omega_e\cos i \quad (7)$$

$$f_D(t)=-\frac{f_c}{c}\frac{ar\sin(\omega(t-t_m))\cos\psi(t_m)\cdot\omega}{\sqrt{a^2+r^2-2ar\cos(\omega(t-t_m))\cos\psi(t_m)}} \quad (8)$$

$$t_m=t_1-\frac{\alpha(t_1)}{\omega} \quad (9)$$

$$\cos\psi(t_m)=-c^2\frac{f_D^2(t_1)}{arf_c^2\omega^2}\left(\frac{\sin\alpha(t_1)\dot{f}_D(t_1)}{\omega f_D(t_1)}-\cos\alpha(t_1)\right)^{-1} \quad (10)$$

where $\alpha(t_1)$ is in eq. (11).

$$f_D(t)=-\frac{arf_c\left\{\left[-\omega_s\sin i\cos\left[\sin^{-1}\left(\frac{\sin T_s(t)}{\sin i}\right)\right]\tan T_s(t)\cos(G_s(t)-G_e)-\left(\frac{\omega_s\cos i}{\cos T_s(t)}-\omega_e\cos T_s(t)\right)\sin(G_s(t)-G_e)\right]\cos T_e+(\omega_s\sin i\cos\theta_s(t))\sin T_e\right\}}{c\sqrt{a^2+r^2-2ar(\cos T_s(t)\cos T_e\cos(G_s(t)-G_e)+\sin T_s(t)\sin T_e)}} \quad (6)$$

$$\alpha(t_1)=\theta(t_1)-\theta(t_m)=\operatorname{atan}\left(\frac{\omega f_D(t_1)f_D^3(t_2)+f_D^3(t_1)\dot{f}_D(t_2)\sin(\omega(t_2-t_1))-\omega f_D^3(t_1)f_D(t_2)\cos(\omega(t_2-t_1))}{\dot{f}_D(t_1)f_D^3(t_2)-f_D^3(t_1)\dot{f}_D(t_2)\cos(\omega(t_2-t_1))-\omega f_D^3(t_1)f_D(t_2)\sin(\omega(t_2-t_1))}\right) \quad (11)$$

On the other hand, in our estimation scheme, the latitude and longitude information of the user terminal's position are derived as follows in order to estimate the Doppler shift at any time t by using the measured Doppler shift, $f_D(t_0)$, $f_D(t_1)$ and $f_D(t_2)$ and Doppler shift rate, $\dot{f}_D(t_0)$, $\dot{f}_D(t_1)$ and $\dot{f}_D(t_2)$ at the initial sampling instants t_0, t_1 and t_2, respectively.

The distance function, $s(t)$, between satellite and user terminal is represented as eq. (12) by using the derivative of cos $\psi(t)$ and the Doppler shift rate, $\dot{f}_D(t)$, which is the derivative of Doppler shift in eq.(1).

$$\frac{d^2}{dt^2}\cos\psi(t)=-\left(\frac{d\theta(t)}{dt}\right)^2\cdot\cos\psi(t) \approx -(\omega_s-\omega_e\cos i)^2\cdot\frac{a^2+r^2-s^2(t)}{2ar} \quad (12)$$

$$s^2(t)-\frac{2c\dot{f}_D(t)}{\omega^2 f_c}s(t)-\left[\left(a^2+r^2\right)-2\left(\frac{cf_D(t)}{\omega f_c}\right)^2\right]=0 \quad (13)$$

We select the adequate results, $s(t_0)$ and $s(t_1)$, based on physical considerations from the solutions of the 2nd order function in eq.(12) at the initial sampling instants t_0 and t_1, respectively, by using the geomatric relation of a satellite and a user terminal. Using $s(t_0)$ and $s(t_1)$, we can obtain cos $\psi(t_0)$ and cos $\psi(t_1)$ and also the user terminal's location.

$$\sin G_e=\frac{(A_2D_1-A_1D_2)-(A_2C_1-A_1C_2)\sin T_e}{(A_2B_1-A_1B_2)\cos T_e} \quad (14)$$

$$\cos G_e=\frac{(B_2D_1-B_1D_2)-(B_2C_1-B_1C_2)\sin T_e}{(A_1B_2-A_2B_1)\cos T_e} \quad (15)$$

$$\sin T_e=\frac{k_2\pm\sqrt{k_2^2-k_1k_3}}{k_1} \quad (16)$$

where the variables are as follows.

$$A_n=\cos T_s(t_n)\cos G_s(t_n)$$
$$B_n=\cos T_s(t_n)\sin G_s(t_n)$$
$$C_n=\sin T_s(t_n)$$
$$D_n=\cos\psi(t_n)$$

$$k_1=(A_2C_1-A_1C_2)^2+(B_2C_1-B_1C_2)^2+(A_1B_2-A_2B_1)^2$$
$$k_2=(A_2C_1-A_1C_2)(A_2D_1-A_1D_2)+(B_2C_1-B_1C_2)(B_2D_1-B_1D_2)$$
$$k_3=(A_2D_1-A_1D_2)^2+(B_2D_1-B_1D_2)^2-(A_1B_2-A_2B_1)^2$$

After estimating user terminal's location, the Doppler shift of the next sampling time is predicted by using eq. (6).

SIMULATION RESULTS

We simulated the performance of the proposed prediction scheme and compared to Ali's method. The LEO satellite systems being considered are of two types, one with a polar orbit and the other with a inclined orbit, and each system has the parameters shown at Table 1. Each satellite was simulated for 500 times rotation around the earth. We assumed that a user terminal is located at longitude E20° and latitude N30°, the satellite passes over the user position 129 times in the case of LEO (A) and 264 times in the case of LEO (B) as shown in Fig.2 and Fig.3, respectively.

Simulation results in Table 2 show that the prediction error resulted from the proposed method is about ±0.4 kHz for LEO (A) system and about ±1.3 kHz for LEO (B) system. Compared to Ali's method, the proposed method produced about 5% and 20% of reduction in prediction error for LEO (A) system and LEO (B) system, respectively.

In Fig. 4 and Fig. 5, the probabilty density fuctions of Doppler prediction error for both methods are shown. It is clearly shown that the proposed method has better performance than that of Ali's prediction scheme.

CONCLUSION

We have proposed a new Doppler prediction scheme and evaluated its performance, which is compared to the performance of Ali's Doppler prediction scheme. For the performance evaluation, we have simulated two schemes not only for the LEO satellite with a polar orbit but also for the LEO satellite with an inclined orbit. The prediction performance compared to Ali's method showed about 5 ~ 20% reductions in average estimation error depending on the satellite orbit characteristics. Although the proposed scheme needs a little bit more calculation loads, the Doppler prediction error range that has to be covered by frequency synchronization curcuit is much smaller. Therefore this proposed scheme makes frequency acquisition and tracking step simple.

In future, we may have to carry out a trade-off analysis between the accuracy of the prediction schemes and the computational efficiency, and H/W complexity of synchronization circuit.

REFERENCES

[1] I. Ali, N. Al-Dhahir and J. Hershey , "Doppler Characterization for LEO Satellites", *IEEE Trans. Commun.*, Vol. 46, No. 3, pp.309-313, Mar. 1998.

[2] M. J. Miller, B. Vucetic and B. Les, *Satellite Communications : Mobile and Fixed Services*, Kluwer Academic Publishers, 1993.

[3] S. Ohmori, H.Wakana and S. Kawase, *Mobile Satellite Communications*, Artech House, 1998.

[4] E. Vilar and J. Austin, "Analysis and Correction Techniques of Doppler Shift for Nongeosynchronous Communication Satellites", *Int. J. Satellite Commun.*, Vol. 9, pp.123-136, 1991.

[5] A. Kajiwara, "Mobile Satellite CDMA System Robust to Doppler Shift", *IEEE Trans. Veh. Technol.*, Vol. 44, pp.480-486, 1995.

[6] Y. Uno *et al*, "Carrier Regenaration in a Block Demodulator for Low Earth-Orbital Satellite Communication Systems", *Proc. the 4th Int. Symp. Personal, Indoor, and Mobile Radio Commun.*, pp.458-462, 1993.

[7] M. H. You and S. I. Lee, "The Characteristics of Doppler Shift over NGSO Satellite Communication Links", *Proc. IEEK Summer Conference '98,* Vol. 21, No. 1, pp 26 – 29, Jun. 1998.

[8] ITU-R, "Recommendation ITU-R M.1225", App.2 to Annex 2, 1997.

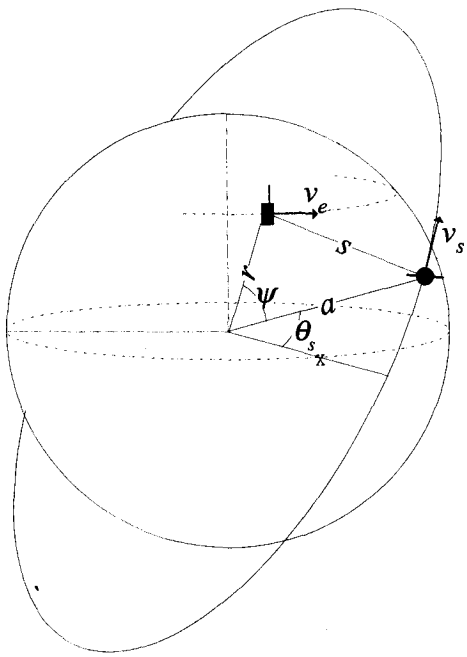


Fig.1 Satellite geometry in the earth centered inertial coordinate frame

Table 1. Simulation parameters

	LEO (A)	*LEO (B)*
Altitude (km)	780	1414
Orbit inclination (deg.)	86.4	52
Min. elevation (deg.)	8.3	10
Carrier frequency (GHz)	1.6	2.4
No. of simulated ground paths for the satellite	129	264

Table 2. Doppler prediction error statistics

Prediction error	*LEO (A)*		*LEO (B)*	
(Hz)	*#1*	*#2*	*#1*	*#2*
Max.	419	20834	1270	5528
Min.	-460	-20790	-448	-5494
Mean	192	3650	332	1610

Note) #1 : Proposed scheme
#2 : Ali's scheme

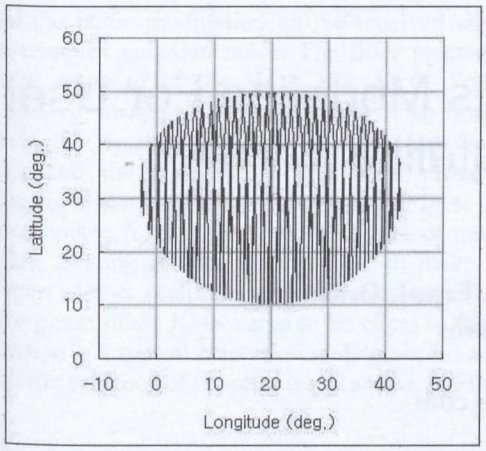


Fig.2 Ground trace of LEO (A) satellite

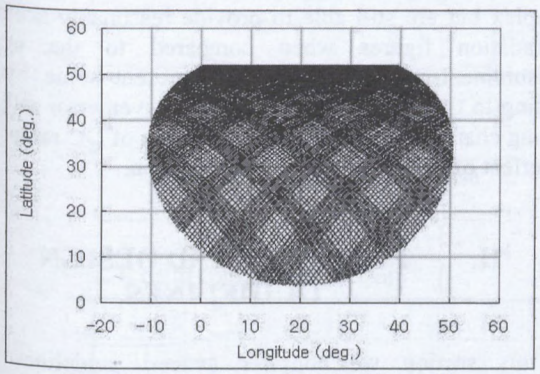


Fig.3 Ground trace of LEO (B) satellite

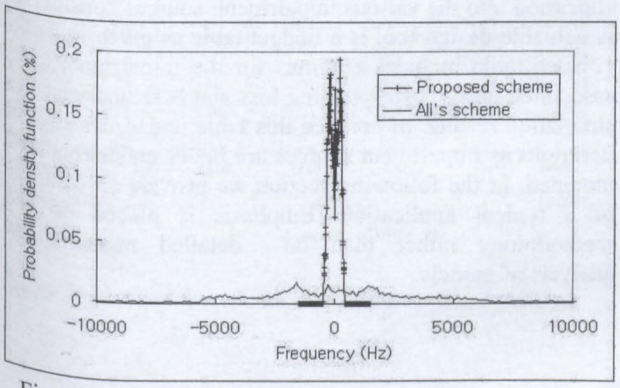


Fig.4 Doppler prediction error probability for LEO (A)

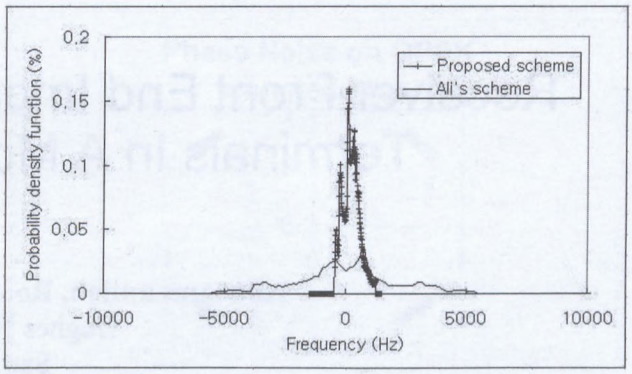


Fig.5 Doppler prediction error probability for LEO (B)

Receiver Front End Impairments Modeling For User Terminals In A Mobile Satellite System

Abu Amanullah, Romeo Velarde, Frank Onochie
Hughes Network Systems
San Diego, CA
Email: aamanullah@hns.com

ABSTRACT

In a satellite communication system, various impairments distort signals as they traverse the gateway, satellite, and terminal communication paths. The design of a good user terminal (UT) require that the impairments in the receiver front end of the UT be modeled correctly. Realistic equivalent model for these impairments are necessary to determine the degradation in SNR caused by such impairments. Due to small link margin in satellite communication, these degradations can have a serious impact on the BER. In this paper, we present modeling of several possible impairments in the UT front end in the context of an end to end simulation system. Planning and design guidelines are given with a view to system validation. We also present results on the degradation in SNR caused by these impairments in Gaussian and Fading Channels.

I. INTRODUCTION

One of the challenging task in the design of a User Terminal (UT) for a Mobile Satellite System (MSS) is to make sure that the UT can function with very small received signal levels. The low levels and the consequent small link margin is due to the distances from the UT to the satellites (longest for Geo-synchronous systems) and the fact power at the satellites is very limited.

The signal from the satellite to the UT is distorted by various impairments whose sources are distributed in both the transmitter and the receiver. The first step in dealing with these impairments is to understand how they contribute to the degradation of a performance measure such as bit Error Rate (BER). In this paper we specifically look at the impairments at the UT receiver front end. These include phase noise, IQ mismatch and DC drift, and fixed point implementation loss. The outline of the paper is as follows. Section II gives some design guidelines. In section III, models of these impairments will be presented along with some typical values in the state of the art receivers. The goal is to present models that are not very complex but are still able to provide reasonably accurate degradation figures when compared to the actual measurements. In section IV, we present some results relating to the degradation in SNR for given error rates in Fading channels. In section V, estimation of DC ramp and the effect of correction algorithm are given.

II. PLANNING AND DESIGN GUIDELINES

In this section we consider general guidelines for considering the effects of impairments in the design of user terminal (UT) segment of a communications system. Given an overall segment budget, the starting point is to address the question of whether this budget is realistic. Next, consideration has to be given to the breakdown or allocation into the various impairment sources considered. A valuable design tool is a budget table as given in Table 1. Such table includes columns for the impairment type, associated units, corresponding loss and both analytical an simulation results. In practice this table undergoes several iterations as impairment sources are better understood and modeled. In the following section we provide an example of a typical application. Emphasis is placed on the methodology rather than on detailed mathematical analysis of models.

III. IMPAIRMENT MODEL

Phase Noise

Phase noise arises in a system due to oscillator instability and thermal noise. However, the magnitude and the shaping of the noise is affected by devices such as filtering and the phase-tracking structure such as the PLL. A widely used model [1] of phase noise is given in Figure 1. It

involves the phase modulation on the received signal by filtered complex gaussian noise. The filter response is a composite response of the VCTCXO, VCO, PLL, and other passive elements in the circuit. Two important factors in the description of phase noise are the noise spectrum and the RMS jitter. A typical phase noise spectrum for L-band UT receiver is given in Figure 2. This can be computed from the component values or measured in the lab. Scaling is done on the filter to make it unit power gain. Other scaling is required to adjust the phase noise (or phase jitter) RMS value to be equal to the given specification . A typical constellation diagram for a QPSK signal in the presence of phase noise is shown in Figure 3.

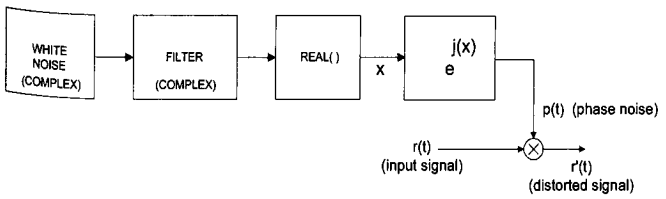


Figure 1: Model of Phase Noise

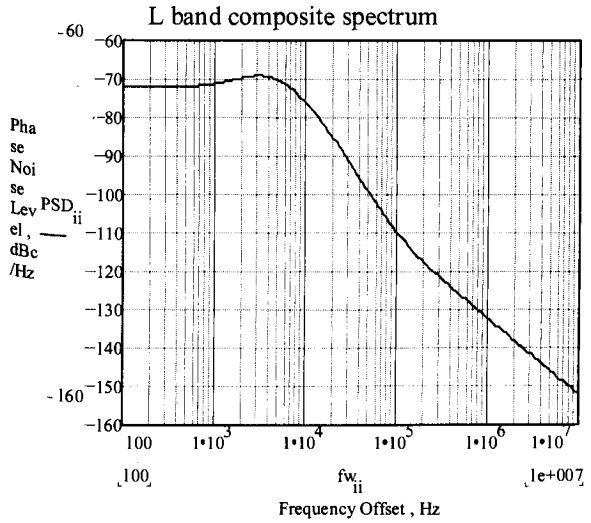


Figure 2: L band Composite spectrum

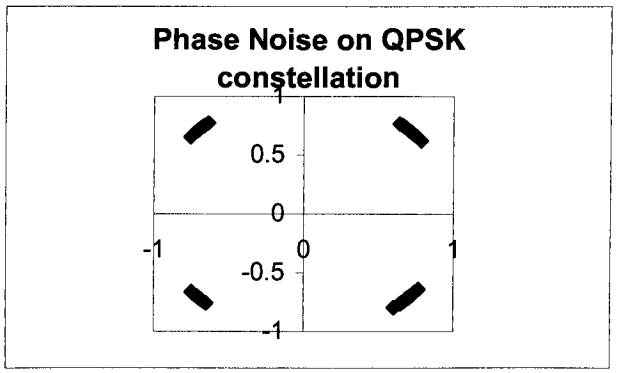


Figure 3: Constellation diagram for QPSK in the presence of phase noise.

IQ Mismatch or Imbalance

For a typical heterodyne receiver with I/Q_phase demodulation on a chip, the gain on the I and Q signals are not exactly the same. In addition, the I and Q signal may not be exactly quadrature, i.e. the angle between the I and Q signals is not exactly 90 degrees. The IQ mismatch may be modeled as in Fig. 4 where the I and Q signal are projected on to two orthogonal axes and appropriate scaling is applied. For an amplitude mismatch of k dB, the scaling on the I signal is by $10^{k/20}$. For a phase angle of (90-θ) degrees between I and Q signal, the scaling on the I signal is by cos(θ). The modified Q component is obtained by adding the projection of I signal on the Q axis to the Q signal.

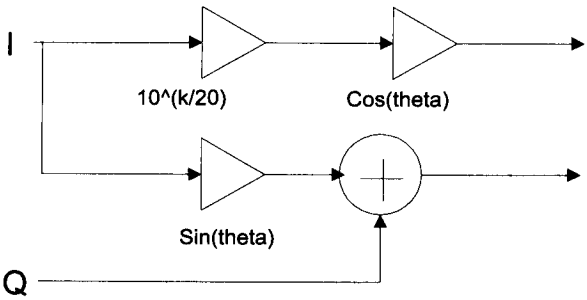


Figure 4: Model of IQ mismatch

DC Offset (Drift)

For certain heterodyne receiver with I/Q_phase demodulation on a chip, a differential DC offset exists a the demodulator's output which is corrected periodically in the chip. Between correction however, this DC drift in the I and Q signal voltage can have severe impact on the signal quality especially at signal levels that are comparable to drift value. The model for the DC offset given in figure 5 contains a ramp voltage added to the baseband I and Q samples.

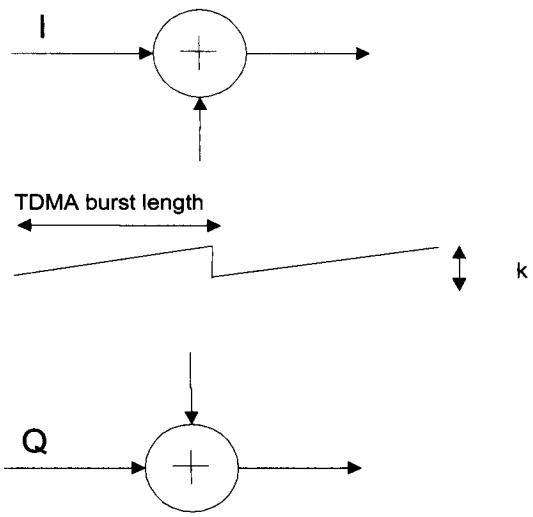


Figure 5: Model of DC Offset

The period of the ramp is equal to the length of the transmitted bursts in a TDMA system since during the DC offset correction in the chip, no demodulation of receive signal takes place. Typical values for DC offset or drift in state-of-the-art receiver is 1mV/ms .

Fixed Point Implementation Loss

The use of 16-bit fixed-point digital signal processors (DSPs) is widespread in the implementation of baseband receivers. Losses due to fixed point algorithms have to be considered. These values can be easily obtained by performance comparison with a reference floating point implementation.

IV. RESULTS

In this section, we present some results from simulation of a QPSK system with models of impairment given in the previous section. The overall system consists of a modulator, transmit shaping filter, channel model, receive filter and a demodulator. The impairments are added after the channel model at the input to the receiver. Figure 6 shows degradation in SNR for a given BER for typical values of IQ mismatch, DC offset, and Phase noise. The results are for phase noise of 2.8 degrees RMS, IQ mismatch of 1.7 dB in amplitude, and 3 degrees in phase, and DC drift of 2mV/ms. The signal level is 6 dB below full scale which corresponds to about .7 V peak. Figure 7 shows the effect of DC offset when the input signal level is low, e.g. 36 dB below full scale.

A typical budget table is given in Table 1. This table is obtained by a consideration of each impairment in isolation. This is a simplistic approach since in practice there is a degree of interaction among the sources. The initial entry for each impairment source can be obtained by intelligent guesswork. As much as practicable analytical results should be derived for increased confidence in the models. A positive margin may be obtained after the initial iteration. However, this margin changes as a result of further model refinement and laboratory measurements. An exit criterion is for example the achievement of a reasonable degree of stability in the table between iterations. It is important to record the corresponding channel conditions.

Source	Loss (dB)	Channel Conditions	
		Analytical	Simulated
Phase noise	X_1	√	√
I/Q mismatch	X_2		√
Fixed-point implementation	X_3		√
DC offset (Before Compensation)	> 3		√
DC offset (After Compensation)	X_4		√
Total	ΣX_i		
Overall Budget (Specification)	X_B		
Margin	X_B-ΣX_i		

Table 1: Error budget table used in design.

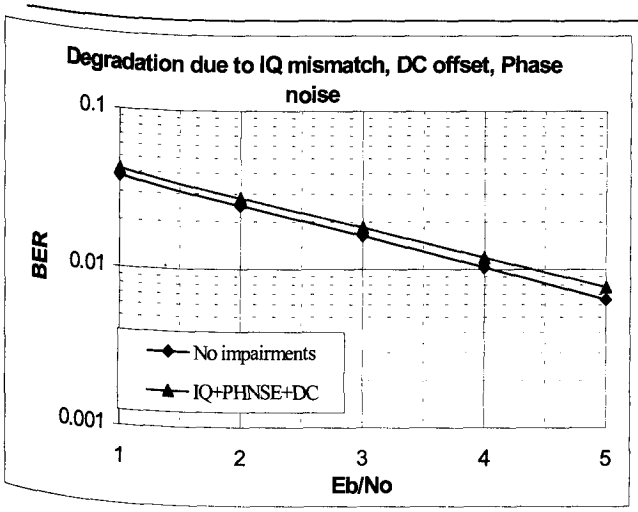


Figure 5: BER vs. SNR with signal level 6 dB below FS

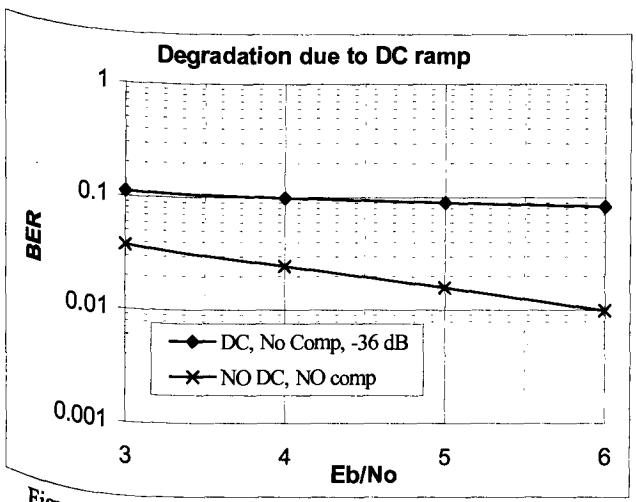


Figure 6: BER vs. SNR with signal level 36 dB below FS No DC Offset Compensation

V. DC OFFSET COMPENSATION

As seen from the results and the budget table in the previous section, the most severe impact in the performance of the receiver is from the DC drift in the I and Q signal between compensation. For I and Q signals that are 36 dB below full scale and comparable to the DC drift, upto 4 dB of degradation in SNR is possible. This stems from the fact that the phase of the received symbol can be completely altered due to the DC drift. This requires that for any practical TDMA burst size, the DC drift has to be estimated at the receiver and compensated for. Figure 7 shows the BER with DC offset after compensation. The estimation and compensation is done in the digital domain with a DSP. Simulation shows that in a fading channel and in the presence of noise, averaging over bursts is sufficient.

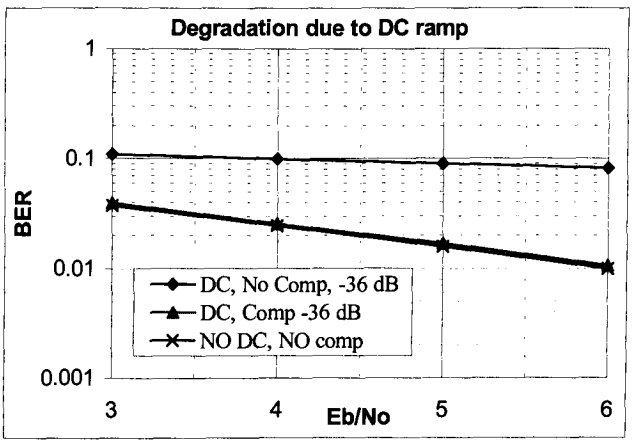


Figure 7: BER vs. SNR with signal level 36 dB below FS DC Offset Compensation.

VI. CONCLUSION

The loss budget table is a good planning and design tool. Although simplistic in nature, it yields tractable results which can lead to good system design and performance. This approach also lends itself for continuous reevaluation of the overall segment loss specification.

REFERENCES

[1] M. C. Jeruchim, P. Balaban, K. Shanmugan, Simulation of Communication systems, New York: Plenum Press, 1992., ch. 4.

Development and Trials of an Inmarsat STD D+ Mobile Satellite Terminal

D. Reveler, A. Parolin, J.Ovtsyn, L.Tibbo, P. Rossiter, G. Eaves
SkyWave Mobile Communications Inc.,
3701 Carling Ave, PO Box 11490, Station "H", Ottawa, Ontario K2H 8S2, Canada.
peter.rossiter@skywavemobile.com

ABSTRACT

This paper describes the development and field trials of SkyWave's new low-cost DMR-200 Inmarsat Standard D+ mobile satellite terminal.

Figure 1 DMR-200 Terminal

The DMR-200 is a low-speed, two way messaging terminal designed to operate within Inmarsat's L-band global mobile satellite communications system. Typical Applications for the DMR-200 include asset tracking, SCADA, covert mobile applications, and low-power untethered tracking.

This paper includes a brief overview of the D+ system, specifications and measured performance characteristics of the DMR-200, and a description of a field trial conducted between Canada and Europe earlier this year.

BACKGROUND

While mobile satellite terminals have been used for at least 10 years for vehicle tracking applications they have, until recently, been limited to use on large highway transports or ships because of their size and cost.

With new integrated circuit devices and other technologies emerging from the wireless and PCS industries it is now possible to design mobile satellite terminals, such as the DMR-200, weighing less than a pound (450g) and costing less than $1000. Such a terminals open up possibilities for high volume applications because they fit well within the size and power constraints of common smaller platforms such as vans, cars, pleasure boats or even motor cycles.

Recognizing this, Inmarsat defined the D+ system in the mid-nineties and encouraged the development of terminals and terrestrial Hub infrastructure.

As of 1998 the D+ system is operational and service is provided to both the Atlantic and Indian Ocean regions via Station 12's LES in Burum, the Netherlands. The DMR-200 terminal is fully Type Approved by Inmarsat, and is in production at SkyWave.

INMARSAT D+SYSTEM OVERVIEW

The D+ system is intended to support short messaging between large numbers of mobile terminals and terrestrial stations. For example, a GPS position report from a mobile can be delivered as an 8-byte message.
Table 1 shows the primary system design parameters.

Forward Link	**Specification**
Modulation	32-ary FSK
Coding	Reed Solomon (31,15)
Symbol Rate	20 sps
User Data Rate	8.75 bps
Frame Duration	2 or 4 minutes
Receive G/T	-25.1 dB/K
Return Link	**Specification**
Modulation	Binary FSK
Coding	R ½, Convolutional
Symbol Rate	4, 16, 32, 64 or 128 sps
User Data Rate	2, 8, 16, 32 or 64 bps
Transmit EIRP	0 dBWic minimum
Message types	Long Data Burst: 64 bits Short Data Burst: 16 or 23 bits Acknowlegment: 1 bit

Table 1 Primary D+ System Design Parameters

TERMINAL SPECIFICATIONS

The DMR-200 consists of a single module housed in a polycarbonate shell and includes the transceiver, antenna, GPS receiver, controller and power conditioning.

The user interface is via an RS232 port. This port is used for both user I/O messages, and for programming the terminal. A full API is supported. Fig 1 shows the DMR-200 terminal. Figure 2 shows a block diagram of the terminal.

Table 2 shows the electrical and mechanical specifications.

Parameter	Specification
Receive Band (MHz)	1525 to 1559
Transmit Band (MHz)	1626.5 to 1660.5
Channel Spacing (KHz)	2.5
GPS Receiver	Integral, 12-channel
User Interface	RS-232, 9600bps
Size	121mm x 121mm x 41mm
Weight	425 g
Supply Voltage	8-32vdc
Power Consumption	
Receive	1.2 W
Idle	0.5W
Transmit	10 W
GPS	1 W
Sleep	500 micro A
Environmental (operating)	-40 to +70 C

Table 2 DMR-200 Specifications

TERMINAL DESIGN APPROACH

Figure 2 shows a block diagram of the terminal.
The design approach is a relatively conventional one with multiple stages of up and down conversion. Transmit and receive frequencies are synthesized in 1 Hz steps across the full transmit and receive band. Great attention was paid to the design of the frequency plan to ensure spurious emissions met the stringent requirements of ETSI specification 300-254.

The system noise temperature is under 300K and the SSPA, is rated at 2W, and operates in a low duty cycle mode to support the time-slotted return link access.

A 12 channel GPS receiver is integrated into the terminal to support the position reporting capability. Time to first fix is under one minute.

The Antenna is a broad-band, circularly-polarized microstrip patch which is designed cover the full transmit and receive bands together with the GPS frequency at 1575 MHz.

The demodulator/decoder is implemented in a proprietary DSP design and provides excellent performance down to a threshold C/No of 16 dB-Hz.

The power conditioning subsystem accepts a wide range of input voltages (8 to 32 vdc) to ensure ease of operation in most environments.

Terminal control functions, message formatting and user programming are implemented in a 16-bit microcontroller.

PROGRAMMING

Operating Modes

The DMR-200 operates in one of two modes - Slave mode or Automonous mode. In Slave mode the DMR-200 functions as an Inmarsat D+ satellite modem under the control of an external controller to support a flexible range of applications.

In Autonomous mode the DMR-200 operates as a standalone unit with no external connections except for power. This mode is designed for low-cost asset tracking applications. Its capabilities are described in detail in the next section.

Autonomous Mode Operation: In order to provide a terminal that can function effectively for different requirements in Autonomous mode, the DMR-200 software was designed to allow users to custom program their terminals. Users can custom program features such as message content, frequency of return link messages, and the operating power modes of the terminal.

Major modules that control the DMR-200 are the GPS, Configuration, and Power Control modules. Other modules such as the Alarm and Timer modules perform Local Data Processing. A common interface for all these modules is a set of read/write memory mapped registers with each register's function defined by its module.

In the forward link Inmarsat-D+ supports long messages and the DMR-200 permits the concatenation of multiple configuration (or write) commands into a single packet. These messages can be used to either control or reprogram the DMR-200. All configurations can be stored in non-volatile memory. For fleet applications the Inmarsat-D+ group messaging can be used to program multiple terminals by sending a single broadcast message to all units.

In the return link, the DMR-200 puts the response to all interrogating messages into short packets that fit within the Inmarsat-D+ return link message format. This allows the user to use the same protocol to query the DMR-200 both over the air and over the RS232 interface.

The common module interface definition facilitates Local Data Processing. Local Data Processing is performed using the Alarm and Timer modules. Up to eight different timers and alarms are defined. The Alarm module interrogates a status indicated by reading a memory-mapped register in any of the DMR-200 modules and testing against a condition. If an alarm is triggered the Local Data Processing causes an "Action" to execute. Like Alarms, Timers also generate Actions when they expire. Timers can either be generated based upon the time-of-day, or at regular, definable intervals. They can also be programmed to either start under control of the terminal or automatically when the terminal is turned on.

A typical Action consists of one or more writes to a memory-mapped register in the DMR-200. The function defined by this write could vary from putting the unit to sleep, to sending a return link message, to resetting a timer.

In order to support Autonomous mode for global applications the DMR-200 supports automatic satellite selection, based upon the current position of the terminal and a user-defined satellite orbital coordinates.

Programming Examples

A couple of simple examples are outlined below. Other more complicated state machines are possible by using multiple Timer and Alarm modules and by allowing Actions to reset timers.

Example 1 - Report Position Once a Day at 9:00 PM
In this example the Timer module is programmed to start when the terminal is turned on and to generate an Action at 9 PM each day. The Action associated with the timer causes a position request to be generated, a position report to be sent, and then puts the terminal to sleep until the next reporting time (twenty-three plus hours later).

Example 2 - Report Position If Changed Since Last Report. In this example the Alarm Module is programmed to compare the current position with the last saved position and compute the difference. If the difference exceeds a predefined limit an Alarm is triggered. This Alarm then initiates two Actions. One Action writes to the GPS port instructing it to update the last saved position. The other Action generates an API Message that causes a position report to be generated.

INITIAL FIELD TRIAL

During the spring of 1999 SkyWave conducted an initial field trial of the DMR-200 between Canada and Europe. The objective of the trial was to demonstrate end-to-end operation of an autonomous vehicle tracking application.

The vehicle was equipped with a DMR-200 terminal and drove around a 100 km course in the Ottawa area at highway speeds. The DMR-200 was pre-programmed to deliver GPS position reports every 4 minutes. The complete link is shown in Fig 5 and includes three segments as follows:

a) Satellite Link: Ottawa <->Station 12 LES Netherlands,

b) Terrestrial Link I: Netherlands<-> Spain via X.25, and ,

c) Terrestrial Link II: Spain<->Ottawa via Internet.

The position reports were received at the GIS mapping server in Spain and entered as dots on a map of the Ottawa area. The annotated map was then forwarded to back to Ottawa via the Internet. The final result, as displayed on a PC, is shown in Fig 5.

CONCLUSIONS

The paper has presented details of the design and field trial of SkyWave's DMR-200 terminal within the Inmarsat D+ System. Both the terminal and the D+ system are intended for a new class of high volume, low cost applications.

Perhaps the most striking conclusion is that an enormous range of useful applications can be served with short message services such as D+.

ACKNOWLEGEMENTS

The development and field trial activity described in the paper are the result of the support and commitments by a people in a number of organizations.

SkyWave wishes to recognize and thank the following participants and supporters:

Inmarsat, Station 12, Zunibal, Industry Canada, the Canadian Communications Research Centre, and The Canadian Space Agency.

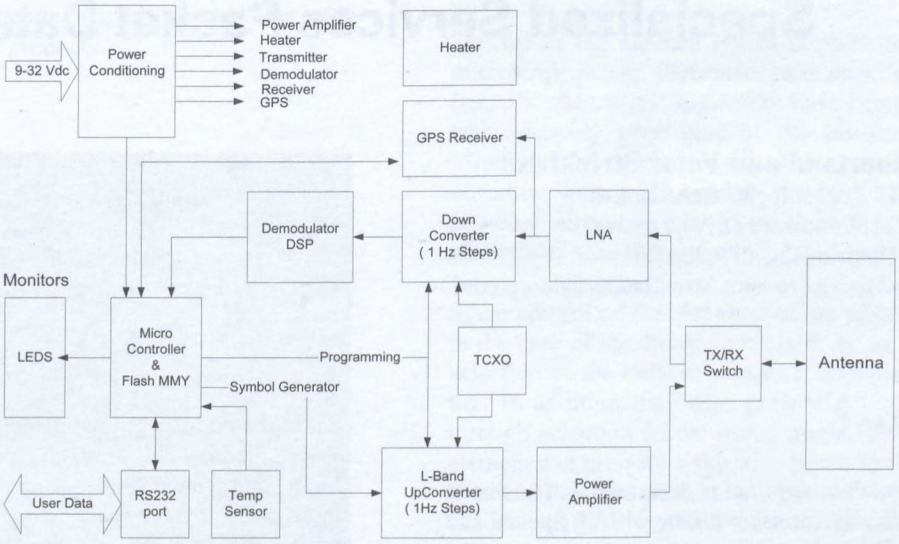


Fig. 2 DMR-200 Block Diagram

Fig 3 Trial Vehicle with DMR-200 Mounted

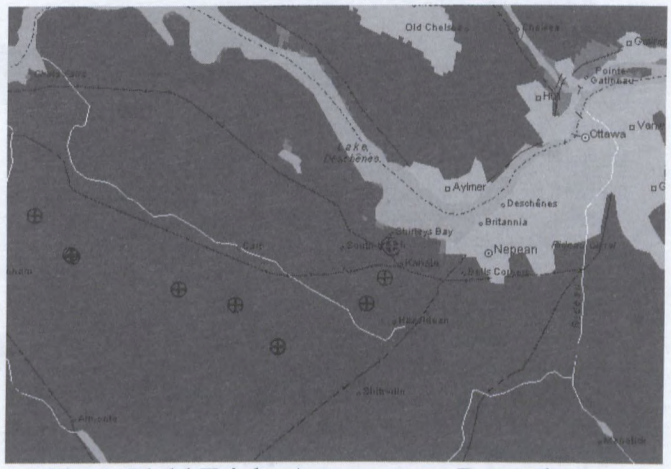


Fig 5 Field Trial - Autonomous Reporting Results

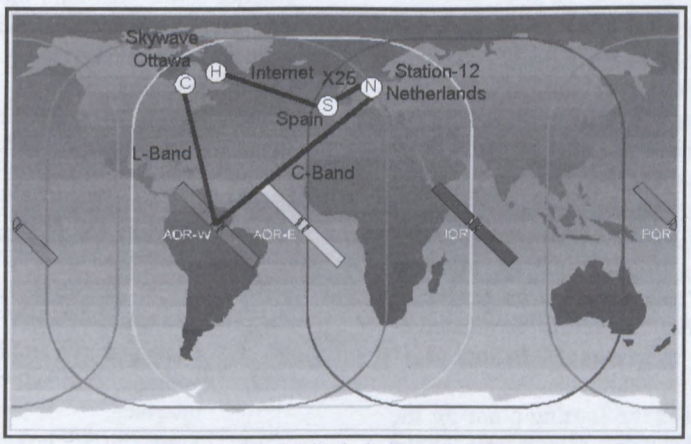


Fig 4 Field Trial Connections

A Low Cost Land Mobile Terminal for MSAT Specialized Services Packet Data

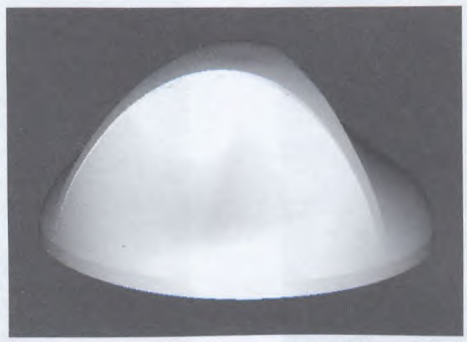
Colin Sutherland and Peter Strickland
EMS Technologies Canada, Ltd
1725 Woodward Drive, Ottawa
Ontario K2C 0P9, Canada
Email: sutherland@calcorp.com strickland@calcorp.com

Michael Moher
Communications Research Centre
3701 Carling Avenue, Ottawa
Ontario K2H 8S2, Canada
Email: michael.moher@crc.doc.ca

ABSTRACT

A low cost land mobile terminal is described that has been developed specifically for use with the MSAT Specialized Services Packet Data. It provides two-way messaging and GPS position determination for vehicles operating anywhere in North America.

When the development program was started in September, 1998, the objective was to produce a terminal that could be manufactured at a fraction of the cost of existing equipment. Several design tradeoffs were necessary in order to meet the cost objective. Firstly, the terminal was implemented as a single, self-contained unit housing all the electronics and the antenna. Only the user interface is, of necessity, external to the unit. Secondly, the half-duplex mode was selected to eliminate the need for a duplexer and two independent frequency synthesizers. Finally, the link budget was tightened to remove excess margin and permit simplification of the design. The most important result of this was the ability to adopt a fixed, omnidirectional antenna.



The availability of a wide range of very small, low-cost parts intended for handsets and other portable devices presents both an opportunity and a challenge to the satellite terminal designer. The opportunity is clear in the form of a plentiful supply of parts intended for the mobile wireless environment. The challenges are to apply highly integrated devices in ways not necessarily envisaged by the manufacturer, to achieve a stable design in the face of rapidly changing product portfolios, and to establish reliable supplier relationships given production quantities that are insignificant in relation to mainstream handset volumes.

The omnidirectional antenna was the starting point for the design, and comprises a crossed drooping dipole at the centre of a circular ground plane. Careful optimization led to an antenna that provides the necessary coverage, is simple and robust, and can be manufactured easily. A single multi-layer circuit board carries all the electronic components and also provides the ground plane for the antenna. The enclosure is a two-piece design with provision for several installation options.

A key element of the design is efficient implementation of the QPSK modem, which has been performed by a team at Communications Research Centre (CRC). Their collective experience in DSP modem architectures spans many years, and the design created for this project reflects this in the performance achieved, and in the minimum hardware requirements.

This paper describes the overall design approach, the antenna design and characteristics, the modem performance, and other interesting aspects of this terminal.

INTRODUCTION

It has been apparent for some time that it is now feasible to produce a low cost land mobile packet data terminal, and that there is considerable demand for such a terminal. The Specialized Services Packet Data (SSPD) offered by TMI Communications supports the two-way transfer of messages between a fixed site such as a dispatch centre and a mobile user anywhere in the MSAT coverage area. The development of the Packet Data Terminal (PDT) was started in September, 1998, and at the time of writing is approaching the Critical Design Review.

Some of the measures adopted to achieve the cost objective are listed below:

- Packaged as a single self-contained unit including antenna and all electronics except the user interface;
- Single circuit board carries all components;

- Fixed, omnidirectional antenna;
- Half duplex protocol; and
- Heavy use of components intended for high volume portable products.

ANTENNA

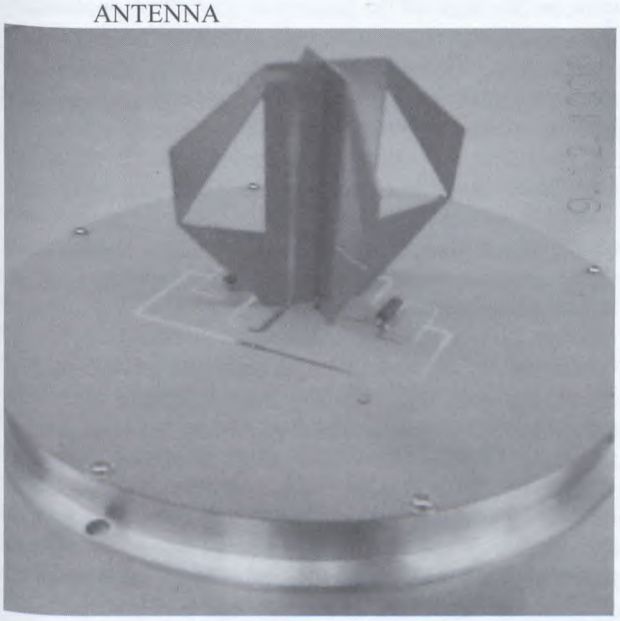


Figure 1: PDT Element

The PDT requires an antenna having right-hand circular polarization and the maximum possible gain in the upper hemisphere, particularly at low elevation angles where multipath losses can be high. The terminal will typically be mounted directly on a vehicle roof-top that is usually metallic and the antenna is required to perform well in this type of installation as well as installations where a conductive roof-top is not present. There are a variety of antenna types that can meet this very general initial criteria including:

- Quadrafilar helices
- MiniCap radiators
- Microstrip patches
- Drooping crossed dipoles
- Crossed slot radiators
- Conical log-spirals
- Dielectric resonators

The quadrafilar helix is the most flexible of these antenna designs, allowing almost unlimited control of the radiation pattern. In general the longer the helix is the better the pattern can be matched to a specific gain mask. This is similar to the general source synthesis problem with constraints on the source norm wherein high spatial frequency components of the desired pattern can be produced with a long source aperture. In this case however a low profile antenna is required and consequently the benefits of a long quadrafilar helix cannot be obtained. In addition the quadrafilar helix is generally more expensive to manufacture than the other antenna options.

Several of the antenna types, in particular the MiniCap, microstrip patch, dielectric resonator, crossed-slot and (usually) the conical log-spiral, have beam peaks at zenith and relatively poor gain at the horizon. Of these the MiniCap and dielectric resonator have the best low elevation gain performance resulting from their small physical size and also have excellent bandwidth for electrically small structures.

Some control of the elevation of the gain peak is possible in the case of the drooping crossed dipole through optimal selection of the radiator elevation above the vehicle roof-top. In addition the beam peak width can be controlled through selection of the droop angle. The crossed dipole element can produce a dip in the gain in the zenith region if a small ground plane is present below the radiator and this dip tends to increase the horizon gain. Note that the PDT circuit board has a ground plane layer which is adequate to provide the horizon gain dip in installations where the unit is not mounted on a conductive roof-top.

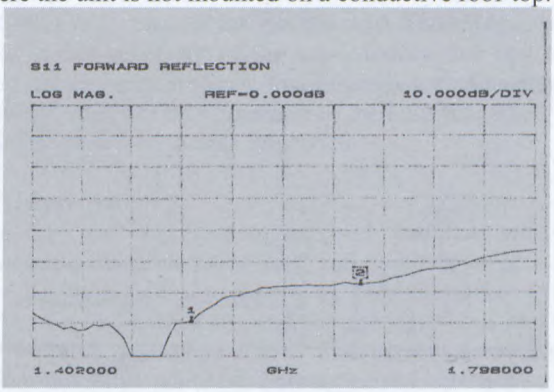


Figure 2: Element Return Loss Without Isolation Resistors in Hybrid

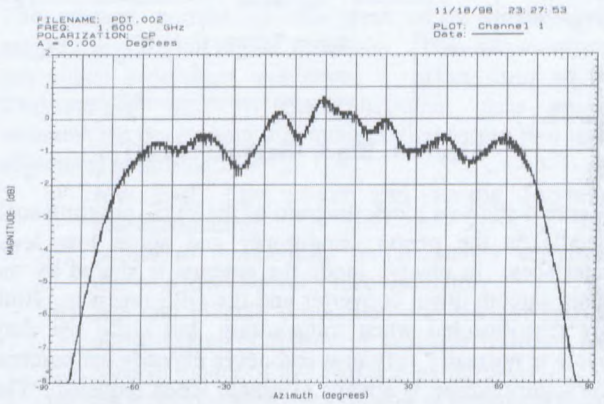


Figure 3: PDT Antenna Radiation Pattern (0 degrees is zenith), Measured on 1m by 1m Ground Plane

EMS Technologies has developed a particularly broadband printed crossed dipole antenna for use in the PDT terminal.

As illustrated in Figure 1 the antenna is manufactured as two printed circuit boards that slide together and drop into slots in the transceiver PCB below. Four 50ohm microstrip lines on the antenna solder directly to 50 ohm microstrip lines on the transceiver board and plated-through holes on the transceiver board provide a ground connection for the lower level of the microstrip transmission lines on the antenna. A microstrip feed was selected for this antenna instead of the usual co-planar waveguide because on the selected substrate microstrip lines have a much lower insertion loss and are less sensitive to the alignment of the crossed circuit boards. The element arms also have an unusual design each having a triangular shape such that the four arm radiating structure forms a crossed pair of drooping bow-ties. The arm shape is a truncated frequency independent structure that provides wider bandwidth than the usual constant width arm (a frequency independent structure is one which remains unchanged when scaled). Several Moment Method analysis tools (WIPL, IE3D and Ensemble) have been used in the optimization of the antenna dimensions.

HARDWARE DESIGN

Architecture

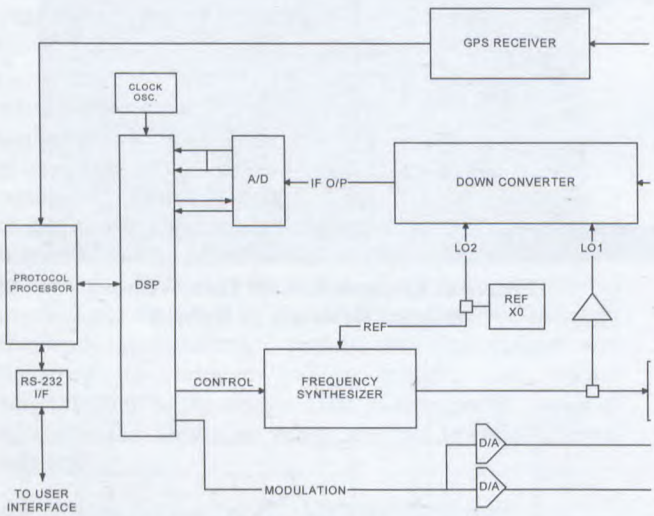


Figure 4: Block Diagram of PDT

Figure 4 shows a block diagram of the PDT, omitting such details as the power conditioner and some low level interfaces. In receive mode the antenna is shared by the main satcom down converter and the GPS receiver. Both are disconnected when transmitting, but since the duty cycle is normally very low and never exceeds ten percent, GPS information is always available when required. The frequency synthesizer is switched between the transmit chain and the down converter, where it provides LO1. The transmitter uses direct I/Q modulation of the L-band carrier generated by the synthesizer, this being simpler and less expensive than the more conventional up converter. A DSP performs all signal processing for both transmit and receive modes. Complete frames of data are passed between the DSP and the protocol processor, which looks after most internal and all external hardware interfaces in addition to executing the SSPD protocol. The primary user interface is provided via an RS-232 serial link to a terminal that facilitates two-way messaging in a format the may be customized to suit each particular application. Basic applications requiring only position reporting can be implemented by the PDT alone without requiring a user interface, once initial commissioning and configuration have been accomplished.

Low Noise Amplifier (LNA)

The PDT low noise amplifier is the state-of-the-art in low cost, high performance design, achieving a 0.4dB noise figure without tuning and using only low-cost surface mount components. The first stage of the LNA uses a discrete PHEMT device that has an outstanding F_{min} of 0.2dB and is offered in an inexpensive plastic surface mount package. Losses in the input matching circuit are typically minimized through the use of a flying lead or single turn custom inductor, often requiring tuning in production test. In this case however high Q tight-tolerance surface mount coil inductors were selected that provide equal performance without any tuning being required. Two versions of the first stage were developed, the first having a measured noise figure of less than 0.28 dB over the PDT receive band with a gain of greater than 15dB, and the second having a noise figure of less than 0.35dB with a similar gain. The version with the higher noise figure has been selected for production because of the lower cost of the PHEMT device used.

The second LNA stage is an unconditionally stable RFIC with a noise figure, when measured in isolation, of less than 1dB. Stability of the second LNA stage is particularly important since it is terminated by a non-absorptive filter. Figure 5 and Figure 6 show the noise figure and gain respectively for the two-stage LNA using the lower cost PHEMT device.

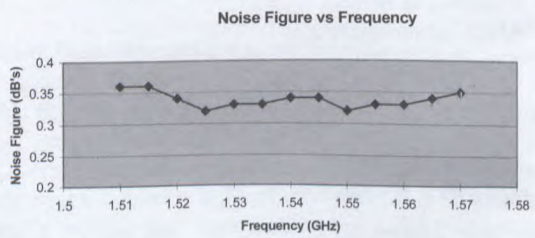


Figure 5: Noise Figure of Two-Stage LNA

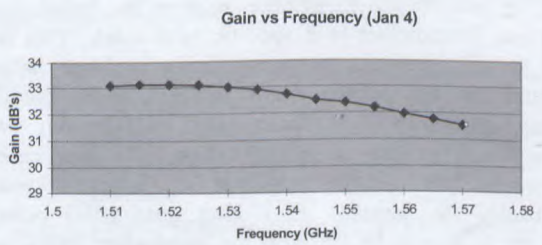


Figure 6: Gain of Two-Stage LNA

High Power Amplifier (HPA)
Since the return link modulation uses filtered DQPSK, a linear HPA is required to meet the spectral mask requirements. Laterally-diffused MOS (LDMOS) devices suitable for L-band use are now available, and one was selected having a nominal P_{1dB} of 12 watts. These devices are remarkably linear even when operated close to compression, and it was found that it was almost possible to achieve the required linearity in a class AB configuration without using additional measures. However, as described later, predistortion was adopted to give margin for temperature and production variations. A three stage monolithic device was selected as the driver, and this is driven directly from the I/Q modulator.

Transmit/Receive Switch
The T/R switch had to combine high isolation, in the region of 50 dB, with very low loss due to the demanding noise temperature budget. This eliminated GaAs switches, so three PIN diodes were used in a series/shunt configuration. The diode parasitics, both series inductance and shunt capacitance, had to be tuned to achieve the required performance, and ultimately a receive path insertion loss of under 0.2 dB was achieved. The transmit loss is about 0.45 dB and the isolation exceeds the required 50 dB across the transmit band.

Down Converter
Direct to baseband down conversion was considered, but rejected for the following reasons. It would require a very high gain at baseband, which would make it susceptible to the high EMI environment of a transport truck, and also it would also be susceptible to microphony in the presence of high vibration levels. A conventional dual-conversion architecture was therefore adopted. The starting point for the down converter design was a coupled resonator SAW filter centred at 83.16 MHz. This component is in high volume production for use in handsets, and has sufficient selectivity to permit the use of a 455 kHz second IF. The reference oscillator frequency was set at 82.705 MHz, enabling it to be used directly for LO2 without the use of an auxiliary PLL. The L-band front end filter is a ceramic resonator type, and the 455 kHz filter is a piezoelectric type selected for a small group delay variation over the approximately 6 kHz signal bandwidth. The main active devices are a combined RF amplifier/first mixer, and a second mixer/AGC amplifier with an AGC range exceeding 50 dB. The real 455 kHz IF output is undersampled at 67.5 kHz yielding a digitized IF of 17.5 kHz.

Frequency Synthesizer
The synthesizer performance required for this type of terminal is quite demanding, and made more so by the decision to use direct I/Q modulation rather than an up converter chain. In addition to channel tuning, the synthesizer must provide fine tuning capability to deal with reference oscillator drift and doppler offset due to vehicle motion. Fortunately, direct digital synthesizer (DDS) chips with an integral D/A converter are now available, providing the starting point for a low cost synthesizer with the required performance. The challenge is to shift the DDS tuning range to L-band without introducing unacceptable spurious products or phase noise. The solution adopted here is to operate the DDS at a very low output frequency to achieve exceptional spectral purity, then mix this signal with a divided version of the reference, which is also very clean, and finally to multiply this signal to L-band using a PLL. The PLL provides coarse steps and the DDS provides fine tuning within each step with a resolution of approximately 3 Hz.

Digital Hardware
The digital hardware comprises two processors with associated RAM and flash ROM, an RS-232 interface, and a few miscellaneous functions. The DSP is an ADSP2185L, selected for compatibility with a platform at CRC that was to be used in the development of the DSP code for the PDT. The protocol processor is an MC68LC302, chosen for its low cost and compatibility with earlier members of the same family that had been used on previous projects. Interprocessor communication uses the internal DMA interface of the DSP processor, and is interrupt driven in both directions.

GPS Receiver
The GPS receiver uses an approach that lies half way between a complete OEM card and a chip level solution. The RF front end is packaged as a small module in the style of a ball-grid array, and the processor and memory are supplied as discrete chips. This permits considerable layout flexibility while minimizing the risk through the use of a fully tested RF front end.

MODEM IMPLEMENTATION

The modem portion of this terminal is implemented entirely in a digital signal processor. This allows one to use signal processing techniques to offset some of the hardware/cost tradeoffs made in other parts of the terminal. In developing this modem, there were five main objectives identified:

i) low cost, low power and simple hardware implementation;
ii) ability to compensate for frequency variations due to a low-cost reference oscillator;
iii) fast and reliable acquisition and reacquisition to allow seamless transition between transmit and receive modes of this half duplex terminal;
iv) high performance demodulator to minimize losses and allow margin for other hardware components in the system; and
v) precompensation of the transmit nonlinearity to meet the transmit spectral mask.

Obviously, the opportunity here is to lower the overall terminal complexity and cost through the use of compensation techniques in the DSP.

The first challenge was met by selecting an Analog Devices 218x series fixed point Digital Signal Processor. This provided not only a flexible, low power solution, but with its on-chip data and program memory, it provides a single-chip hardware solution.

To keep terminal costs low, a reference oscillator with an accuracy of only ±25 ppm had been selected. As a result, initial frequency error on the order of several channel bandwidths is possible. The DSP approach allows division of this uncertainty into a number of bins and rapidly searching through the bins until the desired signal is located. There also may be significant variation in this frequency during the period when the terminal is warming up. Again because the tracking mechanism is implemented in the DSP, short term variations can be tracked accurately in software while longer term corrections are passed on to and corrected by the hardware.

Forward link acquisition is based on the 32-bit unique word associated with each 238ms outbound frame. This unique word is used to provide timing and frequency synchronization. Although the modulation is DQPSK, frequency detection and correction in software permits the demodulation of signals with large hardware frequency offsets with negligible degradation. It also allows tracking of frequency offsets with time whether they are induced by reference oscillator variations due to temperature or Doppler variation due to a change in terminal motion. Initially, acquisition is performed on two successive unique words to provide sufficiently low probabilities of false alarm. Consequently, acquisition occurs in two frame periods with high probability (assuming the correct frequency bin). For reacquisition, frequency and timing of the next frame are estimated from the last known frame boundary. Timing and frequency drift are such that only one unique word is required to confirm frame sync in this case.

The modulation scheme in the forward direction is differential QPSK with rate ¾ convolutional coding. There are several keys to optimizing the performance of this demodulator. The first is maintaining very good frequency and timing synchronization, so the that loss relative to an ideal differential front end is minimized. It was explained above how this frequency and timing synchronization was obtained. The second is maintaining soft decisions throughout the received chain. This implies multiple bits of precision to allow for potential gain variations in the receive chain and also to accurately capture the effects of fading. The Viterbi decoder was implemented in software on the DSP using 16-bit values. This not only reduces the part count but also eliminates the losses that are inherent with 3-bit soft-decision hardware decoders.

On the transmit side of the terminal, one of the key requirements is meeting the transmit spectral mask with minimal terminal cost in terms of hardware and power. The spectral mask specification is representative of quasi-linear operation. The approach taken with this terminal is to choose a low cost amplifier that nominally meets the requirements, and then to precompensate the transmitted signal in the DSP for the nominal amplifier characteristics. In this way, there is sufficient margin to reliably meet the spectral requirements across manufacturing tolerances.

CONTROL & PROTOCOL PROCESSING

The control processor executes the SSPD protocol as required for communication and messaging via the TMI Data Hub. A few changes were found to be necessary to deal with the half duplex format, which had not previously been implemented. In addition to the satellite protocol, the control processor is responsible for many of the hardware interfaces and algorithms, such as:

- Self test
- Downloading code to DSP
- Assisting initial acquisition of DH-D channel
- Crystal temperature compensation
- Setting HPA bias levels
- Transmit/receive switching
- Synthesizer tuning
- Communication with GPS receiver
- Communication with user interface
- Monitoring and controlling discrete I/O lines

The general philosophy has been to use software algorithms to permit reduced cost hardware implementation where possible. The use of predistortion in the DSP is one example, and the simple AT-cut crystal (rather than a TCXO) used in the reference oscillator is another. The crystal temperature is monitored, and its frequency offset is available by comparison with the satellite DH-D channel frequency. For land mobile applications the doppler shift is small compared with the crystal drift. This makes it possible to implement a software TCXO that adapts to the characteristic of an individual crystal, removing to a large extent the effects of initial tolerance, temperature variation and aging. The benefit of this capability is a terminal that always starts up close to the centre of the wanted channel, minimizing the time that must be spent searching and leading to rapid acquisition. This is particularly beneficial when the terminal is programmed to "sleep" to conserve battery power, waking for brief periods to check for messages.

PACKAGING

The circuit board is a laminate of an upper layer of low loss RF substrate, and a multilayer board of conventional FR4 material. Plated vias provide connections to the copper tracks on the RF layer, as well as interconnections within the multilayer board. A die cast aluminum alloy

base provides the ruggedness required for the most demanding land mobile applications, including installations on locomotives. It also provides overall shielding, reducing potential EMC problems. The radome is injection molded plastic, selected for the best compromise among the required characteristics of resistance to infrared and chemical exposure, resilience at low temperatures, and low reflection and loss at L-band. A low cost environmental connector was selected, the choice being simplified by the absence of any requirement for coaxial inserts.

CONCLUSION

The design of a low cost terminal for use with the MSAT SSPD has been described. The integrated packaging, adoption of an omnidirectional antenna, half duplex protocol, and selection of high volume parts were the keys to a low recurring cost. The use of software to reduce hardware cost was also discussed.

Coping With Obstructions to Line-of-Sight in Mobile Satellite Systems: A Comparison of Different Systems

Leonard Schiff
Qualcomm Inc.
10185 McKellar Court
San Diego, CA 92121
Email: lschiff@qualcomm.com

ABSTRACT

In terrestrial cellular systems the path between base station and terminal is seldom line-of-sight. As a result, these systems must be capable of coping with multiple tens of dB. of variability in path loss. By contrast, Mobile Satellite Systems (MSS), which provide communications through satellites to mobile users, depend on the fact that the path between satellite and user is usually clear of obstructions. Nevertheless, a small but significant number of times there are obstructions and each MSS must have a way of coping with these events. This paper compares the approaches taken in the ICO, Iridium and Globalstar systems[1].

I-INTRODUCTION

In terrestrial cellular systems the path between base station and terminal is seldom line-of-sight. There is a large path loss variation caused by variations in distance between base station and user. Then there is substantial variation in the so-called shadow loss even at a specific distance. Finally there are the microscopic variations in signals strength with distances of the order of a wavelength usually modeled as Rayleigh distributions. As a result, these systems must be capable of coping with multiple tens of dB. of variability in path loss.

In a typical MSS things are quite different. The path between satellite and user is usually unobstructed. Assuming the satellite beam shapes and/or power is used to try to compensate for $1/r^2$ loss, the variation in loss at different satellite elevations is not that large. There is some variability due to specular reflections, particularly at low angles. And there is microscopic variation that is usually modeled as Ricean. But taken together these effects produce far less variability than in a terrestrial cellular environment. That's important because it would be difficult to build satellites and handsets able to cope with multiple tens of dB. greater path loss on every link.

Nonetheless, there will be times when the path is obstructed (by vegetation, terrain, man-made structures, etc.) and each system must have a strategy for coping with those times. In section II below we describe the strategies used by ICO, Iridium and Globalstar to deal with such obstructions that occur during a normal call. In section III we contrast the efficacies of these approaches and finally in section IV we discuss the cost, in terms of satellite power and bandwidth, for the different approaches.

II-COPING STRATEGIES

All the systems we will compare make use of power control. So they all try to power each link with the minimum power needed to provide the error rate objectives and no more. This is done to reduce interference to other users. In the Iridium system this power control, normally used to adjust for the small variations in path loss, is used to cope with the larger losses encountered during blockages. It is a simple strategy made possible, in part, by the fact that Iridium flies at a lower altitude and has a smaller average path loss than the other systems. This feature is sometimes described by saying that Iridium has a 16 dB. margin against such fades. The word margin is somewhat misleading. In most satellite systems a margin of X dB. implies that X dB. more than the power normally needed is always being transmitted. In this case it means that power control adjusts for increased path loss and the limit of its ability to correct is an added power of 16 dB. compared to the average case. The Iridium link budget advantage is what enables, for example, the user terminal to come up with 40x more power than it usually needs to overcome a 16 dB. obstruction.

The other two systems are also power controlled and have the ability to cope with some amount of excess loss through that mechanism. But the primary means of coping with obstructions is through the use of diversity. Both systems rely on the fact that when communications is to be established from a user to a terrestrial gateway there are usually multiple satellites in the constellation that are jointly visible to user and gateway; and a call can be simultaneously established through, for example, two different links--one on each satellites. If one of the links is obstructed the other link can carry the call. But even though both systems rely on diversity, important details are different. The Globalstar system is an extension of the IS-95 system used for terrestrial cellular.[2,3]

As such, the ability to diversity combine both paths using maximal ratio combining is built-in. This is the optimal form of diversity combining[4]. The ICO system uses switch combining (i.e., one of the two paths is used--whichever is better). The ICO system uses narrow band TDMA modulation and hence cannot easily derive a timing signal to time align the two signals for maximal ratio combining. The spread nature of the IS-95 type signal used in Globalstar is what permits the time alignment in Globalstar.

III-EFFICACY OF STRATEGY

To make a meaningful quantitative assessment of how well any of these schemes works requires a knowledge of how often paths are obstructed and what their excess attenuation is once they are obstructed. That is something that is not available. It isn't that there haven't been a great deal of measurement campaigns. There have. (See, for example, reference 5 and other tests done since that time). But the problem is that any test campaign can only test a relatively small number of areas. There is no way of knowing what is "typical" on a world-wide basis. Based on a limited amount of testing in the Globalstar program the probability of an obstruction is estimated at 0.1 and most of the time excess attenuation introduced by a blockage is above 10 dB., sometimes considerably above 10 dB. The reader should not treat these as hard numbers but only plausible guesses that are in the right range of values. Using these values allows one to say that for Iridium the probability of dropping a call because of an obstruction is certainly less than 0.1--how much less depending on what fraction of the time the excess attenuation is greater than 16 dB.

For the ICO and Globalstar systems the value of the diversity is dependent on how uncorrelated the probability of obstruction is on each path. The satellite are usually well separated in angle because of the constellation design. Hence if 2 satellites are used and the blocking probability on each is 0.1, then the probability of dropping a call due to obstruction is 0.01. This assumes that both obstructions have sufficient loss that the normal power control cannot overcome the obstruction. There is some degradation due to the small percentage of the time the angle between the satellites is small and blockage is correlated and some degradation due to the fact that two satellite are occasionally not available.

There would also be some degradation if a mixed diversity strategy is used. In this approach the system uses two satellites if the highest satellite is below some elevation angle □□□□If not, only one satellite is used. Hence when elevation angles are low and obstruction is more likely, two satellites are used. And when one is sufficiently high the system drops back to only one. For the ICO system the use of this approach can save a substantial amount of satellite power (see next section). For the Globalstar system it does not and this mode is not planned for Globalstar.

The above describes how effective each system is in ameliorating the effect of obstructions and in particular how diversity is used. The reduction in the probability of a dropped call is the greatest benefit of diversity. But there is a subsidiary benefit in that when diversity is used the amount of satellite power is reduced. This effect is perhaps easiest to see in the ICO system. Consider the case of a terminal driving at high speed. The power control cannot adjust the E_b/N_0 to the precise value needed for a given error probability because the fluctuations are too rapid. It can only adjust the mean of the fluctuation of E_b/N_0 to be high enough so that the error objective is reached. Assume there is only one link with the power adjusted so that the average E_b/N_0 over a frame were exactly right to just meet the desired frame error rate. But if the user had two such independent links of the same power the frame error rate would be better than the desired frame error rate. Because whenever there is a failure on one link the other link may still be successful. Hence it is possible to reduce the average E_b/N_0 on both links and still reach the desired frame error rate. The author is not aware of any place in the literature where the computation of how much power is saved but it is clear qualitatively that there is a power reduction.

For the Globalstar case the equivalent effect for the forward direction is described in reference 6. Results of a simulation are given there. We summarize these results in Tables 1 and 2 below.

velocity (mph) \ k value in dB.	10	15	20
0	2.2	2.2	2.2
5	6.47	3.75	2.82
20	3.81	2.89	2.55

Table 1-Eb/No (in dB.) Needed for single path

velocity (mph) \ k value in dB.	10	15	20
0	none	none	none
5	2.0	0.6	0.2
20	0.7	0.2	none

Table2-Diversity Gain (Eb/No reduction) for 2 Equal Paths

The E_b/N_0 values in Table 1 come from single path simulations for different velocities and values of Ricean k. The values in the second table show the reduction in total E_b/N_0 when you have two equal paths (i.e. the total E_b/N_0 needed for 5 mph and a k of 10 dB is 6.47-2.0=4.47). Since the paths are equal you need 3 dB. less for each of the two paths. The results given in the reference

for two unequal paths show that the diversity gain is only slightly reduced when the paths are unequal in strength.

For the ICO case we do not have any quantitative results and even in the Globalstar case the power reduction is highly dependent on the specific cases so it is hard to estimate an average value. But this reduction in power in the forward direction is important to note because in the next section we will estimate the extra power burden diversity places on the satellites. That burden is an increase in power. But it is the gross increase. To find the net increase one must subtract the diversity gain value above.

IV-COST OF STRATEGY

Each system's coping strategy imposes its own cost on the satellites and handsets. For Iridium the cost of powering through obstructions imposes the price of 40x higher power capability than needed for the average case. But clearly the satellite does not require 40x the power--not all users are simultaneously obstructed and not all require the full 16 dB. of reserve power. Assume that the average power needed for an unobstructed link is s and the average power of an obstructed link is S. Assume that the link is clear with probability q and obstructed with probability p=1-q. Then the average power needed by a call is qs+pS and the power increase factor over a line-of-sight call is q+p(S/s). As an example assume that the probability of blockage is 0.1 and that the average power increase for an obstructed call is 10 dB. Then the power increase factor for all calls is 1.9 or 2.78 dB. The satellite has to be sized larger in power than 2.78 dB. more than what it would need to handle only clear calls because one must consider the variations around the average. But this adds very little to the needed power when the number of simultaneous calls is high. If, for example, the satellite is sized to handle 1000 calls then a two standard deviation allowance will increase power only 3%. This gives a feeling for the resource cost for the Iridium strategy.

In the ICO system diversity extracts a price in power at the satellites. All calls that are in diversity mode get transmissions from two different satellites each of which is powerful enough to supply the objective for frame error rate. That's twice the power needed to supply a call in a non-diversity mode in an unblocked state (less the diversity gain mentioned in the last section). As pointed out in the previous section the use of a mixed strategy will reduce this penalty to a factor of 1 rather than 2 for calls in which the highest satellite is above □□□in elevation. This comes at the cost of slightly degraded protection for the non-diversity calls above □□. So for this mode the satellite increased power is somewhat less than 3 dB., depending on the value of □□□and the probability that the highest satellite is above □□. But the diversity mode used in ICO has an impact on the bandwidth resources as well as power. When two different satellites transmit to a user in diversity mode they must transmit on different frequencies (or on the same frequency in 2 different time slots). Transmission on the same frequency at the same time would cause too much interference for the narrowband TDMA modulation used. This may or

may not be an important factor depending on whether the assigned bandwidth is large enough to allow the system to be power limited rather than bandwidth limited.

The Globalstar system, by contrast, requires no extra bandwidth (although two walsh codes are used which takes up more of the non-power resources). Both satellites transmit on the same frequency; the interference being overcome because of the wideband CDMA modulation. To examine the satellite power penalty we begin by analyzing an idealized system and then make adjustments for some practical implementation details.

Consider a satellite with ideal antennas for each beam for this purpose. This would be an antenna pattern with ideal isoflux correction. The gain would increase with increasing angle from nadir to exactly compensate for the increased space loss caused by the curvature of the earth--i.e., that the earth is a larger distance away with increasing angle. Therefore the EIRP over the whole area on earth illumined by the beam is constant. The EIRP outside the nominal area falls off rapidly. Further the different beams are matched in gain so that not only is the EIRP constant across the beam but that value will be the same in the next beam and all over the satellite footprint. That was the goal of the antenna designers. One exception was that for the far outer portion of the outer forward beams, the beam was rolled off harder than an isoflux correction would indicate because of requirements on power flux density limitations at low user elevation angles to limit interference into terrestrial microwave systems.

Now consider the case where satellites with such beams are in the position where a beam in each of two satellites covers the same area. Assume that this area on the earth has a total of N users active and because all N users are in the beam of each satellite we can consider two disparate ways of serving those customers. The first method is to have N/2 users served from one satellite and N/2 served from the other--a technique that doesn't use diversity. The second is the diversity technique of serving all N users from both satellites. Let us compare the satellite power required for each scheme. Consider the first scheme first. Half the N users are "tuned" to one satellite--i.e., using its PN pattern to despread-- and half are tuned to the other. Since all users are in exactly the same propagation condition each will require the same average satellite power P to produce the same E_b, energy per bit, at each receiver. Each receiver sees a thermal noise density N_o and an interference density I_o which is caused by the interference effect of the N/2 users at power P. The fact that each user is on a different walsh channel guarantees that the receiver doesn't see the interference effect of the N/2-1 other users on its own satellite because of the orthogonality of the walsh functions. Hence the SNR each user receives in this scenario is E_b/N_o+I_o while each satellite expends NP/2 units of power and the constellation as a whole expends NP.

Now contrast this with the case where all N users are served in diversity manner from both satellites. Let the power used by each user on each satellite be P/2. Now each receiver on the ground uses two fingers for reception--one for each satellite. The finger tuned to the first satellite receives an energy per bit value of $E_b/2$ on each finger. Its thermal noise density is N_o and the interference density it sees from the second satellite is I_o because, compared to the scenario above, it sees twice as many interferers, each one of whom is at one half the power. So the SNR on each finger is $E_b/2(N_o+I_o)$. But with maximal ratio combining the SNRs add. Hence the SNR for each receiver is still E_b/N_o+I_o as it was in the last scenario. And like the last scenario each satellite expends NP/2 units of power and the constellation NP overall. The conclusion is that diversity costs nothing extra in terms of satellite power.

Of course the actual situations with beam overlap are more complicated and more complicated to analyze. And the satellite and user antennas cannot be as ideal as we have described. And the operation of the combiner actually involves a couple of tenths of a dB. of combining loss. So this conclusion cannot be taken too literally. In order to gain some understanding of what diversity will cost in more practical situations we begin with the case of diversity with equal path gains and P/2 units of power from each satellite as above and consider variations from the equal gain. The variations can be caused by departures from perfect isoflux correction in the satellite antennas or by non equal gain in all directions by the user antenna or by multipath on one of the paths, etc. Assume that one of the path gains becomes poorer than the other by X dB. Assume that we still apply equal satellite power on each satellite (equal RF power not EIRP--the gain of the satellite antenna in part of the total path gain). The total satellite power must then increase to P' (P'/2 in each one) to maintain the same SNR. It's straightforward to calculate the increase of power as a function of X. The result is shown in Figure 1 below.

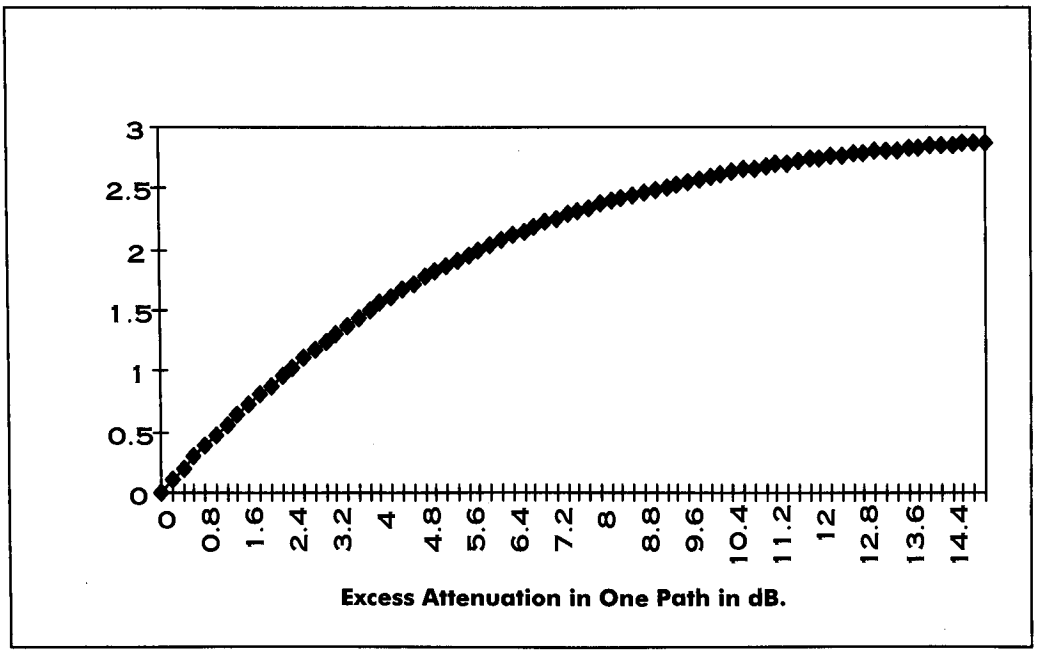


Figure 1: Increase in Satellite Power as a Function of Excess Attenuation in one of the Paths

With this policy the satellite power increases as the excess attenuation in one of them does. For very large values of excess attenuation the satellite power asymptotes to 3 dB. That is quite obvious because a path with large attenuation plays negligible role in boosting the SNR and the good path which used to contribute half the power must now be boosted by 3 dB. to contribute all. It makes no sense to try to carry both paths if one is highly attenuated. As can be seen above it increases satellite power. But also it may not help the goal of redundancy in case of blockage. If the path gain on the poor path is sufficiently low the receiver looking at that satellite will not even be able to capture the pilot. Not only will that mean that the poorer path is unable to contribute to the SNR but if the good path is suddenly obstructed the poorer path cannot serve the purpose of carrying the call. Hence when the weaker one becomes sufficiently weaker than the stronger it should be dropped. The range of excess attenuation (at which the weaker should be dropped) that seems appropriate is 4 - 6 dB. The extra satellite power is then in the range of 1.5 - 2 dB. Note that this is the extra satellite power at the moment before the poorer one is dropped and is the largest that excess power ever is. The average excess power is much less. The beams are designed so that the beam gain varies by 2 -3 dB. across the whole beam area and as the user varies in his position in both beams the extra satellite power varies between 0 and 2 dB. (for a 6 dB. threshold) and when the threshold is exceeded the weaker link is dropped and the extra satellite power drops to 0. When the weaker link rises above the threshold it is restored. Hence, the worst case price in extra power is 2 dB. and the average value is much less.

REFERENCES

[1] B. Miller; "Satellites Free the Mobile Phone", IEEE Spectrum, March'98

[2] P. Monte; S. Carter; "The Globalstar Air Interface Modulation and Access", 15th AIAA International Communications Satellite Systems Conference, San Diego CA, Feb. 1994.

[3] Telecommunications Industry Association, TIA/EIA/IS-95 Interim Standard (TIA, July 1993)

[4] J. G. Proakis, "*Digital Communications*," McGraw-Hill, 1989

[5] J. Goldhirsh, W.J. Vogel; "Propagation Effects of Land Mobile Satellite Systems: Overview of Experimental and Modeling Results", NASA Reference Publication 1274, Feb 1992

[6] J. Nicolas, L. Schiff; "Benefits, Costs & Implementation of Diversity in the Globalstar System", Fourth European Conference on Satellite Communications Proceedings; Rome, Italy, Nov. 18-20, 1997; pp 188-193

A method for evaluating the capacity of non-GEO satellite constellations for mobile communications

Wolfgang Krewel

Bouygues Telecom
51, Avenue de l'Europe
78944 Velizy, France
Email: krewel@enst.fr

Gérard Maral

Ecole Nationale Supérieure des
Télécommunications
31028 Toulouse cedex, BP 4004, France
Email: maral@tlse.enst.fr

ABSTRACT

The number of subscribers that can be served by a non-GEO satellite system is conditioned by the coverage area and allowable traffic load of each satellite in the constellation. Coverage area and traffic load vary with the terminal location due to constellation dynamics that may result in satellite diversity. Satellite diversity is managed thanks to different diversity techniques. Allowable traffic load, local coverage areas, and diversity techniques are discussed within this paper. A method to evaluate the allowable traffic load density (Erlang per km^2) with respect to satellite diversity is derived. An estimation of the total potential number of subscribers of different circular LEO/MEO constellations is given.

INTRODUCTION

The commercial success of candidate non-GEO satellite systems providing personal mobile communications is related to the size of the market with respect to the overall offered capacity. The capacity of a system conditions the call blocking probability, given the traffic load, and the subscriber's acceptance of the service. For the evaluation of the traffic load of non-GEO satellite communications systems, two basic features must be considered: First, the terminal residing time in a cell is conditioned by the speed of the satellite rather then by the user mobility. Second, the satellite resource allocation is a dynamic process so as to provide the required service quality in various regions (according to the latitute) and user environments (urban, suburban and rural). Satellite diversity, dynamic antenna beam forming are examples of such allocation techniques.

This paper presents a method for evaluating the capacity of constellations of satellites in circular orbits, taking into consideration the influence of satellite diversity. This method is then used to evaluate the capacity of the space segment of first-generation non-GEO mobile satellite communications systems, namely Iridium, Globalstar, and ICO. Intrinsic constellation parameters are given in [1].

Section 2 discusses the allowable traffic load. The allowable traffic load per satellite is calculated using a traffic model based on the concept of street of coverage [2][3]. This model leads to a simple mathematical formulation and is considered accurate enough for the purpose [4]. The model takes into account different channel allocation schemes (queued, non-queued, priority, non-priority, pre- and non pre-emptive, guaranteed HO, etc.). In this paper a non-queued birth-death process is assumed for all constellations. Section 3 discusses coverage area according to either fixed or dynamic satellite antenna beam forming. The terminal to satellite allocation in case of simultaneous coverage by multiple satellites (satellite diversity) is discussed in Section 4, and path diversity techniques are introduced. Section 5 derives the allowable traffic load density as the ratio of the allowable traffic load upon the coverage area. The total number of potential subscribers for typical non-GEO constellations is evaluated in Section 6. Finally, Section 7 concludes.

ALLOWABLE TRAFFIC LOAD

The traffic model based on the concept of street of converage (SOC) has been introduced in [3] and [5].

Arrival Rates

The flux equilibrium of newly generated call attempt rate, λ_n, and incoming handover rate, λ_h determines the average number of handovers (HO) per call attempt:

$$G = \frac{\lambda_h}{\lambda_n} = \frac{(1-P_{b1})P_{h1}}{1-(1-P_{b2})P_{h2}}$$

where the probabilities are:
P_{h1} : a newly generated call faces a HO
P_{h2} : a handed-over call faces a further HO
P_{b1} : blocking probability for new call requests
P_{b2} : probability of hand over failure

Service Rates

The mean call duration, T_{call}, (as well as the arrival rate of newly generated traffic) is assumed to follow a negative exponential distribution.
The probability density function (pdf) of the cell residing time, T_c, applying the street of coverage model is given in [3]. The channel holding time $T = 1/\mu$ is either T_{call} or T_c, whatever is less. The probabilities P_{h1} and P_{h2}, are given by :

$$P_{h1} = \gamma(1-\exp(-1/\gamma))$$
$$P_{h2} = \exp(-1/\gamma)$$

where the user mobility γ is given by $\gamma = T_{call}/T_c$. The total traffic intensity per cell $A_t = \lambda_t/\mu$ is given by:

$$A_t = \frac{\lambda_h+\lambda_n}{\mu} = \lambda_n T_{call}[(1-P_{h1})+G(1-P_{h2})] \quad \text{(Eq. 3)}$$

In the non-priority case ($P_{b1}=P_{b2}=P_b$), the amount of newly generated traffic as a function of P_b is given by

$$A_n(P_b) = A_t(P_b)[(1-P_{h1})+G(P_b)(1-P_{h2})]^{-1} \quad \text{(Eq. 4)}$$

where $G(P_b)$ is calculated according to Eq. 1. The factor $X_h = [(1-P_{h1})+G(1-P_{h2})]^{-1}$ in Eq. 4 determines the ratio of newly generated to total carried traffic, thus $X_h = A_n/A_t$. The variation of X_h as a function of the blocking probability is shown in Figure 1 for residing times of different commercial constellations: 600s (Iridium), 1.000s (Globalstar), and 7.000s (ICO).

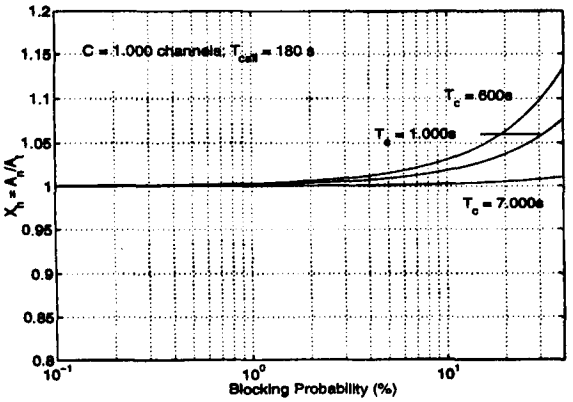


Fig. 1 Newly generated to total carried traffic ($X_h = A_n/A_t$) vs. Blocking Probability for different residing times

It can be seen that X_h varies only slightly with the residing time for small blocking probabilities ($P_b \le 0.01$), and that $A_n \approx A_t$ up to this boundary. The allowable total traffic load A_t is determined by the blocking probability, P_b, and the selected channel allocation scheme. In this paper a non-queued birth-death process is considered, and $A_t(P_b)$ is determined by the Erlang-B formula.

COVERAGE AREA

Permanent coverage of the service area is the considered requirement. The size of the coverage area is determinated by the lower limit of the elevation angle, E_{min}, [6]:

$$F(E_{min}) = 2\pi R_e^2\left[1-\cos(\frac{S(E_{min})}{R_e})\right] \quad \text{(Eq. 5)}$$

where the range S is given by

$$S(E_{min}) = R_e\left[\cos^{-1}\left\{\frac{R_e\cos(E_{min})}{R_e+h}\right\} - E_{min}\right] \quad \text{(Eq. 6)}$$

with the earth radius R_e = 6378 km and h = orbit height. Two coverage area allocation schemes are investigated: (i) pre-defined fixed coverage, (ii) dynamic coverage provided by active satellite antennas.

Predefined fixed coverage

If $(E_{min})_{glob}$ is the minimum elevation angle for permanent coverage all over the service area, then the fixed size of the coverage area is $F = F((E_{min})_{glob})$. This possibly corresponds to a too much demanding case as a higher elevation angle, $(E_{min})_{loc}$, could locally suffice to provide the required coverage. For example, with Iridium the lowest minimum elevation angle providing global coverage is 8.2 deg. A higher minimum elevation angle is obtained for non-equatorial locations, so that $(E_{min})_{loc} \ge (E_{min})_{glob}$, and $F((E_{min})_{glob}) \ge F((E_{min})_{loc})$.
However, predefining a fixed coverage with $(E_{min})_{glob}$ as the minimum elevation angle results in an overlap of several coverage areas, as shown in Figure 2. This provides an opportunity for satellite diversity (see Section 4).

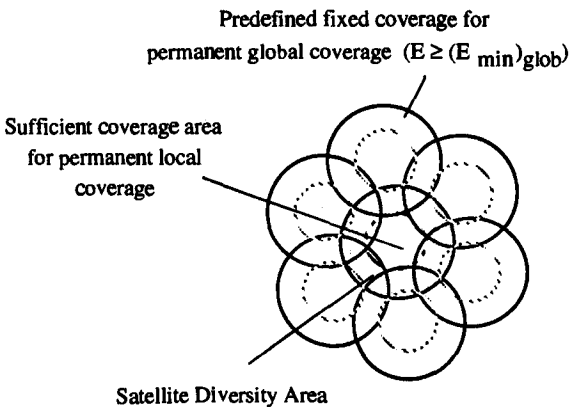


Fig. 2 Predifined fixed coverage of the service area

Dynamic coverage

Active array antennas allow the generation of beams with variable shape and size in time [7][8][9]. Figure 3 shows that the size of the coverage area reduces to the minimum area that provides permanent coverage at a specific terminal location.

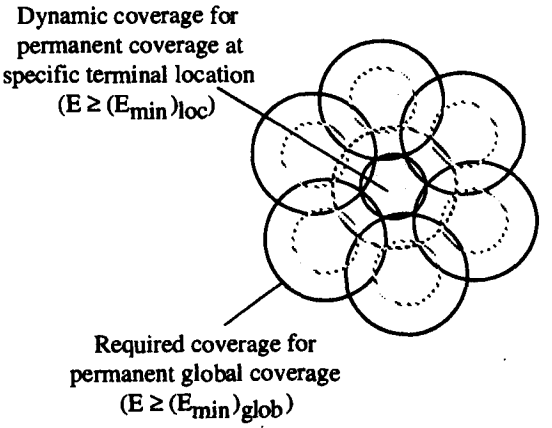


Fig. 3: Dynamic coverage of the service area

The corresponding coverage area is then $F = F((E_{min})_{loc})$, where $(E_{min})_{loc}$ is the minimum elevation angle providing permanent coverage at a given terminal location.

SATELLITE DIVERSITY

Satellite diversity deals with the situation where several satellites are above a specified elevation angle for a given user on the earth [1]. Satellite diversity then allows for either combining signals transmitted via several satellites (*combining* path diversity) or selecting the signal from a non-blocked satellite path among all path from the covering satellites (*selection* path diversity).

Combining path diversity

Combining diversity as a fade mitigation technique imposes the implementation of special means to result in a combined signal superior to any single signal. Satellite systems using those facilities will hereinafter be called *diversity based systems*. Diversity based systems (such as Globalstar, for instance) are characterised by high multiple visibility statistics (dual visibility about 100% of time), whereas non-diversity based systems (e.g. Iridium) are characterised by high fade margins (about 10 dB higher than in diversity based systems).

Combining diversity is of importance in shadowed environments (sub-urban and tree shadowed), where fading mostly does not result in complete blockage. In this case, already dual satellite diversity yields system cost reduction and better user acceptance (smaller size of terminals, for instance) [10]. Due to the smaller link margin, however, the availability of *two* non-blocked satellites are required. The margin reduction resulting from diversity (diversity gain) has to be paid by a decrease of the allowable traffic load density, since simultanous coverage by two non-blocked/non-saturated satellites is required for communications.

Selection path diversity

Communications in rural and urban environments are mainly effectuated under LOS-conditions. Subscriber terminals of all candidate systems (diversity based or not) will then operate with a single satellite in view. Satellite diversity may therefore be available but signal combining is not used, either because fading is negligable or that strong that it cannot be overcome by diversity combining. Diversity results then in the opportunity to select one of all covering satellites, following a selection path diversity scheme. In the simplest scheme (scheme a) the terminal selects the nearest satellite providing the strongest signal and sends an access request. If no channel is available, the access is denied otherwise the call can proceed. The second scheme (b) differs from scheme (a) in that an access is not denied in case all channels of the nearest satellite are busy but the second nearest satellite is then selected.

ALLOWABLE TRAFFIC LOAD DENSITY

Allowable traffic load density is the ratio of the allowable traffic load (see Section 2) upon the coverage area (see Section 3), and expresses in Erl./km^2. Formulas will be derived according to the considered path diversity technique (see Section 4), which impacts the call blocking probability.

Combining path diversity

The blocking probability for each satellite (P_{b2}) will be derived for fixed and dynamic coverage as a function of the overall call blocking probability, P_{bc}.

Fixed coverage. The coverage area is of constant size all over the service area, and provides permanent coverage by at least one satellite. This is given for $(E_{1\min})_{glob}$, and the corresponding coverage area is $F=F((E_{1\min})_{glob})$. Furthermore, two non-blocked satellite channels are required for combining diversity. Call blocking and dual coverage are statistically independent. The allowable traffic load density is therefore:

$$\rho = \frac{A_n(P_{b2} = 1 - \sqrt{\frac{1-P_{bc}}{P_{v2}}})}{F((E_{1\min})_{glob})}$$

where P_{v2} is the probability that at least two satellites are visible. The fixed coverage of an area corresponding to $(E_{1\min})_{glob}$ is a suitable allocation scheme only for systems performing (at least approximately) permanent dual coverage at $(E_{1\min})_{glob}$, since Eq. 7 requires $P_{v2} > (1-P_{bc})$. Assuming a call blocking probability of $P_{bc} = 1\%$ dual visivility of at least $P_{v2} = 99\%$ at $(E_{1\min})_{glob}$ is required. In case of Globalstar, for example, $P_{v2} \geq 99\%$ at $(E_{1\min})_{glob} = 10°$ is given for terminals located within a latitude band of $25° < \iota < 50°$ [11]. If coverage beyond this band is required, a higher call blocking probability has to be accepted. At the equator, for instance, P_{bc} exceeds $(1- P_{v2}) = 0.16$ due to a dual visibility of $P_{v2} = 84\%$ at latitude $\iota = 0°$ for $(E_{1\min})_{glob} = 10°$.

A more reasonable coverage area corresponds to the minimum elevation angle that performs permanent (i.e. $P_{v2} = 1$) coverage from at least two satellites all over the service area. The minimum elevation angle above which at least *two* satellites are permanently visible all over the service area is hereinafter named $(E_{2\min})_{glob}$. The allowable traffic load density is then:

$$\rho = \frac{A_n(P_{b2} = 1 - \sqrt{1-P_{bc}})}{F((E_{2\min})_{glob})} \qquad \text{(Eq. 8)}$$

Dynamic coverage. The disadvantage of the fixed coverage area originates in the fact that a smaller coverage area would be sufficient for permanent coverage, since $(E_{\min})_{loc}$ is larger than $(E_{\min})_{glob}$ almost all over the service area. In case of dynamic coverage the allowable traffic load density then becomes:

$$\rho = \frac{A_n(P_{b2} = 1 - \sqrt{1-P_{bc}})}{F((E_{2\min})_{loc})}$$

where $(E_{2\min})_{loc}$ is the minimum elevation angle above which at least two satellites are permanently visible at a given terminal location.

Selection path diversity (scheme a)

In this scheme, the call blocking probability is directly conditioned by the blocking probability of the nearest

satellite (and only of this satellite, since only the best can be selected). Thus $P_{b2} = P_{bc}$, regardless of whether the service area is covered by one or several satellites.

Fixed coverage. In case of fixed coverage ($F = F((E_{min})_{glob})$) the allowable traffic load density is

$$\rho = \frac{A_n(P_{b2} = P_{bc})}{F((E_{min})_{glob})} \qquad \text{(Eq. 10)}$$

Dynamic coverage. Dynamic coverage takes into account the local value of E_{min}, so that

$$\rho = \frac{A_n(P_{b2} = P_{bc})}{F((E_{min})_{loc})} \qquad \text{(Eq. 11)}$$

Selection path diversity (scheme b)

Fixed coverage. Similarly to the combining satellite allocation scheme, the size of the fixed coverage area is determined by $(E_{1min})_{glob}$ or $(E_{2min})_{glob}$ with consequences on the blocking probability. We will first consider the allocation to fixed single coverage. The probability that a call is blocked is then the sum of the probability to be covered by only one satellite that is blocked, on the one hand, and the probability to be covered by two satellites, both being blocked, on the other hand. The allowable traffic load density is therefore:

$$\rho = \frac{A_n(P_{b2} = \sqrt{\frac{P_{bc}}{P_{v2}} + \left[\frac{1-P_{v2}}{2P_{v2}}\right]^2} - \frac{1-P_{v2}}{2P_{v2}})}{F((E_{1min})_{glob})} \qquad \text{(Eq. 12)}$$

Figure 4 visualises the allowable traffic load per satellite as a function of the multiple visibility probability, P_{v2}, for a call blocking probability of $P_{bc} = 1\%$ (C = 2400 channels). In this chart, the minimum value ($A(P_{v2} = 0)$) indicates the allowable traffic load for single visibility, whereas the maximum ($(A(P_{v2} = 1))$ is the allowable traffic load for permanent coverage of at least two satellites.

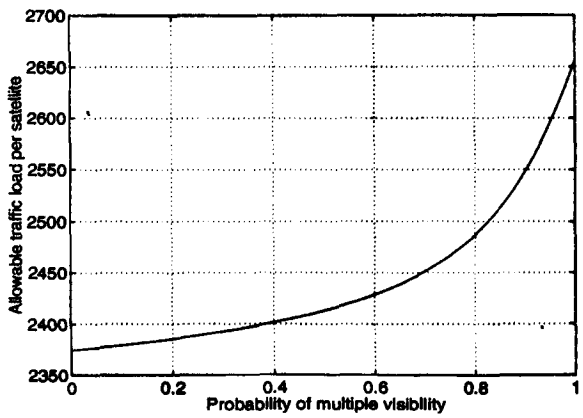


Fig. 4 Allowable traffic load per satellite for the Erlang-loss system (C = 2400 channels per satellite, $P_{bc} = 1\%$)

We now consider the coverage area corresponding to $(E_{2min})_{glob}$, i.e. permanent coverage from two satellites. The call blocking probability is then the probability that both satellites are blocked. The allowable traffic load density is given by:

$$\rho = \frac{A_n(P_{b2} = \sqrt{P_{bc}})}{F((E_{2min})_{glob})} \qquad \text{(Eq. 13)}$$

Dynamic coverage. In the case of dynamic coverage area allocation to $F = F((E_{1min})_{loc})$, the allowable traffic load increases to (since $F((E_{1min})_{loc}) \leq F((E_{1min})_{glob})$)

$$\rho = \frac{A_n(P_{b2} = \sqrt{\frac{P_{bc}}{P_{v2}} + \left[\frac{1-P_{v2}}{2P_{v2}}\right]^2} - \frac{1-P_{v2}}{2P_{v2}})}{F((E_{1min})_{loc})} \qquad \text{(Eq. 14)}$$

The dynamical allocation to $F = F((E_{2min})_{loc})$, as discussed for the case of combining satellite diversity allocation, results in

$$\rho = \frac{A_n(P_{b2} = \sqrt{P_{bc}})}{F((E_{2min})_{loc})} \qquad \text{(Eq. 15)}$$

Example

In Table 1 , numerical values of the allowable traffic load density are given for Globalstar at latitude $\iota = 40°$ (see Table 2)

Table 1: Allowable traffic load density in 10^{-5} Erl/km^2 (Globalstar, $\iota = 40°$, $P_{bc} = 1\%$)

Coverage		Combining	Selection (a)	Selection (b)
Fixed	$F((E_{1min})_{glob})$	*(8.7)*	8.8	9.8
	$F((E_{2min})_{glob})$	**5.9**	**5.9**	**6.6**
Dynamic	$F((E_{1min})_{loc})$	-	30.4	30.6
	$F((E_{2min})_{loc})$	**12.4**	**12.5**	**14.0**

Combining diversity in case of fixed or dynamic dual coverage leads to an allowable traffic load density of $5.9 \cdot 10^{-5}$ Erl/km^2 and $12.4 \cdot 10^{-5}$ Erl/km^2 respectively. Non-shadowed users apply either the path diversity scheme (a) or (b). Should there be dynamic beam forming, the allowable traffic load density would increase up to about $12.5 \cdot 10^{-5}$ Erl/km^2 and $14.4 \cdot 10^{-5}$ Erl/km^2, respectively. These values are lower than the theoretical values from permanent local single coverage ($30.4 \cdot 10^{-5}$ Erl/km^2 and $30.6 \cdot 10^{-5}$Erl/km^2), but are still higher than values obtained from fixed coverage. Thus, dynamic coverage increases the allowable traffic load density significanly (by a factor of nearly two), but the diversity gain due to signal combining has to be paid by a decrease of the allowable traffic load density. The slight increase (about 10%) of the allowable traffic load density with scheme (b) probably does not justify the increase in the system complexity.

COMPARISON OF TYPICAL SYSTEMS

Table 2 presents the parameters for the considered systems Figure 5 shows the local single coverage area for Iridium and ICO, and local dual coverage area for Globalstar.

Table 2: Parameters of considered systems

	Iridium	Globalstar	ICO
No of satellites	66	48	10
Orbit height (km)	780	1414	10355
Channels/satellite	550	2400	4500
$(E_{1min})_{glob}$ (deg)	8.2°	10° [2]	18°
$F((E_{1min})_{glob})$ (km²)	1.53E7	2.64E7 [2]	9.37E7
Min. offered traffic density[1] (10^{-5}Erl/km²)	3.4	5.9[2,3] / 8.8 [2,4]	4.8
Diversity-based	no	yes	on demand
Satellite use	nearest only	combining/ nearest only	nearest only

[1] P_{bc}= 1% assuming Erlang-B
[2] in latitude band -70°≤ ι≤ 70°
[3] if combining
[4] if no combining

The service area of the diversity based system Globalstar has to be covered permanently by two satellites. The comparison is therefore based on guaranteed single coverage in case of the non-diversity based systems (Iridium, ICO), and on guaranteed dual coverage for Globalstar. Scheme (a) as discussed in Section 5.2 is considered.

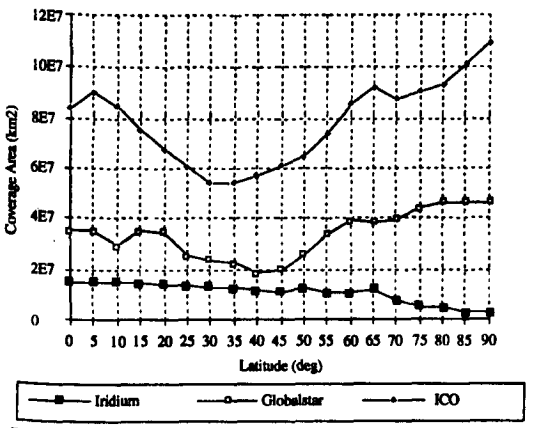


Fig. 5 Coverage area at minimum elevation angles of the non-GEO constellations in Table 2

Allowable traffic load density

The allowable traffic load density is calculated by Eq. 11 for all constellations and shown in Figure 6.

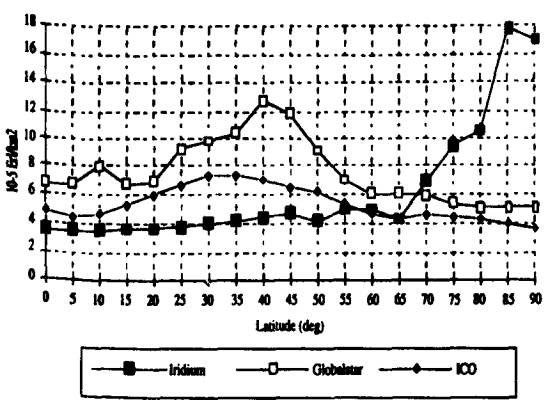


Fig. 6 Traffic load density

Maximum number of subscribers

Deviding the allowable traffic load density at a given terminal location (latitude, ι) as given in Figure 6 by the subscriber activity in Erlang per subscriber leads to the density of subscribers that can be served with the reference quality of service (P_{bc} = 1%). The traffic generated by urban and rural subscribers is taken equal to 7.87E-4 Erl. and 3.25E-3 Erlang per subscriber, respectively [12]. The type of subscriber is decided upon the population density at the terminal location as shown in Table 3:

Table 3: Type of subscriber

	urban	rural	non-populated
Population per km²	> 25	1 - 25	< 1
Subscriber Activity (Erlang/ Subscriber)	7.87E-4	3.25E-3	0

A coarse estimation of the percentage of each category for every latitude band has been taken from [13]. Multiplying the subscriber density by the area of a latitude band between latitude ι_1 and ι_2 gives the maximum number of subscribers as a function of the latitude band that can be served per system. The surface of a latitude band between latitude ι_1 and ι_2 is given by:

$$\Delta F = 2\pi R_e^2(\sin(\iota_2)-\sin(\iota_1)) \qquad \text{(Eq. 16)}$$

This procedure results in an estimation of the maximum number of subscribers per system. Figure 7 shows the number of subscribers per latitude band, and the cumulated number up to latitude of ±90 deg, for both the northern and southern hemisphere of the ICO constellation.

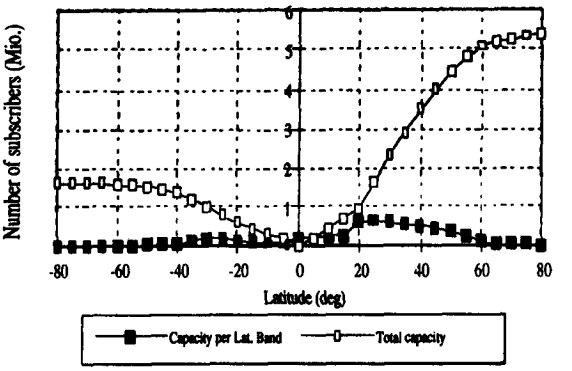


Fig. 7 Estimated maximum number of subscribers (ICO)

Total capacity estimations of the analysed systems under the condition of system saturation in all latitude bands are shown in Table 4.

Table 4 Estimated Capacity of some mobile satellite communications systems (subscribers in millions)

Hemisphere	Iridium	Globalstar	ICO
Northern	4	7.8	5.5
Southern	1	2.4	1.5
Total	5	10.2	7

Figure 8 shows the potential regional share of the traffic load of mobile satellite communications services for all systems:

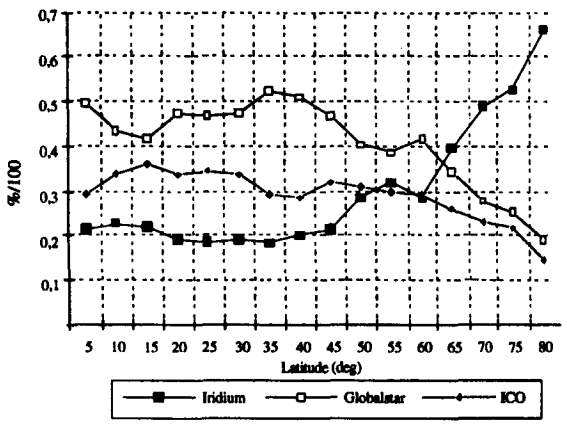


Fig. 8 Potential regional proportion of mobile satellite communications traffic load

Globalstar appears to be in the position to serve the largest part of the mobile satellite communications traffic load (somewhat about 46%). Globalstar officialy anticipates capturing about 2.7 million subscribers by the year 2002, growing to 16 million subscribers by 2012 [14]. However, equatorial subscribers in shadowed environments will face higher link blockage probability since dual visibility at the required minimum elevation angle of about 10 deg is performed with a probability of P_{v2} = 84 %, only. Assuming a fading-probability of P_f = 0.05 and a call blocking probability of P_{bc} = 0.01, as well as these events being not correlated, results in service availability of the order of

$$A = P_{v2} \cdot (1-P_f) \cdot (1-P_{bc}) = 0.84 \cdot (1-0.05) \cdot (1-0.01) = 0.79$$

Iridium aims at a subscriber base of 5 million subscribers until year 2002 [15]. This is feasible under the condition of system saturation in all latitude bands.

CONCLUSION

This paper has outlined a method for evaluating the capacity of non-GEO satellite constellations for mobile communications, taking into account both fixed coverage and dynamic coverage, the latter implying reconfigurable satellite antenna patterns. The method considers the coverage from several satellites in the constellation, and therefore provides means for comparing the capacity for diversity and non-diversity based systems. Dynamic coverage is shown to bring a significant increase in the capacity, and therefore such a feature is an appealing one for future system design. Finally, the paper has presented an estimation of the share of the overall mobile satellite communications traffic load between candidate first generation systems according to their respective capacity. The estimated capacity of all considered systems is about 22.2 Million subscribers world-wide. This number is much below the amount of 426 Million subscribers that has been forecast for the year 2000 by the UMTS-Forum [16]. This shows that there will be enough room in the mobile satellite communications arena for all first-generation mobile satellite systems.

REFERENCES

[1] W. Krewel, G. Maral, 'Single and Multiple Satellite Visibility Statistics of First-generation Non-GEO Constellations for Personal Communications', International Journal of Satellite Communications, Wiley, vol. 16, pp. 105-125, 1998

[2] R. Lüders, 'Satellite Networks for continous zonal coverage', American Rocket Society Journal, 1961

[3] E. Del Re, R. Fantacci, G. Giambene, 'Performance analysis of a dynamic channel allocation technique for satellite mobile cellular networks', International Journal of Satellite Communications, Vol 12, pp. 25-32, 1994

[4] W. Krewel, 'A method for evaluating the capacity of first generation non-GEO satellite constellations for mobile communications', Note Technique 99/1, Ecole Nationale Supérieure des Télécommunicat-ions, 1999

[5] E. Del Re, R. Fantacci, G. Giambene, 'Efficient Dynamic Channel Allocation Techniques with Hand over Queuing for Mobile Satellite Networks', IEEE Journal on Selected Areas in Communicaitons, Vol 13, No 2, pp. 397-404, 1995.

[6] M. Davidoff, 'The Satellite Experimenter's Handbook', The American Radio Relay League, 1994

[7] Zaghloul, A., 'Advances in multibeam communicat-ions satellite antennas', Proc. IEEE, vol. 78, no. 7, pp. 1214-1232, 1990

[8] J-P. Cances, G. Maral, B. Coulomb, R. Lenormand, "Coverage reconfiguration for dynamicallocation in a multibeam satellite system", 15th AIAA International Communications Satellite Systems Conference (ICSSC-15), pp 1032-1041, San Diego, Feb 28 -March 3, 1994

[9] Rao, K.s., Goyette, G. , Gauvin, H., Richard, S., 'Reconfigurable L-Band active array antennas for satellite communications', Can. J. Elect. & Comp. Eng., Vol. 17, No. 3, 1992

[10] J. Goldhirsh, W. Vogel, 'Propagation Effects for Land Mobile Satellite Systems: Overview of Experimental and Modelin Results', NASA Reference Puiblication 1274, 1992

[11] Globalstar, 'Description of the Globalstar System', GS-TR-94-0001, 1997

[12] ETSI Technical Report,'Overall requirements on the radio interface(s) of the Universal Mobile Telecommunication System (UMTS)', ETSI, 1996

[13] W. Cleveland, 'Britannica Atlas', Encyclopaedia Britannica, Inc. 1995

[14] Mobile Communications International, 'Global Personal Satellite Communications', 1994

[15] Bouygues Telecom, '01 Reseaux', no. 51, p.66, July 1998

[16] UMTS Forum, 'A Regulatory Framework for UMTS', Report no. 1 from the UMTS Forum, May 1997

THE ELLIPSO™ SYSTEM – AN OPTIMAL SOLUTION FOR THE CANADIAN MOBILE SATELLITE COMMUNICATION MARKET

John E. Draim
Director, Constellation Design & Launch Vehicles
Ellipso, Inc.
1133 21st St. NW, Suite 800
Washington, DC, 20036, U.S.A.
jdraim @ellipso.com

Cecile S. Davidson
Director, Strategy and Planning
Ellipso, Inc.
1133 21st St. NW, Suite 800
Washington, D.C., 20036, U.S.A.
cdavidson@ellipso.com

ABSTRACT

The ELLIPSO system is one of five so-called Big-LEO[1] commercial satellite communications systems licensed by the US Federal Communications Commission. ELLIPSO is unique among these systems in that it uses a patented constellation of elliptical orbits that permit it to selectively bias coverage and capacity to desired latitude bands and to selected local times of day or night. Thus, this system can more efficiently match designed system performance to the actual or projected market requirements with significantly fewer satellites than are needed by its circular-orbit competitors. The two inclined, elliptic, sun-synchronous planes of ELLIPSO (referred to as the BOREALIS™ planes) both have apogees that remain at a high Northern latitude, where they provide excellent coverage of all of Canada. They are also biased towards to lean towards the sunlit hemisphere of the earth, so that they provide augmented daytime coverage of Canada (as well as the rest of North America). The advanced design of the ELLIPSO system promises to bring user costs per minute down to affordable levels. The ELLIPSO system will provide mobile cellular-like communications from a hand-held phone at any location in Canada to any public-switched, cellular, or ELLIPSO phone at anywhere in Canada or overseas.

INTRODUCTION

Big-LEO mobile satellite service providers will provide wireless regional and global telecommunications. Major market segments include: (1) extension of terrestrial cellular networks in regions with low population density, (2) telephony service in regions with no access to wireless or wireline communications, and (3) international travelers visiting countries with various cellular modes/bands and unreliable/congested networks. Big-LEO service providers include Iridium, Globalstar, ICO, ELLIPSO, and Constellation. The first three received their FCC license in January 1995. ELLIPSO and Constellation (ECCO) were granted their FCC license in July 1997. All these systems rely on circular orbit constellations except ELLIPSO. By definition, circular orbits imply equal Northern Hemisphere versus Southern Hemisphere coverage, and additionally, equal day versus night coverage. ELLIPSO has two sub-constellations in three planes. See Fig. 1. The two inclined, elliptic, sun-synchronous planes are called BOREALIS. The circular-orbit equatorial plane sub-constellation is called CONCORDIA; it serves the lower Northern latitudes, the tropics, and populated areas of the Southern Hemisphere. (Elliptic orbit satellites will be added to Concordia at a later date.) In serving the Canadian market, ELLIPSO will rely primarily on BOREALIS, with some nighttime coverage by CONCORDIA of lower portions of Canada. For that reason, this paper will concentrate mostly on the two BOREALIS planes of the ELLIPSO System.

[1] The European terminology is **G**lobal **M**obile **P**ersonal **C**ommunications **S**ystem (GMPCS).

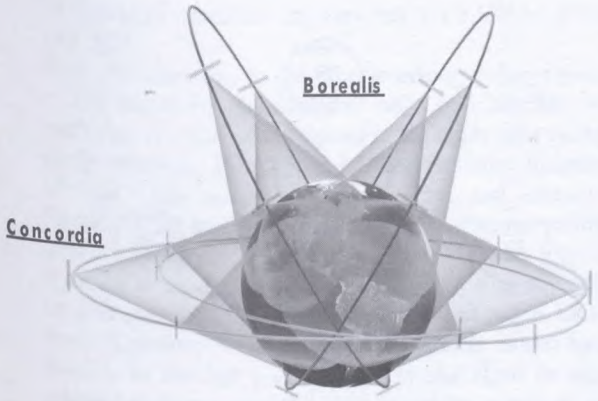


Fig. 1, ELLIPSO Constellation

ELLIPSO is unique in the group of Big-LEOs in that it employs elliptical orbits to better target the satellite cellular market. By placing the satellite apogees over more heavily populated areas during daytime hours, the system uses fewer satellites than are required by competing circular-orbit systems. Additionally, advanced design in the space segment, ground segment and user network, drawing upon recent developing technologies, assures high voice quality and seamless handovers with low probability of dropped calls. All of these factors translate directly into a lower cost per minute to the customer since fewer satellites also means fewer launch vehicles- both of which are major components in any total system cost. The rather heavy biasing of ELLIPSO towards Northern Hemisphere coverage will strongly benefit ELLIPSO customers in Canada, where they will be assured of high quality, ubiquitous coverage through at least two satellites at high elevation angles. Using a hand-held ELLIPSO phone (similar to a cell phone), customers will be able place a call from anywhere in Canada to any other telephone in-country or internationally through the Public-Switched Telephone Network (PSTN), or to any other ELLIPSO mobile phone anywhere in the world.

THE ELLIPSO CONSTELLATION

Constellation Overview

The unique ELLIPSO constellation, covered by US and foreign patents, employs a combination of elliptic and circular orbits to ensure efficient application of satellite resources to real-world communications requirements. ELLIPSO's constellation allows considerable flexibility in tailoring overall system capacity through its use of several forms of elliptic orbits. One primary consideration is the very uneven distribution of population between the Northern and Southern Hemispheres. It makes eminently good sense to bias communications capacity in favor of the Northern Hemisphere (while still maintaining adequate continuous coverage of the Southern Hemisphere). Also, it makes good sense to have extra capacity available during daytime communication peaks with lesser capacity at night when the need for communications drops off dramatically. In sum, the coverage (number of satellites in view and their elevation angles) and the capacity (throughput per satellite multiplied by the number of satellites in view) provided by the ELLIPSO constellation is optimized for both latitude and time-of-day.

BOREALIS - Northern Hemisphere Coverage:

Coverage of most of the Northern Hemisphere by ELLIPSO is accomplished by the BOREALIS sub-constellation - (you might have guessed this from its name!) BOREALIS comprises two planes, each having five elliptic sun-synchronous satellites, in approximately three-hour orbits. They are aligned so that the orbit planes are edge-on to the sun. Another way of stating this is that one has a noon ascending node, and the other a midnight ascending node. Being sun-synchronous, they maintain this condition of being edge on to the sun year in, year out. Like the Russian Molniyas, they have apogees high in the Northern Hemisphere, giving them considerable dwell time to better serve Northern Hemisphere (including Canadian) customers. On average, each Borealis satellite spends two-thirds of each orbital period in the Northern Hemisphere. Unlike the Molniyas (that are <u>not</u> sun-synchronous) BOREALIS satellites can be biased towards daylight hours due to their inherent sun-synchronicity. We do this merely by leaning the axes of both sets of these elliptical orbits towards the sun about 10°. In effect, this tilt puts the apogees more on the sunny side of the earth, thus biasing coverage towards daytime hours (local time).

The design of BOREALIS ensures that it will inherently provide superior coverage of Northern Hemisphere countries, such as Canada, Norway, Sweden, the UK, Russia, Northern China and Japan. Furthermore, the much higher minimum and average elevation angles attained in this region will permit reliable communications for users in rugged or mountainous terrain, or in heavily forested areas, as well as during periods of heavy precipitation. Wireless satellite communications typically degrades markedly at the lower elevation angles common to most of the other Big-LEO systems. A comparison of the minimum elevation angles, as a function of latitude, is revealing in this regard. See Fig. 2.

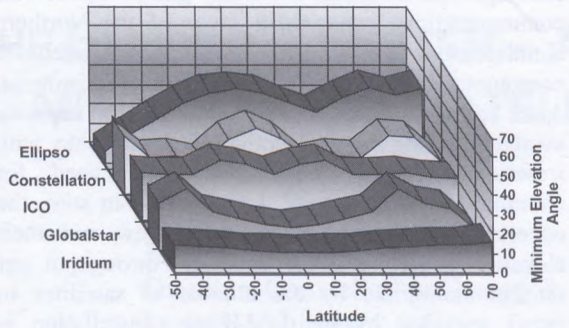


Fig. 2, Minimum Elevation Angle Comparison

CONCORDIA - Tropical and Southern Hemisphere Coverage :

In order to cover the tropics and the remainder of the Southern Hemisphere, ELLIPSO makes use of the CONCORDIA sub-constellation. The baseline, initial deployment of this plane uses seven circular orbit satellites of approximately 4.8 hour periods. This ensures both day and night coverage down to 55° South latitude. A later augmentation will be accomplished using several elliptical satellites whose motion is integrated with the encompassing circular satellites. This is the so-called CONCORDIA "Gear" array (US Patent Pending). See Fig. 3. The elliptical portion of this array employs the Apogee Pointing To the Sun (APTS) concept. The total combination will then provide both tropical and Southern Hemisphere coverage of populated areas, with augmented coverage during daylight hours. It will still not, however, provide coverage of Antarctica (impossible for equatorial plane satellites).

CONCORDIA can provide lower elevation angle coverage of the southern portions of Canada. However, ELLIPSO operators will generally use the BOREALIS satellites for coverage above 30° North latitude during daytime, and above 40° North during nighttime.

ELLIPSO COVERAGE IN CANADA

For reasons just described, the BOREALIS satellites will provide the vast majority of the coverage service to Canadian customers. They will cover Canada ubiquitously, filling in all regions not covered by existing terrestrial cellular service. The high elevation angles and long dwell time of the satellites will ensure single satellite connectivity between the more densely populated areas of Canada with the vast areas of the country presently uncovered by wire-line or terrestrial cellular services.

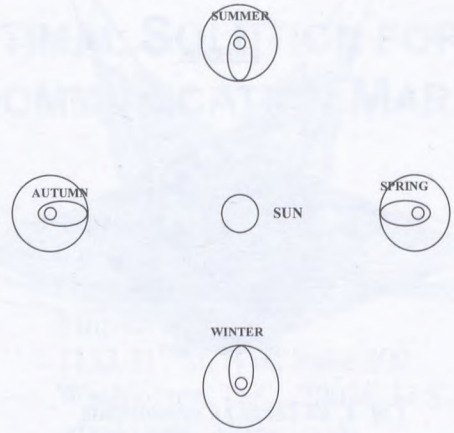


Fig. 3, ELLIPSO GEAR Array
(US Patent Pending)

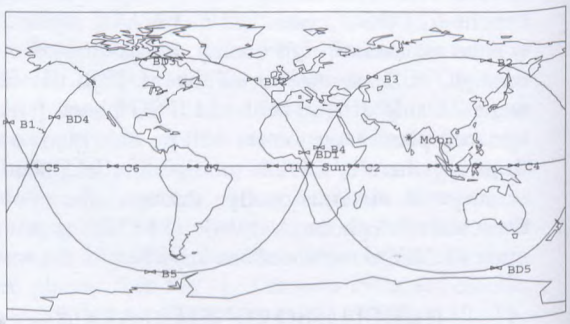


Fig. 4, BOREALIS Ground Tracks;
6 a.m. Ottawa Time

The satellite sub-points (tracks) for the two BOREALIS rings are shown on a Cartesian map in Fig 4. The satellites, being in retrograde 116.5° orbits, move from East to West. Those satellites marked "BD" are in the noon ascending node ring (note the position of the sun). Those satellites marked "B" have a midnight ascending node. At the time shown in Fig. 5 (about 6 am local Ottawa time) there are two

BOREALIS satellites in view of OTTAWA- BD2 and BD3.

Note also that, due to the elliptic orbits with perigees in the Southern Hemisphere, only one satellite in each ring is south of the equator; the other four being in the Northern Hemisphere! Since Borealis apogees of 7605 km occur at about 60°N, the satellite footprints in this region are quite large, covering approximately one-fifth of the earth's surface. Figure 5 shows an azimuth-elevation (az-el) view as seen by an observer looking up from Ottawa at noon local time. ***[Caution: North is up, and South is down; but, East is to the left and West is to the right in this ground-to-space view!]*** Four satellites are now in view, all at rather high elevation angles. Also shown is the locus of all the geostationary satellites, as well as the locus of the equatorial CONCORDIA satellites. None of the satellites are near either the ring of geostationary satellites, or the ring of CONCORDIA satellites. Thus, no inter-satellite interference is possible. At this point, it might be noted that at local midnight, it was found that only two satellites were in view from Ottawa, thus demonstrating the relative capacity bias on favor of day-time.

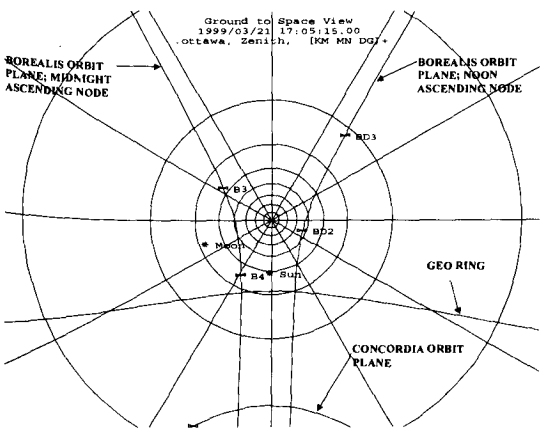


Fig. 5, Az-El View (Center is at zenith)
Ottawa local noon

A 24-hour time plot at Ottawa for all ten BOREALIS satellites' elevation angles, is shown in Fig. 6. Satellites will pass directly overhead when the BOREALIS orbit plane contains the geocentric radius of Ottawa. This will occur twice each day for each plane. The two peaks in elevation angles for each plane are clearly shown in Fig. 6. Some idea of the envelope of minimum elevation angle (to the highest satellite) can be gained by a closer study of this Figure.

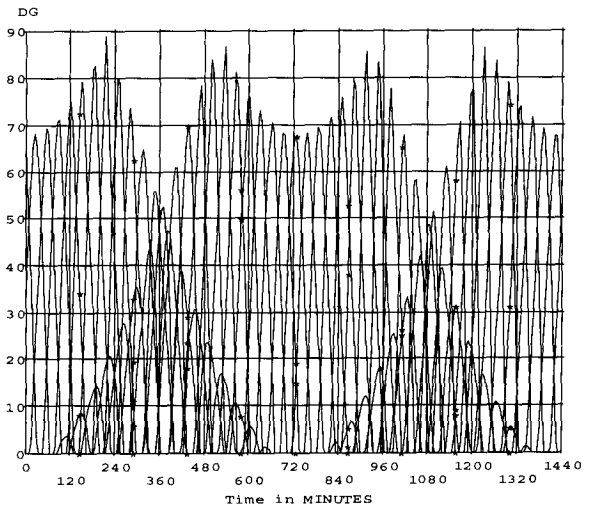


Fig. 6, Elevation Angles to BOREALIS Satellites from Ottawa, over a 24-hr period

If a Canadian ELLIPSO user is located farther North near Yellowknife, at 62° 29' N, coverage is similar to that seen at Ottawa. See Fig. 7. Also, note that he will now be always able to access satellites in both of the two BOREALIS planes. In addition, since his latitude almost matches that of the Borealis apogees, there will be only one peak per day per Borealis ring. Looking North over the pole towards the opposite BOREALIS ring satellites, he will find their elevation angles can fall as low as 7 or 8 degrees. Still, he will have at least two, and sometimes three, satellites from the nearest ring visible at much higher elevation angles.

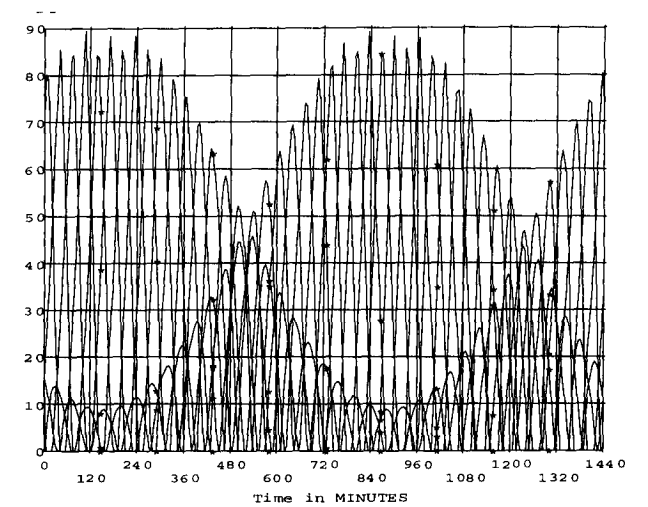


Fig. 7, Elevation Angles to BOREALIS Satellites from Yellowknife, over a 24-hr period.

On the average, two to four satellites will be in view from any particular location within Canadian territory. Both the average and the minimum elevation angles to these usable satellites will be quite high, ensuring high quality voice transmissions, not adversely affected by terrain or atmospherics. In addition, due to their lower velocities near apogee and their higher operating altitudes, the BOREALIS satellites will remain in view for longer periods of time and the frequency of handovers is greatly reduced.

A final chart showing the minimum and average elevation angles for both BOREALIS and CONCORDIA over a two-week period is shown as Fig. 8.

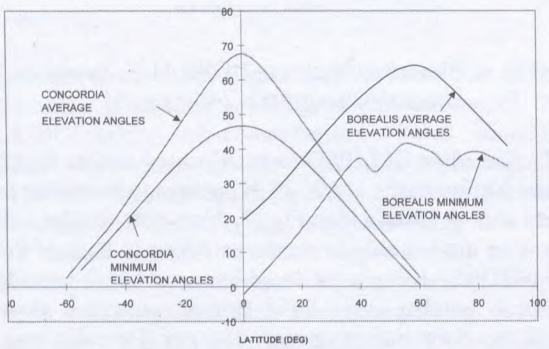


Fig. 8, Minimum and Average Elevation Angles as a Function of Latitude

THE ELLIPSO SATELLITE

All ELLIPSO satellites (whether for BOREALIS or CONCORDIA) are identical in design. They all enjoy relatively small beta angles[1] of less than 23.5° . This is because CONCORDIA lies in the equatorial plane, while BOREALIS is sun-synchronous with noon and midnight ascending nodes. This use of a single design saves money in design and manufacturing costs.

The Ellipso satellite is a three-axis stabilized vehicle using advanced composite materials in its structure. Its mass is approximately 1350 kg, and it derives its power from two large solar array panels having a total area of approximately 40 square meters. See Fig. 9. The satellite carries a communications payload designed to furnish cellular telephone like service, using broad-band CDMA signals. The user up-link(from the hand-held or fixed user terminal to the satellite) operates in the L-Band at 1610-1621 MHz. The user down-link from the satellite operates in the S-Band at 2484 to 2500 MHz.

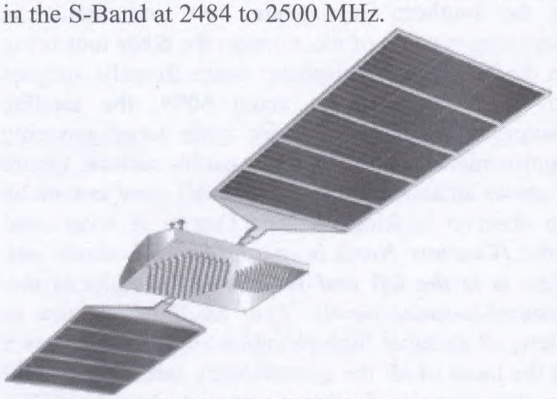


Fig. 9, ELLIPSO Satellite

ELLIPSO'S ECONOMICS AND MARKET

Ellipso will provide wireless domestic, regional and international connectivity. Major market segments include: (1) extension of terrestrial cellular network services in regions with low population density, (2) telephony service in regions with no access to wireline communications, and (3) global wireless connectivity to international travelers visiting countries with differing cellular modes/bands and unreliable networks.

Coverage – Most of the Canadian territory remains uncovered by terrestrial infrastructure. Canada's northern provinces, prairies, mountainous and forested lands, offshore fishing areas, mining sites, offshore oil surveying, drilling, and producing sites, extend through vast distances. These areas have a low population density and the cost per user/line of laying cables or building out cellular networks is prohibitive (Figure 10 highlights the gaps in Canadian cellular coverage and its concentration in metropolitan areas around the United States and coastal borders.) Ellipso satellite footprints will cover the entire Canadian territory, providing seamless connectivity to users traveling or residing outside of cellular and wireline coverage. Ellipso satellites will additionally provide high quality transmission even in forested or mountainous areas: the satellites will hover over the user at a high elevation angle for a long period of time, reducing the likeliness of signal obstruction.

[1] Beta Angle is defined as the angle between the orbit plane and the sun line.

Fig. 10, Canada's Cellular Coverage

User Profile – By providing universal and seamless connectivity, ELLIPSO will serve travelling as well as semi-fixed and fixed communities. These include: (a) residents of remote towns, hinterland and offshore mining and exploration sites, timbering sites, pipeline monitoring centers; (b) outdoor tourists, mountaineers, hunters, campers; (c) geological and ecological monitoring, and exploration centers; (d) park, emergency, health, search and rescue, and security forces, as well as the RCMP and the Canadian Forces; (e) cross-country transportation fleets, as well as maritime, Coast Guard, and fishing fleets, and (f) commercial, general, and private aviation. ELLIPSO will enable customers to place secure emergency, personal, and professional calls from anywhere in Canadian territory, and will connect them to a switch from where their call will be routed locally, nationally, or internationally.

Services – ELLIPSO will offer two baseline services: mobile and fixed. Mobile services will be provided via a multi-mode handset, similar in size, weight, and functionality to state-of-the-art cellular phones. The mobile handset will be mounted on vehicles with a rooftop external antenna for improved signal quality, and a vehicular cradle for hand-free usage. Ruggedized versions will be available for maritime and military uses, in compliance with all relevant standards. Fixed and semi-fixed services will be provided through public and private phones. Ellipso public payphones will be solar-powered, coin or card-operated, and will provide a single or multiple lines in a phone booth. Ellipso private residential or business phones will consist of a rooftop-mounted antenna wired to building phone jacks to which regular phones will be connected. A ruggedized transportable version to be used for moving sites will also be available.

Cost of Service – As ELLIPSO extends the reach of the terrestrial infrastructure, the cost to the end-user of this last, or first, mile of a call will come at a premium over terrestrial rates, *assuming terrestrial options were available,* and be added to PSTN tail charges.

Mobile Services – ELLIPSO plans to wholesale its mobile services at CAN$0.71/mn on average (including monthly fee) to service providers. Retail price before landline and tax charges is estimated at CAN$0.92/mn on average (including monthly fee). As a benchmark, Telus Mobility's median airtime rate is CAN$0.18/mn (including monthly fee and before landline and tax charges.)

Fixed Services – ELLIPSO plans to wholesale its fixed services at CAN$0.12/mn on average to service providers. Retail prices before PSTN tail charges and taxes are estimated at CAN$0.18/mn on average. As a benchmark, Telus's Domestic Direct Dial (DDD) rates for Canada range from CAN$0.05/mn to CAN$0.56/mn.

It is expected that these last or first mile costs are affordable to the user who has limited or no alternative option available to him. They are made possible by ELLIPSO's BOREALIS planes, which are uniquely designed to efficiently serve the Canadian territory. Initial service by ELLIPSO BOREALIS satellites is expected by 2003.

CONCLUSIONS

In conclusion, the design and implementation of the ELLIPSO GMPCS provides efficient, moderately priced, satellite cellular telephone coverage to all parts of Canada. The efficient use of elliptical sun-synchronous orbits with apogees at approximately 60° North latitude ensures satellite visibility at high angles of elevation. This translates directly into efficient, ubiquitous coverage in difficult mountainous terrain, in fairly densely forested areas, and in conditions of bad weather. Moreover, the biasing of the ELLIPSO BOREALIS satellites towards daytime hours ensures that there will be augmented circuit capacity in the day, when it is most needed. Besides the many Canadian civilian and business uses for ELLIPSO services, appreciable military, police and government market segments may also be well served by ELLIPSO BOREALIS satellites after the turn of the century.

The Role of Mobile Satellite Systems in Third Generation Wireless Mobile Communications

Leslie A. Taylor, Roger LeClair
Leslie Taylor Associates
1333 H Street, NW, Suite 1100W
Washington, DC 20005
E-mail: ltaylor@lta.com

ABSTRACT

This paper reviews the proposals for second generation Big LEO systems, including a description of the role of these systems in third generation mobile communications. Spectrum requirements, standards issues and network architecture issues are examined.

INTRODUCTION

The growth of wireless mobile communications has exceeded the most ambitious estimates, with a projected 285 million users worldwide at year-end 1998. Mobile satellite systems have had a small share of this market, with more than 400,000 users globally. However, the advent of satellite cellular systems (e.g., Big LEO systems) beginning in the fall of 1998 demonstrates the utility of mobile satellites in extending the range and utility of wireless mobile communications. For example, the satellite cellular systems will allow users to seamlessly utilize mobile communications globally, using terrestrial mobile systems where available, and accessing the satellite systems when the terrestrial signals are not present. These systems also will provide an important mechanism for instant telecommunications infrastructure in locations where neither terrestrial wireline nor wireless facilities exist. The satellite cellular systems will work compatibly as a component of second generation (e.g., digital) mobile systems.

Third generation wireless mobile communications systems (terrestrial) are under development. These systems will provide higher data rate connectivity to e-mail and the Internet, as well as interactive services such as videoconferencing. The satellite cellular systems are already planning enhanced second generation systems that will be a full-fledged component of third generation wireless, making mobile wireless Internet connectivity a reality on a global basis.

SPECTRUM AND SERVICES FOR FIRST GENERATION MSS

Mobile satellite (voice) services are currently available from Iridium (low earth orbit, (LEO)), and from geostationary systems Inmarsat, American Mobile Satellite, TMI, Solidaridad and OPTUS. LEO systems under construction include Globalstar, ICO Global Communications, Constellation Communications and Mobile Communications Holdings Inc. Ellipso. Their scheduled implementation dates and data rates are presented below.

First Generation LEO MSS Systems

System	Date	Data rates
Iridium	1998	9.6 kbps
Globalstar	1999	9.6 kbps
ICO	2000	38.4 kbps
Ellipso	2002	9.6 kbps
Constellation	2003	Not indicated

There are also a number of regional mobile satellite systems either under construction or in the planning stages; these systems will operate using geostationary satellites, and are called Regional GSOs (Geostationary Satellite Orbit). Examples include ACeS (Asia), EAST (Europe, Middle East and parts of Asia), and Thuraya (Middle East, Europe and parts of Asia). [1]

The frequencies being used by these mobile satellite systems are noted in the table below.

Frequencies	Direction	Current or planned use
1610-1621 MHz	Earth to space	Globalstar, Constellation, Ellipso
1621-1626.5 MHz	Earth to space and space to earth	Iridium
2483.5-2500 MHz	Space to earth	Globalstar, Constellation, Ellipso
1980-2010 MHz*	Earth to space	ICO, US 2 GHz applicants
2160-2200 MHz	Space to earth	ICO, US 2 GHz applicants

Frequencies	Direction	Current or planned use
1626.5-1660.5 MHz	Earth to space	Inmarsat, Am. Mobile, ACeS, other regional GSOs
1525-1559 MHz	Space to earth	Inmarsat, Am. Mobile, ACeS, other regional GSOs
2670-2690 MHz	Earth to space	N-Star, many terrestrial systems
2500-2520 MHz	Space to earth	N-Star, many terrestrial systems

*1990-2025 MHz/2165-2200 MHz in the US

The first generation mobile satellite systems operating in low earth orbit will provide data rates only up to approximately 9.6 kbps. However, ICO has announced plans to expand its service offering to include rates up to 38.4 kbps. As will be seen below, for their second generation systems, the Big LEOs plan to move into the 2 GHz or other frequency bands and dramatically improve their data rate capabilities.

THIRD GENERATION WIRELESS SERVICE DEFINITIONS

International Mobile Telecommunications 2000 (IMT-2000) will provide access to the global telecommunications infrastructure through both satellite and terrestrial systems, serving both fixed and mobile users in public and private networks. It is being developed on the basis of the "family of systems" concept, a federation of systems providing IMT-2000 service capabilities to users of all family members in a global roaming offer. [2]

Key features of IMT-2000 include a high degree of commonality of design worldwide; compatibility of services within IMT-2000 and with fixed networks; use of a small terminals; and capability for a wide range of services.

For the terrestrial component of IMT-2000, the International Telecommunication Union (ITU) has identified various services within the IMT-2000 menu. Services are expected to include "toll" quality voice, simple messaging (rates up to 14 kbps) and switched data (rates up to 64 kbps). In addition, asymmetrical multimedia services are expected to available, including high speed service with a bit rate of 2000 kbps in one direction and 128 kbps in the other. Another service will operate at 384 kbps in one direction, and 64 kbps in the other. Applications include file download, Internet browsing, full motion video and non-interactive telemedicine.

Symmetrical services would include high fidelity audio, video conferencing, telemedicine and various video conferencing applications.

The ITU has identified four broad categories of mobile satellite services in the IMT-2000 offering, including voice services (likely to meet "toll" quality expectations), messaging services (enhanced over current capabilities to include extended message length, better delivery rates, and two-way paging), medium speed multimedia services, up to 144 kbps, and highly interactive multimedia services. The latter are symmetric applications requiring the highest available speeds, up to 384 kbps. [3]

Satellite "Third" Generation Wireless Services

Service	*Bit Rate Required*
Voice	Up to 8-16 kbps
Messaging	Up to 40 kbps
Medium multimedia	Up to 144 kbps
Remote offices*	144 kbps
Highly Interactive Multimedia**	384 kbps

*Access Intranets, downloads, electronic commerce.
**Symmetric high-speed connections including image transfer and video.

POTENTIAL MARKET FOR IMT-2000 SERVICES

When WARC-92 identified spectrum for third generation wireless services (see below), voice service was considered to be the major source of traffic. Now, however, with the introduction of the Internet and skyrocketing demand for multimedia services, IMT-2000 has been redefined as a range of services, including voice, data and multimedia applications. While demand for voice services will continue to grow over the next several years, by 2002/2003, data services are expected to comprise an increasingly larger percentage of overall wireless traffic. Higher data rate services are expected to grow at a rapid rate between 2005-2015.

SPECTRUM FOR THIRD GENERATION WIRELESS

In 1992, the World Administrative Radio Conference identified (but did not allocate) 230 megahertz of spectrum (1885-2025 MHz/2110-2200 MHz) that could be used by administrations wishing to implement third generation wireless systems, IMT-2000. The spectrum is allocated to several services, including fixed, mobile and mobile satellite services (MSS). The intention was to enable administrations to plan for implementation of future generation wireless communications services in these frequency bands. A portion of these bands (1980-2010 MHz/2170-2200 MHz) was allocated internationally for MSS.

The table below shows the current allocations where the IMT-2000 frequencies overlap with Mobile Satellite allocations.

THIRD Generation Spectrum, Terrestrial/Satellite

	Terrestrial, MHz	MSS, MHz
ITU	1885-2025/2110-2200	1980-2010/ 2170-2200
CEPT	DECT:1880-1900 UMTS:1900-1980/ 2110-2170	1980-2010/ 2170-2200
Japan	PHS:1895-1918 IMT-2000: 1918-1980/ 2110-2170	1980-2010/ 2170-2200
US	PCS:1850-1990 No spectrum identified for 3rd gen. wireless	1990-2025/ 2160-2200

Note: the only common spectrum which includes the US is 1990-2010 MHz (earth to space) and 2170-2200 MHz (space to earth) for MSS.

While Europe and Asia have largely observed the WARC-92 designations for IMT-2000, the US, in 1994, allocated 1850-1990 MHz for second generation wireless personal communications services (PCS) to be licensed through competitive bidding for the right to implement service in discrete service areas. [4] No national licenses were made available. This had the result of rendering unusable, in the US, a portion of the international MSS allocations. The US allocations for MSS are 1990-2025 MHz/2165-2200 MHz. [5] This could significantly reduce the spectrum available for global MSS systems, unless they are designed to operate in frequency bands that vary by region.

In addition, the implementation of PCS in geographically discrete areas by multiple licensees raises questions about whether and how third generation wireless systems will be implemented in the US, even terrestrially.

As terrestrial wireless has matured, the benefits of utilizing common frequency bands and a global standard have become evident. Users of GSM wireless systems, operating in the 1800 MHz band, can roam to 90 or more countries. However, US users of digital or analog cellular or PCS (including AMPS, GSM and CDMA standards) cannot readily roam beyond the US. This marketing aspect may work to the advantage of the global mobile satellite systems such as Iridium, Globalstar and ICO, who may find an eager market segment consisting of US travelers to overseas destinations and non-US travelers coming to the US, who wish to use a single phone number. Smart cards provide solutions to some users but these are not widely implemented as of 1999.

Unfortunately, now that the US has gone its separate way with the PCS allocations and licenses, the stage is set for the US isolating itself once again when third generation wireless is implemented. Ironically that could again create opportunities for mobile satellite service providers who will provide at least some of the enhanced features and functions of third generation wireless.

In Europe, access to certain of the MSS frequencies is restricted. The CEPT band plan, as currently reflected in the CEPT decisions, does not provide an opportunity for US space segment licensees/applicants for global systems in the 2 GHz licensing round to obtain access to 2 GHz MSS spectrum Europe before 2005. [6]

REGULATORY SITUATION – THE ITU

At the ITU, Study Group 8/1 is exploring the spectrum requirements for IMT-2000. The goals of the group include improving capabilities of wireless systems by supporting greatly increased bit rates and use of common frequency bands and identified standards to facilitate global roaming. Within TG 8/1, radio transmission technology (RTT) (air interface) proposals have been submitted. Ten terrestrial RTT proposals have been submitted and there are six satellite RTT proposals. Below is a chart of the satellite RTT proposals:

Satellite TG 8/1 RTT Proposals

Proposal	Description	Service
SAT-CDMA	49 LEO satellites, 2000 km alt.	S. Korea TTA
SW-CDMA	Satellite wideband CDMA	ESA
SW-CTDMA	Satellite hybrid CDMA/TDMA	ESA
ICO RTT	10 MEOs, 10,390 km. Alt	ICO Global
Horizons	Horizons satellite system	Inmarsat
INX	Iridium Next Gen.	Iridium

With regard to the terrestrial RTT proposals, Europe is primarily promoting UTRA, which is an evolution of GSM and WCDMA that has been developed by Ericsson. The US has submitted several RTTS, but its cdma2000 proposal, developed by Qualcomm, has generated the most controversy. Technical groups are now meeting to develop a compromise solution to accommodate both CDMA approaches.

From the standpoint of TG 8/1, most, if not all of the proposed RTTs, both terrestrial and satellite, are likely to be reflected in its final recommendation. However, pursuant to ITU policy, the submitting entities must either waive IP rights or agree to license the IP to others on reasonable terms. This is an element in the conflict between the Ericsson and Qualcomm proposals as both companies have suggested they may withdraw their RTTS if no consensus is reached on the CDMA standard. Based on second generation wireless trends, it is reasonable to assume that one TDMA standard and one CDMA standard will predominate in third generation systems.

The satellite situation is somewhat outside this debate as the satellite systems now being implemented are able to operate using the proprietary IP of the satellite system and switch to one or more terrestrial systems, standards and frequencies as needed. So the main concern for potential satellite systems which plan to participate in third generation wireless is whether some administrations will require that their RTT be contained in an ITU recommendation as a prerequisite for licensing authority in that country. While the two may have little or no relationship, many administrations use the ITU processes and recommendations as a proxy for their own evaluation. This is especially the case when an administration lacks sufficient resources to evaluate technically a proposed system.

While RTTs were required to be submitted in 1998, it is possible the process will be extended or reopened and other satellite RTTs permitted to be filed. This would appear reasonable as the first generation personal communications satellites are just now being implemented and system operators have had little opportunity to look ahead to their second generations.

REGULATORY SITUATION: THE US

In the US, the Federal Communications Commission (FCC) has initiated a proceeding soliciting input from industry on spectrum requirements for both the terrestrial and satellite components of IMT-2000. [7] Contributions to the proceeding were received in fall 1998, and the FCC is studying the contributions in preparation for developing the US position on spectrum requirements.

SECOND GENERATION MSS PROPOSALS IN THE US

Apart from its solicitation of views on IMT-2000, the FCC also is conducting a rulemaking to develop licensing procedures and service rules for provision of MSS in the 2 GHz bands. [8] A processing mechanism is required to evaluate applications and letters of intent filed by nine companies to use the 2 GHz frequencies for mobile satellite applications. [9] [10] Included in the filings are applications from Iridium, Constellation, Mobile Communications Holdings, Inc., and Globalstar to launch second generation mobile satellite systems. In addition, ICO Global Communications filed to provide its first generation mobile satellite services in those frequencies. Geostationary system operators Inmarsat and TMI (Canada) filed notices of intent to provide second generation mobile satellite services using higher-powered, next generation (geostationary) satellites. The Boeing Company filed an application to build and launch a satellite-based air navigation and communications system.

The chart below summarizes the applications.

SYSTEM.	SAT/ ORBIT	SERVICES
Am Mobile	1/GEO	Mobile voice and data, point-to-multipoint applications, data up to 384 kbps
Boeing	16/ MEO	Aeronautical navigation
Celsat	1/GEO	Mobile voice, data, paging, messaging, video
Constell.	46/ LEO	Mobile voice, data up to 28.8 kbps, multimedia
Globalstar GS-2	64/ LEO 4/MEO	Voice, data, paging, data up to 144 kbps
Globalstar GS-40	80/ LEO	Cellular trunking, LAN, Internet connections
ICO	10/ MEO	First generation voice and data
Inmarsat Horizons	4/GEO	Data up to 144 kbps with small terminals
Iridium Macro-Cell	96/ LEO	Voice, data up to 384 kbps
MCHI Ellipso 2G	26/ LEO	Voice, data up to 64 kbps
TMI	1/GEO	Voice and data, including Internet

American Mobile Satellite, currently licensed to provide mobile satellite services in the US using the L-band frequencies, proposed to launch a more powerful satellite to offer enhanced services, including the use of non-directive antennas. American Mobile expects to offer higher-speed packet data services for connections to the Internet or other data networks, offering transmission rates up to 384 kbps. The American Mobile system would be capable of providing service in the Western Hemisphere.

The Boeing Co. proposed to launch a 16 medium earth orbit satellite system to provide a range of communications, navigation and surveillance air traffic management services, on a global basis.

The Celsat system, comprised of 3 geostationary satellites, would offer voice, data, message, image and video services at variable bandwidths up to 144 kbps.

The Constellation system would be comprised of 46 satellites, in low earth orbit, offering high speed digital transmission services at rates up to 28.8 kbps, including high speed file transfer, Group 3 facsimile, Internet access and other multimedia services on a global basis. Constellation is in the process of building its first generation mobile satellite system, and plans to integrate its two satellite systems. The ground network will include satellite control stations and gateways located around the world.

Globalstar filed applications to build a second generation mobile satellite system using either of two frequency bands, the 2 GHz and/or the V-band frequencies. The GS-2 system, operating at the 2 GHz frequency, would consist of 64 low earth orbiting and 4 geostationary satellites. The system could provide global voice and mobile data communications, facsimile, tracking and monitoring, position location and paging services. The terminals, which will include handheld, transportable and fixed units, will be designed to support various data rates up to 144 kbps. The handheld and vehicle-mounted terminals will use quasi-omni-antennas, and the fixed terminals will use high gain directive antennas. The ground network will include satellite operations centers and gateways

The V-band system, GS-40, will include 80 satellites in low earth orbit, providing an array of services including trunking and Internet access. The GS-40 is seen to augment Globalstar's first generation system, currently being launched.

ICO Global Communications is building a 12-satellite, medium earth orbiting system to provide voice and data services, on a global basis, to mobile and fixed terminals. It plans to launch its first satellite later this year, and to begin providing satellite-based communications services in 2000. The ICO ground segment includes 12 satellite access nodes, located around the world, interconnected with a fiber optic network.

Inmarsat, the global mobile communications entity, is planning to launch four satellites to geostationary orbit to provide personal multimedia communications and high speed services. The ground network would include access networks linking to points of presence in each country. Subscriber terminals would have an antenna size of 0.25 meters, and the high speed data terminal would have an antenna of 0.75 meters.

The Iridium Macrocell system will consist of 96 low earth orbiting satellites. The system will, according to Iridium, be interoperable with IMT-2000, and will provide voice, e-mail retrieval, fax transfer, video and Internet connectivity and one- and two-way messaging services. A range of data rates are expected to be offered, including from 4.6 kbps (for basic voice) up to 384 kbps. The ground segment will include the satellite operational center and local gateways.

Mobile Communications Holdings, Inc. plans to build a second constellation of 26 satellites; as is the case for MCHI's first generation, the Ellipso 2G will operate in elliptical orbit. The Ellipso 2G system will offer data transmission at speeds up to 64 kbps and Internet access. The ground segment includes a satellite control center, several regional network control centers, ground control stations and switching offices.

TMI Communications plans to launch a single geostationary satellite to provide sophisticated mobile satellite communications in North America. It plans to offer a mix of services, including voice, circuit-switched and packet-switched data, fax and paging. Internet connections will be available.

The Commission is considering a variety of options for handling the multiple applicants, including dividing up the spectrum a priori, licensing all the systems and allowing negotiations after licensing, or even after launch, and spectrum auctions.

NETWORK ARCHITECTURE

Plans for the network architecture for the second generation Big LEOs will be similar to those already being built for the first generation. The ground segment includes satellite operations centers, located where necessary, which will have a certain number of antennas and associated equipment to communicate with the satellites. A network control center will monitor and control the entire network, and will be interconnected with the gateways located strategically around the world.

The gateways handle call interconnection to and from the terrestrial network, either fixed or wireless, and route the call through the satellite network.

The gateways are responsible for mobility management (tracking subscribers and routing calls to and from them), call set-up, maintenance and tear-down, security and fraud management and network operations. The gateways also provide the interface to the terrestrial networks, including the Public Switched and Public Data Networks.

INTERCONNECTION WITH TERRESTRIAL NETWORKS

The gateways of MSS systems provide the interface between the ground segment and the switch of the local network operator. Interfaces will be provided to the Public Switched Telephone Network and data networks.

Assuming that the Big LEO operators will use their first generation ground segment as the basis upon which to integrate the second generation services, the gateway interface will be enhanced to meet the new service requirements. This may mean adding additional switching equipment, as well as an Internet node, to provide appropriate interconnections. Various subsystems may be installed, depending on service requirements, including switching subsystems for circuit services, and packet services subsystems and/or messaging subsystems for data services.

As the table below shows, each of the system operators will be using a variety of modulation techniques in their second generation MSS:

Modulation Techniques, 2nd Gen MSS

System	Modulation
Constellation	CDMA
Globalstar	CDMA, TDMA, FDMA
ICO	TDMA
Horizons	TDMA
Iridium	CDMA, TDMA
MCHI	CDMA

The systems intend to be interoperable with third generation digital cellular standards, which will require the appropriate switching equipment located at each gateway to perform the conversion between the satellite system and the cellular network. For calls going into the fixed public network, a different conversion process is required.

TECHNICAL STANDARDS

In the first generation mobile satellite systems, all systems are being built to be compatible with the GSM digital cellular network, as it is a global standard. In addition, the system operators are specifying user terminals that will be compatible with other second generation wireless standards such as D-AMPS and CDMA. None of the current generation of Big LEO systems is compatible with each other. Each of the five systems has developed a proprietary air interface and specifications for handsets. As an example, Iridium handsets will not work on Globalstar, nor will the ICO handset operate with the Constellation or MCHI systems.

The situation will not change in the second generation LEO mobile satellite systems. Each satellite operator is again developing proprietary air interfaces. Global roaming within each system will be possible, but roaming between systems will not.

Two of the regional geostationary mobile satellite systems, ACeS and EAST, have agreed upon a common air interface to permit seamless roaming between both systems. However, neither system has announced plans regarding its approach to providing third generation wireless services.

CONCLUSION

The dramatic growth in Internet use has changed the way we work, learn, access information and entertainment and interact. To date, ready access to these capabilities has been restricted to those with fixed communications and a PC. Third generation wireless envisions a world in which access to the Internet, to e-mail, and the ability to take advantage of multimedia communications will be possible regardless of the subscriber's proximity to a fixed communications port.

To extend those opportunities to every location on earth, mobile satellite services will be key. Second generation mobile satellite systems will play a crucial role in providing ubiquitous coverage, making mobile wireless Internet connectivity a reality on a global basis.

[1] For details on the systems, see Leslie Taylor Associates' The Complete Book on Mobile Satellites: Systems, Services and Markets, Bethesda, Phillips Business Information, Inc., 1999, Chapters 9-12.
[2] ITU IMT-2000 web site, at www.itu.int.
[3] US IMT-SPEC Working Document Revision 5.0, Draft Contribution, Document 8-1/USA98-26, Working Document Towards a Draft New Report ITU-R M. [IMT.SPEC.], Spectrum Requirements for IMT-2000, 8 October 1998, Table 10 at 20.
[4] Amendment of the Commission's Rules to Establish New Personal Communications Services (PCS Proceeding), GEN Docket No. 90-314, Memorandum Opinion and Order, 9 FCC Rcd 5947 (1994).
[5] Amendment of Section 2.106 of the Commission's Rules to Allocate Spectrum at 2 GHz for Use by the Mobile Satellite Service, ET Docket No. 95-18, First Report and Order and Further Notice of Proposed Rule Making, 12 FCC Rcd 7388 (1997).
[6] CEPT decision ERC/DEC/(96)SS.
[7] Public Notice, Commission Staff Seek Comment on Spectrum Issues Related to Third Generation Wireless/IMT-2000, Report No. IN 98-48, August 26, 1998.
[8] FCC Adopts Notice Proposing Policies and Services Rules for the 2 GHz Mobile Satellite Service, IB docket 99-81, Report No. IN 99-12, March 18, 1999.
[9] Public Notice, Report No. SPB-119, March 19, 1998.
[10] For details on the systems, see LTA study, Op. Cit., Chapter 13.

The ICO System for Personal Communications by Satellite

Dr. N. Bains
ICO Global Communications,
1 Queen Caroline Street,
London W6, United Kingdom
Email : nav.bains@ico.com

THE OPPORTUNITY

By the end of this decade more than 80% of the world's land surface, and about 40% of its population, is likely to be without cellular coverage. There will also be many different cellular standards, both analogue and digital, and even when standards are common, there may be limitations on roaming between different systems. The result is a non-uniform service availability for the customer. The gradual introduction of new digital standards will only partially relieve this, and until well into the next century there will be many situations where a single-standard cellular/PCS phone will not obtain service everywhere.

The fragmentation of standards and the remaining unserved geographical areas represent an opportunity for satellite phones. ICO does not believe that a satellite handheld phone service will ever be deliverable at prices competitive with well-designed terrestrial systems, but it will form an ideal complement for use in regions where terrestrial services are not compatible with the user's home region, or where there simply is no terrestrial coverage.

SUMMARY OF ICO SOLUTION

ICO is a satellite-based mobile communications system designed primarily to provide services to handheld phones. The system will offer digital voice, data, facsimile, and a suite of messaging services anywhere in the intended environments world-wide.

Overview

The system design integrates mobile satellite communications capability with terrestrial networks and employs, among others, handheld mobile telephones which offer services similar to normal cellular phones, in outdoor environments. It will route calls from terrestrial networks through ground stations (called Satellite Access Nodes or "SANs") which will select a satellite through which the call will be connected. Calls from a mobile terminal will be routed via the satellite constellation to the appropriate fixed or mobile networks or to another mobile satellite terminal. Handsets will be produced by major manufacturers of telecommunications equipment, benefiting from terrestrial cellular/PCS technology. Single-mode satellite-only versions will be available, but most are expected to be capable of dual-mode operation with both satellite and terrestrial cellular/PCS systems. Dual-mode handsets will be able to select either satellite or terrestrial modes of operation automatically or under user control, subject to the availability of the satellite and terrestrial systems and the user's preferred service arrangements.

The Space Segment

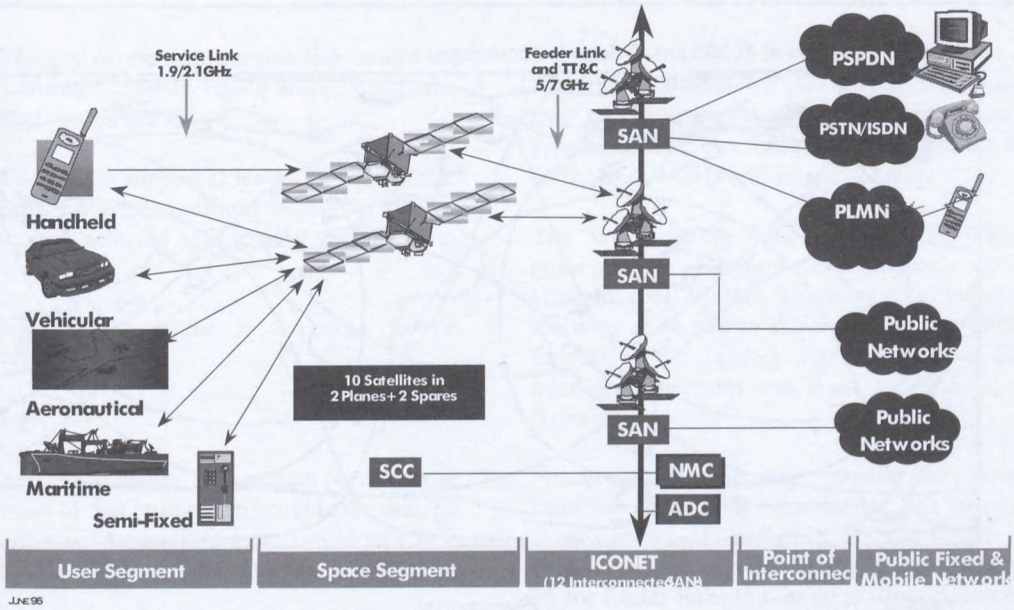
A constellation of 10 satellites in medium earth orbit (MEO), 10,390 km above the earth's surface will be arranged in two planes of five satellites each, with one spare satellite in each plane (i.e. 12 in orbit). Hughes Space & Communications International, Inc., is currently building the satellites under a contract signed in July 1995.

- 10 operational satellites in two orbital planes; 10,390 km (6 hours) orbit
- Each orbital plane inclined 45 degrees to equator
- One spare satellite in each plane; making 12 total launched
- 12 Satellite Access Nodes located globally

ICO Satellite Constellation

The configuration has been designed to provide coverage of the entire surface of the earth at all times and to maximise the path diversity of the system. Path diversity is the availability to a user of more than one satellite at the same time, and provides an alternative path for transmission in case one satellite is obstructed, increasing the likelihood of uninterrupted calls.

The satellites will be linked to a ground network (the ICONET) which will interconnect twelve SANs located throughout the world. SANs comprise earth stations with multiple antennas for communicating with satellites, and associated switching equipment and databases. The ICONET and SANs will implement the selection of call routings to ensure the highest possible quality and availability of service to system users. Gateways are the points of interconnection between terrestrial networks and the ICONET, and are located throughout the world.



ICO System Overview

SELECTION OF ORBITAL CONFIGURATION

The generally known technically feasible options of providing a communication service to handheld satellite phones are:

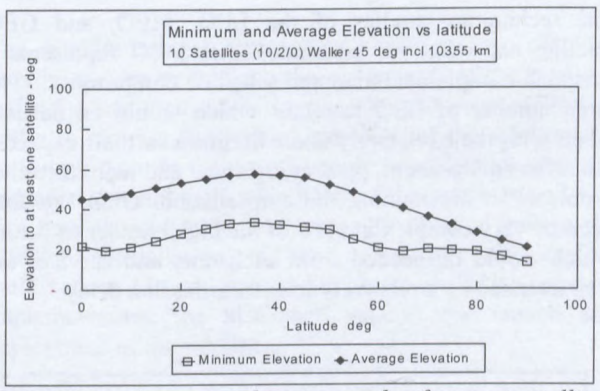
(i) low earth orbit (LEO - up to 2,000 km altitude);
(ii) medium earth orbit (MEO - 8,000 to 20,000 km); or
(iii) geostationary orbit (GEO).

To cover the earth fully, LEO requires around 40-70 satellites, MEO needs 6-20 satellites, and GEO needs 3-6 satellites.

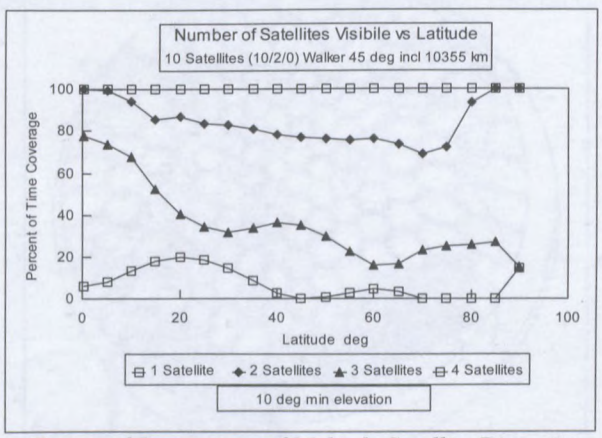
The choice of orbital configuration has to take into account not only the quality of service which will be delivered to the user, but also the feasibility and technical risk of the satellites themselves, and the problems of procuring and managing the required number of satellites.

It was concluded that the MEO configuration could offer best overall service quality for the desired market. This is because of the orbital properties, which confer, with a reasonable number of satellites, the following benefits:

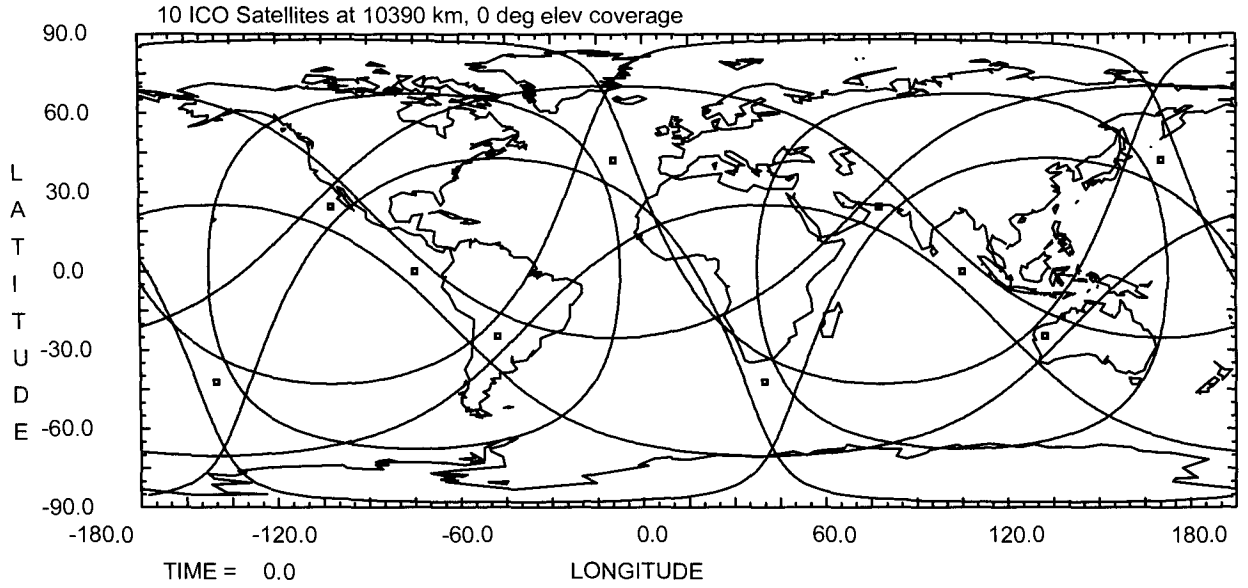
(i) high average elevation angle from user to satellites, minimising probability of blockage
(ii) a user being in the field of view of more than one satellite hence offering good satellite path diversity
(iii) slow-moving satellites (about 1 degree per minute across the sky as perceived by the user).



Minimum and average elevation angle of nearest satellite.

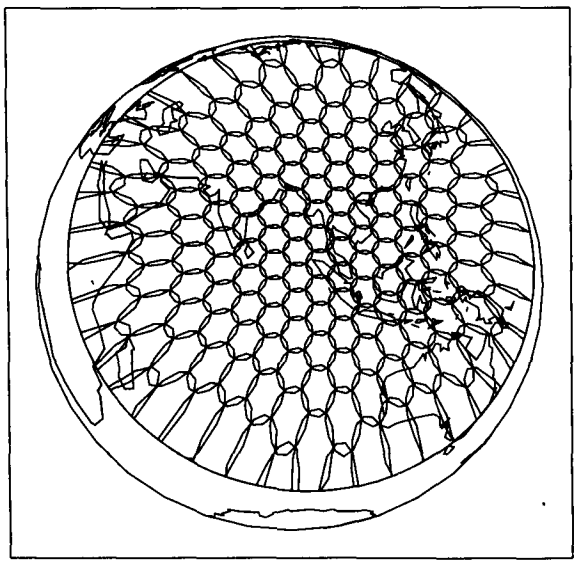


Temporal Percentage of Multiple Satellite Diversity



Instantaneous view of Global System Coverage

The technology studies of the LEO, MEO, and GEO satellite constellations concluded that MEO represents a reasonable implementation and schedule compromise. The large number of LEO satellites which would be needed, taken with their relatively short lifetimes in their expected radiation environment, present logistical and manufacturing problems in maintaining the constellation. GEO satellites become very complex in view of the high number of beams which would be needed from each one, and services are characterised by a relatively long transmission delay.



Service coverage of one ICO satellite.

SPACE SEGMENT

Satellite Constellation

The orbital pattern is designed for significant coverage overlap, ensuring that usually two but sometimes three and up to four satellites will be in view of a user and a SAN at any time. Each satellite will cover approximately 30 per cent of the earth's surface at a given time. The satellite orbits have been selected to provide coverage of the entire globe on a continuous basis, while allowing high elevation angles to users, averaging 40-50 degrees. An instantaneous view of the coverage of the 10 satellites is shown below.

Satellite Design

The satellites are based on the proven HS601 geostationary satellite bus. The communications payload is of transparent design, allowing flexibility to transmission format, using a high degree of digital technology for functions such as channelisation and beam generation that have traditionally been performed by analogue technology. The digital technology provides a very flexible satellite configuration, while having significant advantages over analogue technology in terms of production and manufacture for the comparatively large production run as compared with more conventional geostationary satellite orders.

Another key feature of the design is the separate transmit and receive antennas for the service link antennas, allowing easier manufacture and better intermodulation protection than a combined transmit/receive antenna.

Links between individual users and satellites will be established via service antennas mounted on each of the satellites. To provide robust radio links with handheld units, the satellites use antennas with an aperture in excess

of two metres. The use of multiple service link beams on each satellite also allows frequency re-use and increases the efficient use of spectrum allocation.

Each satellite is designed to support at least 4,500 telephone channels using time division multiple access (TDMA). TDMA technology was selected after careful consideration of other technologies.

The life span of ICO satellites is expected to be approximately twelve years.

Satellite Technology

Service link and number of beams: The system performance will be well in excess of that required to provide the desired level of service. The 163 transmit and receive service link beams will provide links with a minimum power margin in excess of 8 dB.

Feeder link antennas: Feeder link antennas support the link between the satellites and the SANs. At any time, each satellite will usually be in direct contact with between two and four SANs. Before a satellite falls outside the line of sight of one SAN, it will establish contact with another SAN. This SAN will then track the satellite whilst it is in its line of sight.

Service coverage of one ICO satellite

Satellite mass and power estimates: The total satellite launch mass, for a 'direct injection' into the final orbit, is about 2600 kgs allowing multiple launch vehicle capability. Direct injection allows some simplification to the HS601 as no apogee motor is needed to achieve final orbit. The solar arrays will use the latest Gallium Arsenide cells to provide end of life powers in excess of 8,700 W.

Service link spectrum requirements (for connection between user terminals and satellites): The choice of bands for the provision of service links for MSS systems include 1.6/1.5 GHz, 1.6/2.4 GHz and around 2 GHz.

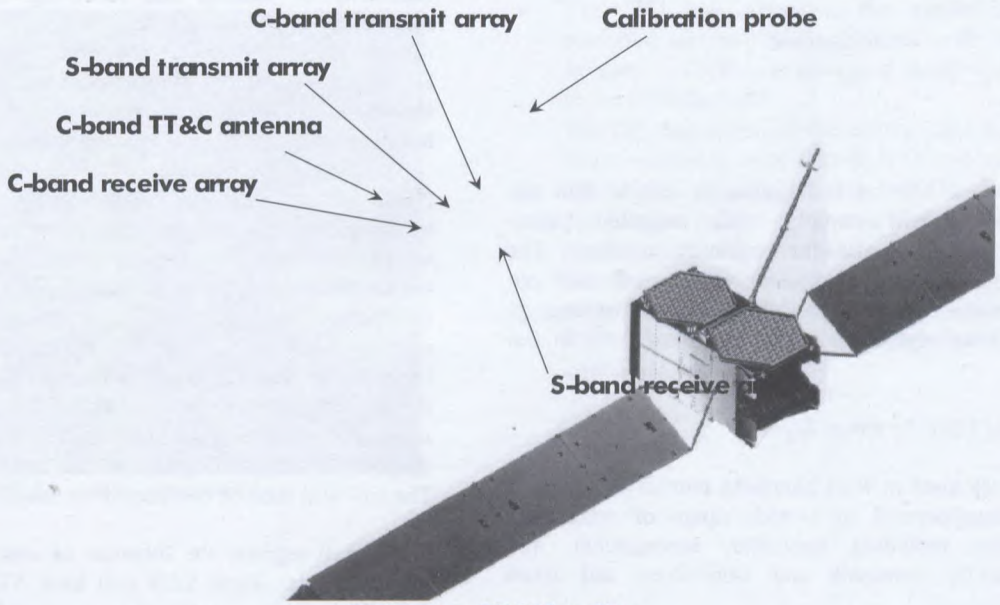
The World Radio Conference 1995 (WRC-95) made a number of important modifications to the original allocations at 2 GHz. The most relevant of these was that the date of access to the bands by the MSS was brought forward to 1st January 2000 (previously 2005), and that additional spectrum was made available in Region 2 (the Americas).

Feeder link spectrum requirements (for connection between satellites and SANs): For feeder link operation, ICO has chosen to operate in the 5 GHz and 7 GHz bands. These bands form part of a pair of new allocations made by WRC-95 for feeder links to non-geostationary satellites providing MSS.

TT&C

The satellite control centre (SCC) will manage the ICO satellite system by tracking the movements of the satellites and adjusting their orbits to maintain the constellation. It will also monitor the general condition of the satellites by collecting data on the power supply, temperature, stability and other operating characteristics of the satellites and will also have the ability to manoeuvre satellites to realign the satellite constellation in the event of any satellite malfunctions. In addition, at the outset of the ICO system implementation, the SCC will support the launch and deployment of the satellites.

The SCC will control the transponder linkages between the feeder and service antennas on the satellites. This process



Indicative ICO Satellite

will dictate, among other things, frequency reconfiguration within feeder link beams and optimal channel allocation between high and low traffic spot beams.

USER TERMINALS

Handheld Phones

The large majority of ICO user terminals are expected to be handheld, pocket-sized telephones, capable of dual-mode (satellite and cellular or PCS) operation and very similar in size and appearance and in voice quality to today's handheld cellular/PCS phones. The price of ICO dual mode phones, on the basis of high volume production, is expected to be competitive with other comparable satellite systems at service introduction.

The ICO handheld phone is planned to have optional features including external data ports and internal buffer memory to support data communication, messaging functions, facsimile, and the use of smartcards (SIMs).

ICO phone

Safety

The ICO system has been designed to ensure that the handheld phone will comply with expected safety requirements in respect of radio frequency radiation. The average transmitted power typically during use will not exceed 0.25 watts. Existing cellular phones have average transmitted power typically in the range of 0.25 to 0.6 watts.

Other Derived User Terminal Types

The technology used in ICO handheld phones is expected also to be incorporated in a wide range of other user terminal types including vehicular, aeronautical, and maritime mobile terminals and semi-fixed and fixed terminals, such as rural phone booths and community telephones.

ICONET AND SATELLITE ACCESS NODES

SANs will be the primary interface between the satellites and the terrestrial networks. They will also house the equipment which will route the satellite signals for distribution to the appropriate Gateways. A SAN will comprise three main elements:

(i) five antennas, with associated equipment to communicate with the satellites;
(ii) a switch to route traffic within the ICONET and to Gateways; and
(iii) databases to support mobility management.

Each SAN will contain a database to hold details of user terminals currently registered to that SAN (in GSM terminology, this is a Visitor Location Register or VLR).

Each SAN will track the satellites within its sight, direct communications traffic to the optimal satellite for the most robust link, and subsequently, as appropriate, will execute hand-offs to maintain uninterrupted communication.

The ICONET will be managed by the Network Management Centre.

USER MOBILITY MANAGEMENT

A critical feature of ICO will be its integration into public land mobile networks (PLMNs). In most instances the satellite network will be viewed as a complementary service into which PLMN subscribers who wish to have the capability of making and receiving calls in areas not serviced by their PLMNs may roam.

In order to provide global roaming, the ICONET will include a system for management of global user mobility based upon the existing digital cellular standard, GSM.

HLRs in co-ordination with VLRs will verify user information and status, and locate the user anywhere in the world. Any handset which is turned on will send a signal via satellite and SAN to the user's HLR which will verify the user's status and allow access to the system. The system will communicate this clearance to that SAN and register it in its VLR. The HLRs second function is to communicate the VLR location of any user to the SAN through which an incoming call is originated. This will enable the call to be directed to the SAN closest to the intended call recipient. The call will then be completed via satellite link.

VLRs will register the location of users outside of their home regions. Each SAN will have VLR capability. As described above, at registration, a signal from the handset

will be sent through the nearest SAN to the HLR and back to the VLR at that SAN. Incoming calls, initially signalling the HLR, will be sent to the SAN whose VLR is registering the current location of the user to complete the linkage via a satellite in that area.

ICO SERVICES

The Company intends to offer a wide range of satellite-based mobile telecommunications services through handheld, speciality and semi-fixed ICO Phones in the four market segments discussed above. These services includes voice, data, facsimile and value-added services such as voicemail, messaging, high penetration notification; three-way calling and call forwarding. Satellite-based access to computer systems, corporate networks and the Internet will also be offered by means of data communications at various speeds. The ICO System is designed to accommodate various data rates, initially at speeds of up to 9.6 kbits per second. The Company intends to offer specialised services such as vehicle location and fleet management services to customers in the ICO Speciality Mobile segment. The Company also intends to offer, in conjunction with its Service Partners, customer support packages in addition to its primary services. The Company is also considering offering a packet data service within its product portfolio.

An important feature of the ICO service will be a high penetration notification function intended to deliver alerts to ICO Phones to inform users of incoming voice calls or data messages in circumstances where normal communication is not possible. Unlike typical paging services, ICO's high penetration notification service is expected to provide confirmation to the call that the notification has been delivered.

CURRENT STATUS

SAN Site under construction

ICO is in its fourth year of implementation and on track to launch a full commercial service in the year 2000:

- Satellite construction by Hughes Space and Communications International since July 1995.
- Satellite System Critical Design Review completed.
- Launch vehicle procurement is complete. The satellites will be launched by a selection of vehicles: the US Atlas IIA and Delta III, Russia's Proton and the Ukrainian Zenit. The first launch is scheduled for 1999.
- A consortium comprising NEC, Hughes Network Systems and Ericsson has begun work on the design, construction and delivery of the ICONET.
- First set of SAN RFT equipment shipped.
- Sites for 12 SANs have been selected in the USA, Brazil, Chile, Mexico, South Africa, Germany, UAE, India, China, Australia, Korea and Indonesia.
- Site infrastructure and operations contracts have been signed in the USA, Chile, South Africa, Germany, UAE, India, Australia, Korea and Indonesia.
- Establishment of the Network Management Centre in Japan and the Satellite Control Centre in London is in progress.
- Contracts have been signed with NEC, Samsung and Mitsubishi for the design, development and manufacture of handsets. Negotiations are in progress with other manufacturers.
- Wavecom of France has been contracted as ICO's user terminal technical reference partner.
- DVSI has been contracted as the vocoder provider.
- Exclusive and non-exclusive distribution agreements have been signed with national service partners in over 90 markets.
- Service licences have been awarded in five countries and other jurisdictions have assured ICO investors that licences will be granted.
- The ITU has advanced the availability of ICO's preferred service-link frequencies in the 2GHz band to January 1, 2000, and allocated feeder-link frequencies in the 5/7GHz band.
- The ITU has allocated the country and mobile network codes needed to route calls to ICO customers.
- CEPT has confirmed that ICO has demonstrated compliance with first five of eight milestones. (Submission of relevant data for international frequency co-ordination, evidence of satellite manufacturing contract, satellite launch agreements, SAN agreements and completion of the spacecraft critical design review).

Interference Cancellation for TDMA-Satellite Systems

Harald Ernst
DLR (German Aerospace Center)
Institut für Nachrichtentechnik,
Oberpfaffenhofen, D-82234 Weßling , Germany
Harald.Ernst@dlr.de

ABSTRACT

This paper aims at the investigation of interference for a multibeam antenna satellite system and a method to reduce the impact of the interference.
First, the theoretical background is outlined and then a method is derived to combat interference. Simulations with the interference cancellation technique are shown. At the end possible restrictions due to imperfect knowledge and effects like Doppler are discussed.

INTRODUCTION

For a TDMA satellite systems there are two major limitations on the capacity of a satellite system: the achievable signal to noise ratio (SNR) due to the link budget and the signal to interference ratio (SIR) owing to the interference between users of different cells. The link budget influences mainly the capacity of one cell, whereas the interference in a TDMA system limits the possible number of cells with the same frequency.

Regarding interference in a TDMA system, the uplink is the worse case, since the transmitter of a user is allowed to be positioned anywhere in its cell, and therefore the distance between a transmitter and its interference sources can become smaller than for the downlink case. For that reason, the article deals only with the uplink channel to the satellite.

The investigated system is a geostationary satellite with a multibeam antenna realized with a fixed beam forming network. It is also assumed that synchronization is done in the digital domain. Therefore the different phases/delays between all signals are accessible. In order to emphasize the main points of the idea a simplified model will be used, assuming that all users are bit-synchronous and have a random phase shift.

Most of the aspects of the interference cancellation will be shown for an ideal regular 61-cell beam antenna, see fig. 1. The grey shaded cells indicate the later used cells for the 3 frequency TDMA system.

For this pattern, the antenna elements are assumed to be as defined by ITU-R S.672.3, with quadratic increasing attenuation in dB, –30 dB side lobe-level and cell radius at the –3dB limit.

INFORMATION THEORETIC ASPECTS

A first approach for this system is to compute the capacity of the whole system. For simplicity, the satellite channel is assumed to be an AWGN channel with a defined interference between different beams.

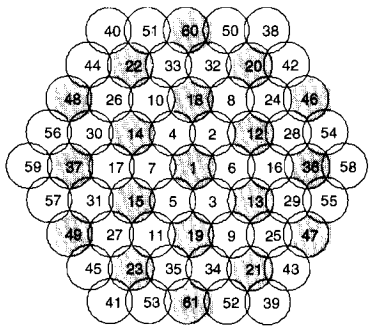


Fig. 1: Regular 61 cells pattern

As stated in Suard et. al. [1] the sum capacity of all users of a multibeam antenna can be written as :

$$\sum_{k=1}^{P} R_k \le \frac{1}{2}\log\det\left(\mathbf{I} + \mathbf{A}\mathbf{R}_S\mathbf{A}^H\right) \qquad (1)$$

Where P is the number of users, the columns of $\mathbf{A}$, $\mathbf{a_k}$ are defined as the response vector of the antenna beam for the k'th user. So the element $\mathbf{A}(i,k)$ is the impulse response of the k-th user on the i-th antenna beam.

$$\mathbf{R}_S = \mathrm{diag}\{\sigma_1^2, \cdots, \sigma_P^2\} \qquad (2)$$

is the SNR of the different uncorrelated user signals. Formula 1 assumes that all user signals are jointly decoded.

If the signals are independently decoded, as normally done, the signal power of all other users in the beam is treated as noise. The sum capacity can then be expressed as:

$$\sum_{k=1}^{P} R_k \le \frac{1}{2}\sum_{k=1}^{P}\log\left(1 + \frac{|A(k,k)|^2 \cdot \sigma_k^2}{1 + \sum_{h=1,h\neq k}^{P} |A(k,h)|^2 \cdot \sigma_h^2}\right) \qquad (3)$$

Therefore the possible increase of capacity by using joint detection can be computed. In the case of the regular 61-

cell pattern with 61 users, joint detection would more than double the sum capacity of the satellite, depending on the individual SNR and the positions of the users. Additionally the capacity could increase further, since even more users than beams could be allowed.

The main drawback for joint decoding is the complexity of the receiver. The optimal maximum-likelihood decoder described by Verdu in [2] has to combine all available knowledge from every terminal's signal. This is done, roughly speaking, by using a Viterbi-like decoding algorithm where each state defines the combination of the signals sent by all users. Therefore, there are 2^{USERS} states and the complexity of the decoder grows exponentially.

Thus, different suboptimal schemes were developed, one of them is interference cancellation.

INTERFERENCE CANCELLATION

In the case of interference cancellation the information is independently decoded in a first step, just like in the conventional scheme. After demodulation and channel decoding there exists an estimate of the transmitted signals. This information is then used to eliminate the interference by subtracting the estimated influence of the signals from the interfering users from the signal which is looked on. This improved signal is then again decoded, giving a better estimate of the signal. This is done for all signals.

These 2 last steps (decoding and cancellation) can then be iterated with ever increasing reliability on the bits, until near single user performance has been reached, see e.g. [3].

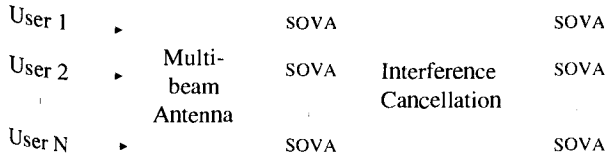


Fig. 2: One stage interference cancellation

In order to generate a good estimation of the influence of the signal, the channel decoder does not only compute hard decisions but soft-information about the quality of the information, so that reliable decisions are taken into account to a greater extent than unreliable. An example for a soft decoding algorithm is the soft-output-Viterbi-decoder (SOVA) [4].

TDMA AND INTERFERENCE REDUCTION

Even, though these sub-optimal schemes need a much lower complexity than the optimal decoder, their complexity still increase significantly with the number of users.

For TDMA an interesting aspect is the frequency re-use scheme used for a typical regular TDMA system. There are always 6 nearest neighbors, which are responsible for the main inter-cell interference. Therefore the complexity of interference cancellation for only 6 next neighbors seems manageable. Additionally the signals are fairly synchronized and have a common frequency slot.

In the satellite case, there are two points which make inter-cell interference cancellation especially attractive. First, there is only one "base station" , i.e. the satellite, where all the information from the different user from different cells is present. Secondly the attenuation of the antenna drops only with the square of boresight deviation, which is far less than in the terrestrial case, resulting in a higher influence on the cell pattern.

In the following part of the paper the improvement is shown, which can be obtained by a simple 1-stage interference cancellation, for an regular TDMA pattern with cluster-size 3, as indicated in figure 1.

SIMULATIONS

In the simulation a soft-output-Viterbi-decoder (SOVA) is used with a memory 4 Rate R=½ convolutional code. BPSK is chosen and the phase between users is assumed to be random. Additionally, the TDMA-frames of the users are individually interleaved. The matrix **A**, which defines the correlation between the users, is assumed to be known at the receiver.

RESULTS

The first simulation is done for the case of the TDMA system indicated in figure 1. First, the worst case is considered: the positions of the users are shown by dots in figure 3. The carrier to interference ratio (C/I) for the user in the center cell of the pattern is then 1.1 dB.

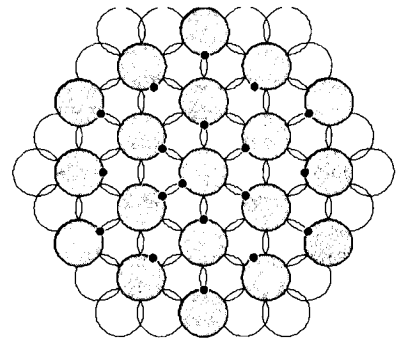


Fig. 3: User position for worst case

The overall signal to noise and interference ratio (SNIR) for independent decoding and the assumption of equal power for all user is then:

$$SNIR = \frac{E_b}{N_0 + \frac{R}{C/I}} \qquad (4)$$

Figure 4 shows the bit error rate over E_b/N_0 for the center user, first in the case of independent decoding. The second curve marks the joint decoding, taking only into account the 6 nearest neighbors, the third curve is full joint decoding. Finally, it is shown what can be achieved by iterating the interference cancellation once more with the single user curve as reference.

As can be seen in figure 4, the bit error rate of the independent decoding is only decreasing very slowly, since the interference is equivalent to a second noise source at 4.1 dB (due to the rate R= ½ convolutional code) and therefore the independent decoding at E_b/N_0 4.1 dB is equivalent to a single user at SNR =1.1 dB. For higher SNR the BER is then asymptotically reaching the single user SNR = 4.1 dB point.

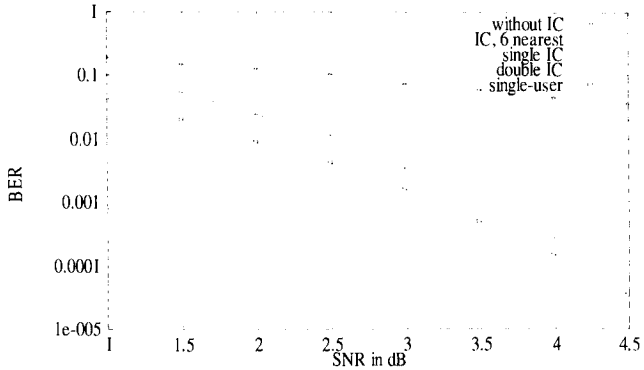


Fig. 4: BER for worst case single user

In contrast to this, the joint decoders have only a small loss of less than 0.5 dB with respect to the single user case. Also there is only a negligible loss, if only the 6 nearest neighbors are taken into account and only a small gain, if the interference cancellation is iterated.

Since the reduction of interference is so successful, it can be used to design a pattern with even higher capacity. Therefore an scenario is assumed with the users at the center of their cells but using only a single frequency, see figure 5.

Actually this pattern is equivalent, regarding interference, to a standard 3 frequency scenario as in figure 1, with the only difference, that the cell radius is not defined to be at the –3 dB edge, but at -1.3 dB. This would result in tripling the number of cells, whereas the diameter of the antenna reflector would stay constant.

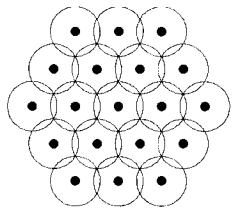


Fig. 5: Single frequency scenario

The C/I at the center cell is 0.55 dB. The results for this user can be seen in figure 6.

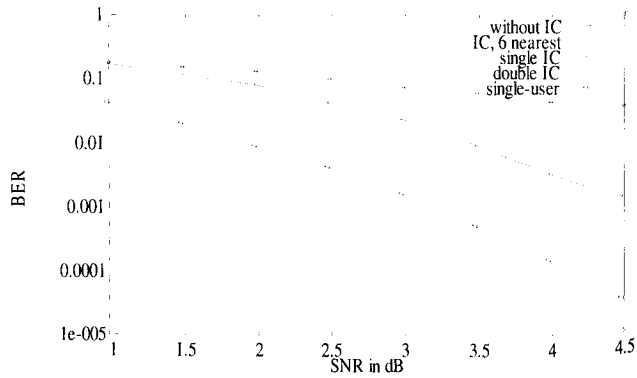


Fig. 6: BER for single frequency system

Since there is now a higher interference between the signals, the gain by doing an additionally interference cancellation stage is increased. A characteristic of repeated interference cancellation is the fact, that there is a threshold, after which it can reach single user BER with enough iterations, see [3]. But at lower SNR, the improvement due to iterations is not so high.

For interference cancellation it is important, that there is an independent interleaving between the users. Figure 7 shows the loss, if the convolutional encoders are synchronous.

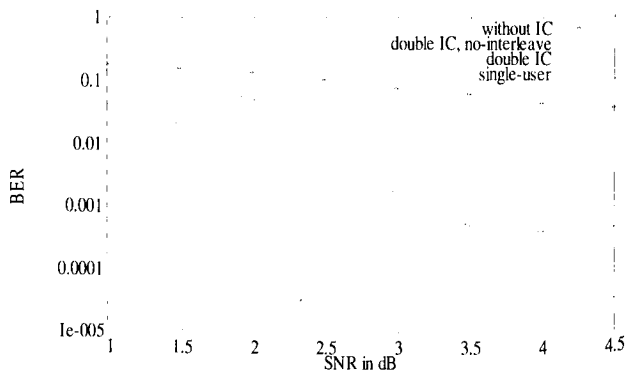


Fig. 7: With and without interleaver for single frequency

UNEQUAL POWER DISTRIBUTION

In the mobile satellite case, the power difference between the received signals of line-of-sight user and a shadowed user can be quite high. Therefore a high link margin is necessary to combat this effect (or alternatively satellite diversity is needed). But a high link margin increases the effect of the interference, since the link margin doesn't include the effect, that each line-of-sight user generates a relative much higher interference for the signal of the shadowed user.

The Lutz-model [5] was used as channel model. We simulated a highway environment with a Rice-factor of 11.9 dB, a shadowing probability of 0.25 a mean attenuation of –7.7 dB for the lognormal fading and a sigma 6 dB for the fully interleaved Rayleigh part. In contrast to the last paragraphs here the BER is computed over all users.

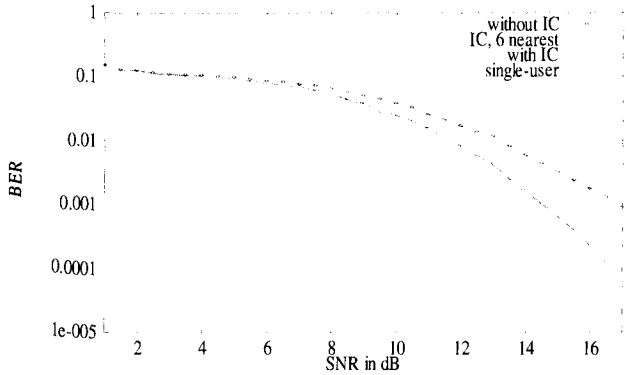


Fig. 8: BER for best case in an highway environment

In figure 8 the best possible situation is assumed, that every user is exactly in the center of its beam, resulting in a C/I of ca. 20 dB. But even then the interference cancellation results in a gain of roughly 2 dB at 10^{-3} BER.

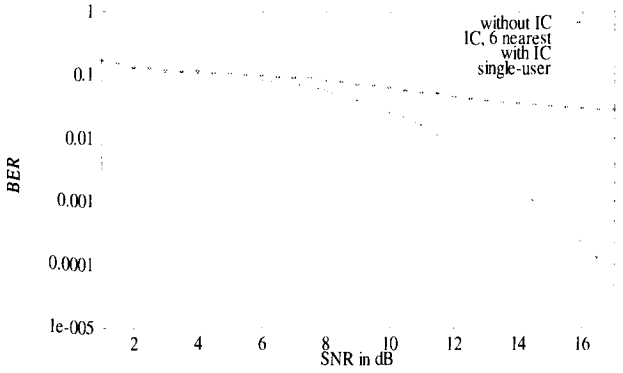


Fig. 9 BER for worst case in an highway environment

In the second figure the worst case from figure 3 is used, but the BER is computed over all users. It can be easily seen that even a high static link margin is useless without interference cancellation. In a real system this effect would be of course be diminished by the power control of the system, but the effect still exists for the transition time, until power control can react.

A second effect is, that there is noticeable interference from cells which are farther away than the next 6 cells. But the neighboring cells still dominate the interference, as can be seen in the small loss in figure 7 in the curve for the interference cancellation for only the 6 nearest neighbors.

SECOMS-PATTERN

In the frame of the European ACTS project SECOMS/ABATE[6] a multi-beam antenna was designed to cover Europe. In the project a 4-frequency reuse pattern was devised for the uplink [7]. In this paper we assume a frequency reuse of 3. A fixed pattern for the positions of the users was selected, which is shown in figure 11. The C/I of the different cell can be seen in the following table:

Cell #	1	3	6	8	11	14
C/I	19 dB	19 dB	18 dB	19 dB	12 dB	16 dB
Cell #	17	20	23	26	29	31
C/I	12 dB	10 dB	19 dB	19 dB	7 dB	10 dB

Fig. 10: C/I of the different cells

What is typical, is the fact that for a given pattern and user distribution, some links have a C/I as high as 20 dB, whereas only a few have a C/I worse than 10 dB. It is often easy to find a user distribution that ensures, that a specific link has a really bad C/I, but then other cells must have good C/I. This eases the task for the interference cancellation, since there are always signals which can be subtracted with a very high reliability.

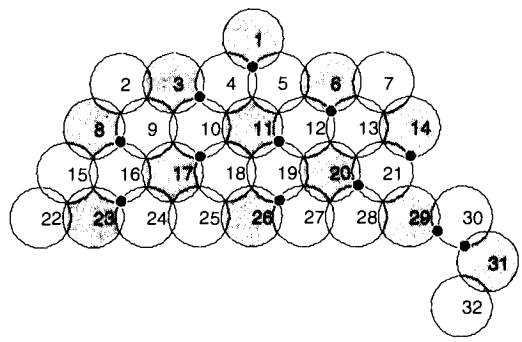


Fig. 11: Cell-Pattern for SECOMS project

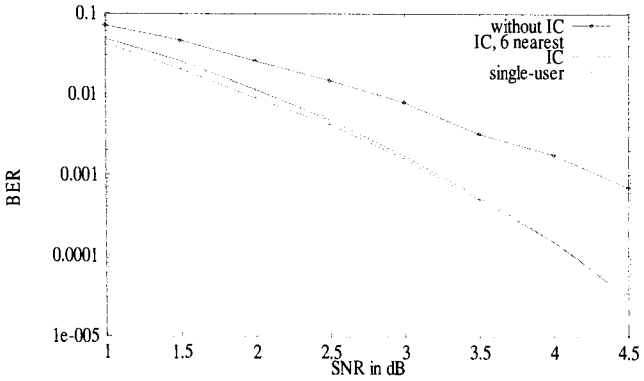


Fig. 12 BER-curve for SECOMS-antenna

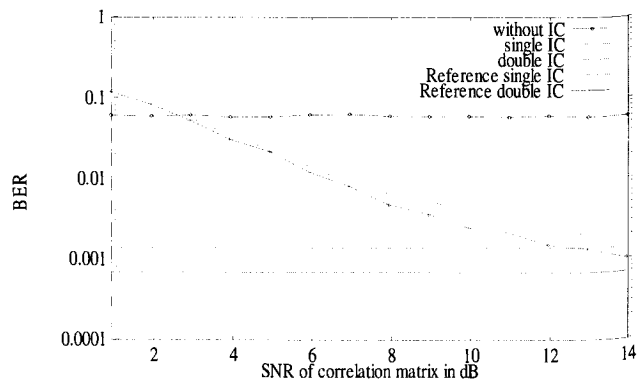


Fig. 13 Effect of noise on BER

Since there is only a small interference, even interference cancellation, taking into account only the 6 nearest neighbors, is near optimal, but on the other hand the gain is small and only increasing for low bit error rates. A second problem for interference cancellation for small interference is the fact that the estimation of the correlation matrix **A** will become more problematic, since the relative influence of error sources will increase.

PROBLEMS: DOPPLER AND FREQUENCY-OFFSET

The main problem of the interference cancellation scheme is getting a correct estimation of the correlation matrix **A**. Therefore not only the power and attenuation of the different signals have to be computed, but also the phase shift between the different users.

Unfortunately this phase difference is not fixed for one block, because of the frequency difference of the signals and the possibility of Doppler shift. Therefore the correlation for every bit of the block has to be phase shifted according to the frequency difference.

Since the frequency difference and the Doppler shift has to be estimated, inaccuracies are introduced to the correlation matrix.

Further error sources can be incorrect knowledge on the position, resulting in an incorrect estimated phase and amplitude value for the antenna pattern, and the error due to synchronization inaccuracy. Since in the case of a geo-stationary satellite, even a coarse knowledge of the position of the user results in a very good estimate on the boresight angle to the satellite antenna, the main problem is in the synchronization and frequency difference.

For the evaluation of the effect of imperfect knowledge on the interference cancellation, the following simulation assumes not anymore perfect knowledge on the correlation matrix **A** between user signals, but Gaussian noise is added to the correlation matrix, according to a given SNR. The SNR in the channel is held constant at an E_b/N_0 of 3.5 dB.

The effect of the noise is a degradation of the interference cancellation. Interestingly the algorithm is only in a small region worse than independent decoding and even this region could be reduced, if the algorithm would be aware of the bad correlation estimations. Summarizing the result, interference cancellation is a relative robust algorithm.

CONCLUSION

It has been shown how and to what extent interference cancellation may be used to improve TDMA satellite systems.

In contrast to CDMA, interference cancellation for TDMA needs less complexity, since only a small number of neighbors dominate the interference and it can be done with less iterations, since the difference between the user with worst C/I and users with very good C/I is typically very large. Also there exist for a given user position always some antenna beams, where the interference can be neglected and which can be decoded successfully in the first step.

On the other hand the computation and estimation of the correlation coefficients may be more difficult, due to the fact that for an overlapping zone more than 1 bit of the interfering user's signal has to be taken into account.

One alternative method to the full computation of the correlation coefficients could be a measurement of its influence in the registration phase, if e.g. one TDMA frame is especially reserved for registration purpose.

But even in this case, important parts of the demodulation and filtering have to be done in the digital domain to make adjustments for frequency differences and the Doppler effect.

As a further area, the application of interference cancellation for satellites is not limited to this topic, but can also be used e.g. to get an increased cross-polarization attenuation and in CDMA based systems.

REFERENCES

[1] B. Suard, G. Xu, H. Liu and T. Kailath, Uplink Channel Capacity of Space-Division-Multiple-Access Schemes, IEEE Trans. on Information Theory, vol. 44, pp. 1468-1476, July 1998

[2] S. Verdu, Minimum Probability of Error for Asynchronous Gaussian Multiple Access Channels, IEEE Trans. on Information Theory, vol. 32, pp. 85-96, Jan. 1986

[3] P. Alexander, A. Grant and M. Reed, Iterative Detection in Code-Division Multiple-Access Interference Suppression Algorithms for CDMA Systems, European Trans. on Telecommunication, vol. 9, pp. 419-426, Sep. 1998

[4] Hagenauer and P. Hoeher, A Viterbi algorithm with soft-decision outputs and its applications, in Proc. IEEE GLOBECOM'89, Dallas, TX, Nov. 1989, pp. 47.1.1-47.1.7.

[5] Lutz et. al , The Land Mobile Satellite Communication Channel – Recording, Statistics and Channel Model, IEEE Trans. on Vehicular Technology, vol. 40, pp. 375-386, May 1991

[6] G. Losquadro, M. Luglio and Vatalaro, A geostationary satellite system for mobile multimedia applications using portable aeronautical and mobile terminals, Proc. IMSC'97, pp. 427-432,1997

[7] Deliverable D 35, Aeronautical Link Design Report, SECOMS/ABATE project, August 1997

CAPACITY ENHANCEMENT OF A CDMA BASED MOBILE SATELLITE SYSTEMS BY BAND SHARING

H.M.Aziz, R.Tafazolli, B.G.Evans
Mobile Communications Research Group
Centre for Communication Systems Research (CCSR)
University of Surrey
Guildford Surrey GU2 5XH UK
E-mail: M.Aziz@ee.surrey.ac.uk

ABSTRACT

In this paper, we present an approach to maximize the capacity of a CDMA based mobile satellite system by band sharing. The overall increase in total system capacity due to band sharing will enhance system performance by utilising the available bandwidth efficiently. The results of the investigation on the total system capacity achievable by band sharing between the two mobile satellite systems (48 satellites LEO system and 12 satellites MEO system) are presented. Results for a single satellite system operating in a band segmentation approach are also presented for performance analysis and comparison.

INTRODUCTION

The next generation of mobile satellite systems (MSS) will use constellations of satellites in non-geostationary orbits, such as LEO (low earth orbit) or MEO (medium earth orbit) for provision of mobile and personal services with global coverage. These new MSS will provide communications to both developed and developing areas of the world where there is little or no telecommunications infrastructure or where it is not economically viable to offer terrestrial cellular coverage due to low population density. The proposed satellite personal communication networks (S-PCNs) are Iridium, ICO, Globalstar and Ellipso. The MSS will use L-band (1610-1626.5MHz) for uplink and S-band (2483.5-2500MHz) for downlink to provide world wide services. These bands were allocated to the MSS on a primary basis at 1992 World Administrative Radio Conference (WARC-92). However, the scarcity of free spectrum, together with the bandwidth required means that MSS will have to share these bands with other systems and services. The first two systems propose to use a Time Division Multiple Access (TDMA) scheme and the others Code Division Multiple Access (CDMA) scheme. It is claimed that CDMA based S-PCNs can share the allocated frequency bands due to the interference resistant property of CDMA [1][2]. The advantages of using CDMA access scheme include: low user terminal power flux density emission, full frequency reuse, soft hand-off (improves the call drop probability), path diversity exploitation capabilities (combining of multiple signals using a RAKE receiver) and intra-system interference control capabilities.

SYSTEM MODEL

A mobile terminal monitors the best two satellites in view for each constellation. The link is established with the satellite from which the maximum power is received. The received power from each satellite is a function of elevation angle and shadowing (channel state). Once the link is set up between the mobile terminal and the selected satellite spotbeam, the interference will be caused by the following:

- multiple access interference from other users in the coverage area (same spotbeam)
- interference from adjacent spotbeams (same satellite),
- interference from other satellite spotbeams (all the satellites that are visible to the mobile terminal) and
- external interference from other MSS sharing the spectrum.

The received carrier from the desired satellite spotbeam and interferences from the above mentioned sources are illustrated in Figure 1.

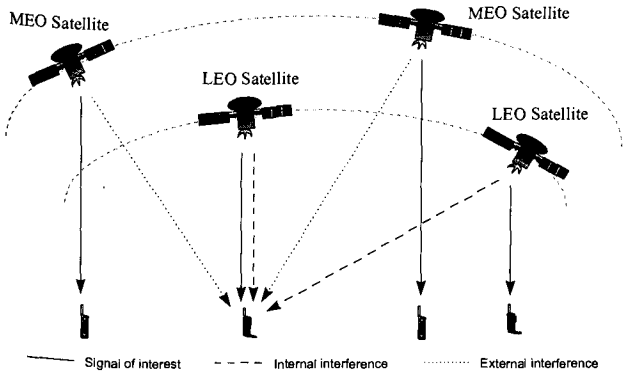


Figure 1: Received carrier power and interference from adjacent satellite spotbeams

The following assumptions are used for the carrier-to-interference (C/I) calculations:

- The selected satellite for the communication link is the one from which maximum power is received.
- Any mobile will have the same carrier power at the receiver input. Hence power control is required to compensate the variations in losses, which are dependent on the mobile location relative to the satellite.
- The mobile antenna is omni-directional.
- Maximum interference is caused by the interferer when the frequency spectrum is totally overlapped.

Using the above assumptions the C/I equations are as follows:

The C is the carrier power received at the mobile terminal.

$$C = \frac{P_{TW} G_{TW}(\theta) G_{RW}(\alpha)}{L(d) P_e a m} \qquad (1)$$

The pseudo noise density Io_1 , is the multiple access interference resulting from (m-1) interferers (i.e. m users are communicating simultaneously per carrier in each spotbeam) in the same spotbeam and can be written as:

$$I_{o_1} = \frac{Ca(m-1)F_1(10)^{\frac{\Delta}{10}}}{B} \qquad (2)$$

The pseudo noise density Io_2, is the beam-to-beam interference resulting from m interferers in each adjacent spotbeams assuming frequency re-use of '1' and can be written as:

$$C_I = \frac{P_{TI} G_{TI}(\theta) G_{RW}(\alpha)}{L(d) P_e a m} \qquad (3)$$

$$I_{o_2} = \frac{C_I a m F_2(10)^{\frac{\Delta}{10}}}{B}$$

The C/I levels are added as thermal noise. The total interfering signal power is the sum of the powers from all 'N' visible satellite spotbeams in the interfering system.

$$\left[\frac{C}{I_o}\right]_T = \left[\frac{C}{I_{o_1} + \sum_{k=1}^{N} I_{o_{2,k}} + I_{o_3}}\right] \qquad (4)$$

where:

P_{TW} : Wanted satellite spotbeam power
P_{TI} : Interfering satellite spotbeam power
$G_{TW}(\theta)$: Wanted spotbeam gain in direction θ
$G_{TI}(\theta)$: Interfering spotbeam gain in direction θ
$G_{RW}(\alpha)$: Mobile terminal antenna gain in direction α
L(d) : Free space path loss
P_e : Propagation effects which takes into account the shadowing/fading loss for the link which is a function of elevation angle and environment.
a : Voice activity ratio
m : Number of users per carrier
Δ : Power control error
F_i : Correlation factor
B : Sub-band bandwidth
Io_3 : External interference in band shared scenario.

Isoflux satellite antenna design is assumed to compensate the differential path loss (i.e. to compensate for the path loss variations due to the slant range differences from the satellite to the earth).

MOBILE SATELLITE SYSTEM CHANNEL

Correlated MSS channel model

As the mobile terminal moves from one location to another, the environmental properties change. Hence the received signal is represented by a model with varying parameters. This type of model is known as non-stationary. Although the channel characteristics vary over large areas, propagation experiments have shown that channel characteristics remain constant over areas with identical environmental features [3]. Therefore a land mobile satellite channel can be modelled with constant parameters over these areas. A channel model for a large area of interest can be modelled by a finite state Markov model [3][4].

In a Markov model, the whole area of interest is divided into M different areas with constant environmental characteristics. Then each of the M areas is represented by a stationary channel models. Particular channel states are characterised by one of the models, Rician, Rayleigh and Log-normal. The

probability of a mobile terminal moving from one state to another is described by the transition probability matrix. During each state the mobile terminal is in, the transition probabilities are assumed to be constant. The probability of switch from one state to the next is dependent on the time a mobile terminal was in present state, the elevation angle (i.e. at high elevation angles, probability of satellite visibility is high) and the satellite constellation dynamics.

In this model, we assume two statistically dependent land mobile satellite channels. The combined shadowing behaviour of the two land mobile satellite channels can be modelled by a four state Markov with azimuth correlation [6]. The correlation coefficient is dependent on the elevation angles and on the azimuth separation of the two satellites. The correlation decreases as the azimuth separation between the two satellite increases. Figure 2 shows the possible combinations of good and bad states of channels 1 and 2. The two channels are used in the simulation for the two highest satellites. If one of these satellites is shadowed, there is some chance for another satellite to be still in view of the user and maintain service. In this way, service availability can be substantially improved (the percentage of time when the service is available). The channel states 1,2 and 3 correspond to good channel, where at least one satellite is available; the bad channel state 0, represents a situation where the signal from both satellite is blocked by an obstacle in the propagation path.

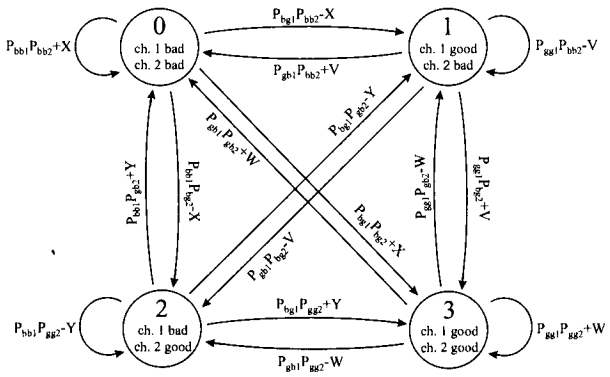


Figure 2: Correlated four state Markov model

The land mobile satellite channel used in the simulation is for the suburban environment. The parameters of the model used are taken from the measured data for low elevation angle from [4] and high elevation angle from [5].

Cross polarization characteristics

In [7], the channel measurements were carried out to assess the isolation between co- and cross-polarized transmission using the Japanese ETS-V and Inmarsat-POR satellites. The isolation at the different fade levels by defining the "equal-probability isolation" as the ratio of cross-polarized to the co-polarized signal levels were derived. The isolation between the polarization reduces as the fading level increases. Figure 3 shows the cross polar isolation.

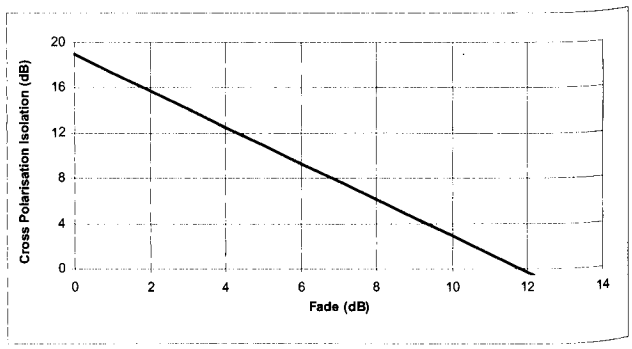


Figure 3: Cross polarization isolation

Based on this land mobile satellite channel, the total system capacities are calculated for band segmented and band shared scenarios.

BAND SHARING SCENARIO

The total available bandwidth is shared between a LEO system and a MEO system. The constellation parameters for the two systems are tabulated in Table 1.

Table 1: Constellation Parameters

Parameters	LEO-48	MEO-12
Orbit altitude	1414km	10354km
Orbit Period	113min	6hrs
Number of satellites	48	12
Number of planes	8	3
Inclination	52°	50°
Minimum elevation angle	10°	20°
Number of spotbeams/sat	19	61
Available Bandwidth	11MHz	11MHz
Carrier Bandwidth	1.25MHz	5.5MHz
Downlink frequency	S-Band	S-Band
Uplink frequency	L-Band	L-Band
Satellite power	1000W	3500W
Voice activity ratio	0.4	0.4

Band segmentation

The total available bandwidth of 11 MHz is divided into two 5.5 MHz sub bands. Both MSS operate in different bands without interfering with each other.

Band shared

The total available bandwidth of 11 MHz is shared between the MSS. In the first scenario, both system share the frequency bandwidth by fully overlapping [8][9]. In the second scenario, each system share the total bandwidth by operating in different polarization. The LEO system uses the RHCP, while the MEO system uses the LHCP. Figure 4 presents both sharing scenarios. The external interference between the MSS is unavoidable as there is no co-ordination assumed between the two systems. To minimize the external interference between the systems, each system must control its transmitted carrier power.

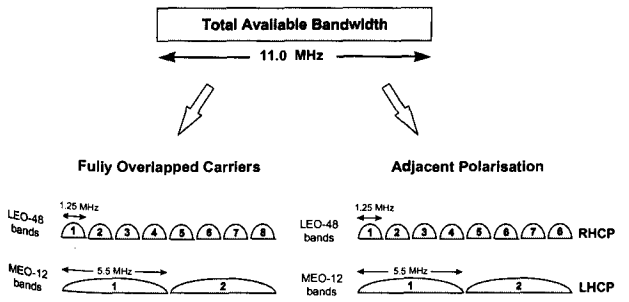


Figure 4: Bandwidth sharing scenario

Carrier-to-interference calculations

Computation of C/I values involves calculating the satellite antenna gain, free space path loss (i.e. distance between the satellite and mobile) and propagation losses. These parameters vary with time according to the satellite motion dynamics.

The SPOC+ (Simulation Package for Orbital Constellation Plus) simulation package was used to predict the positions of each satellite in a constellation at any instance. Hence knowing the position of a mobile terminal on the earth , the software calculates all the satellites that are visible (including the one's over the horizon) to a mobile terminal. Once these are known the software calculates the elevation angle (ε) and slant range (d). The two highest satellites are selected for the communication link. The communication link is established with the satellite spotbeam from which the mobile terminal receives the maximum power. The service area of the spotbeam varies with time, according to constellation dynamics. Hence the service link is transferred from spotbeam to spotbeam. When the mobile terminal is out of satellite coverage area (i.e. the elevation angle is less than the minimum elevation angle), the link is then transferred to another satellite. The intra and inter satellite interference is calculated from all the satellites that are visible to a mobile terminal.

RESULTS

In this section, we present numerical results on total system capacities of a band segmented and a band shared scenarios. A wanted mobile terminal is selected on the surface of the earth and C/I values are calculated by measuring the carrier level and the corresponding interference levels. The worst case scenario for the C/I will occur when the mobile terminal is in the vicinity of several satellites (i.e. the location where the mobile terminal has the highest satellite diversity). The C/(No+Io) of 44.8dB is assumed for a required bit rate of 9600 bps at a bit error rate (BER) of 10^{-3}. The C/(No+Io) statistics calculated correspond to those values that are exceeded for 95% of the time.

MSS carrier-to-interference statistics

Figure 5 and Figure 6 show the carrier-to-interference, C/(No+Io) profile and the C/(No+Io) cumulative distribution functions for the link into the LEO MSS mobile terminal receiver. The C/(No+Io) profile shows the C/(No+Io) against time as the satellite moves away from the mobile terminal. The worst interference event occurs when the mobile terminal is in the vicinity of several interfering satellites and the mobile terminal is shadowed (deep fade) from the wanted satellite. At various times the C/(No+Io) level changes in a discontinuous manner, this is due to the mobile terminal switching from one LEO system satellite to another. Satellite handover is determined by the system design and takes place either when the mobile terminal is at the edge of a satellite spotbeam coverage, or when the wanted satellite goes into deep fade due to shadowing. Similarly, Figure 7 and Figure 8 present the carrier-to-interference, C/(No+Io) profile and the C/(No+Io) cumulative distribution functions for the link into the MEO MSS mobile terminal receiver. It is seen in the C/(No+Io) profiles, the variation of C/(No+Io) in the LEO system is more frequent than the MEO system. This is due to the nature of the LEO and the MEO orbit dynamics.

The distribution functions shows that the required quality of service is achieved for 95% of the time. However, there are times when the C/(No+Io) drops

below the required protection ratio and the link outage will occur. This happens due to higher interference received from adjacent users or when the carrier level drops below the threshold due to shadowing.

MSS capacity

The capacity of the MSS is computed such that the required quality of service (C/(No+Io) is achieved for 95% of the time. The system capacity for the band segmented scenario is used for comparison and performance analysis with the other band sharing schemes. In Table 2, the capacity per satellite for the LEO and the MEO systems are presented. Capacities of 1637 and 3841 users per satellite can be supported for conventional band segmentation approach for the LEO and the MEO system, respectively. For fully overlapped band shared scenario, the satellite capacity of each system is reduced to 1632 and 2063. This corresponds to reduction in capacity of 0.3% and 24.6% for the LEO and the MEO system, which translates into total capacity reduction of 18.4%. The reduction in capacity for MEO system is severe due to multiple LEO system carriers falling within the MEO system bandwidth, whereas in the LEO system only fraction of the MEO carrier power is received by the LEO system receiver. In the second scenario, adjacent polarization is used for each system to enhance capacity. Using band segmentation approach with adjacent polarization, capacities of 1664 and 4167 users per satellite are achieved for the LEO and the MEO system. To further enhance the overall system capacities of each system, the total available bandwidth is shared, with each system operating in a single polarization. The results presented showed an increase in system capacities of 1776 and 4271 users per satellite for the LEO and the MEO system. This represents an enhancement of capacity by 6.7% and 2.5% for the LEO and the MEO system compared with capacity achieved using band segmentation approach with adjacent polarization. However, when this capacity is compared with the conventional band segmentation approach, the increase is 8.5% and 11.2% for the LEO and the MEO system, respectively. This translates into overall system capacity gain of 10.5% for both the LEO and the MEO system.

The total system capacities of the LEO and the MEO system are shown in Figure 9 for band segmented and band shared scenarios.

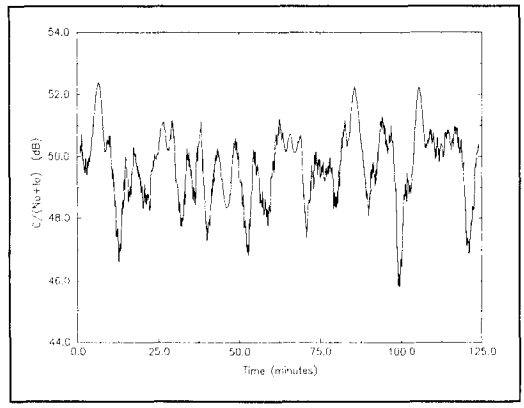


Figure 5: Variation of C/I for a LEO system

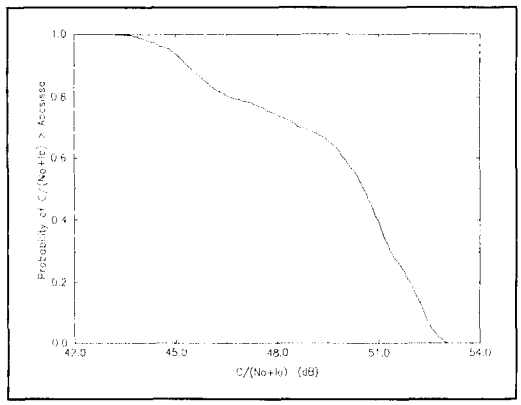


Figure 6: Distribution of C/I for a LEO system

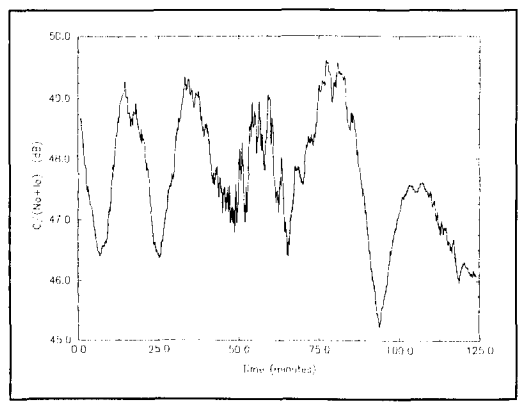


Figure 7: Variation of C/I for a MEO system

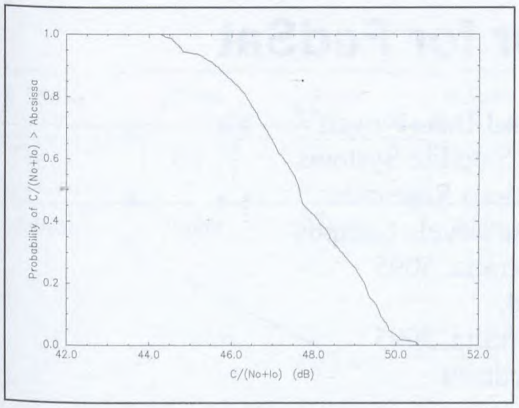


Figure 8: Distribution of C/I for a MEO system

Table 2: Capacity estimate per satellite for LEO and MEO system

Scenario/ System	Full Overlapping		Adjacent polarization	
	Band segment	Band Shared	Band segment	Band Shared
LEO	1637	1632	1664	1776
MEO	3841	2063	4167	4271

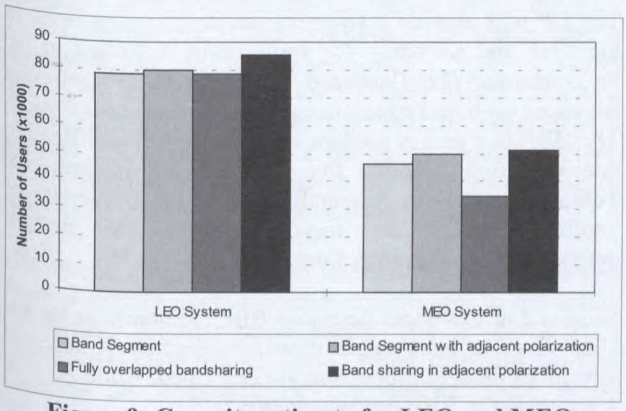


Figure 9: Capacity estimate for LEO and MEO systems

CONCLUSIONS

In this paper, we assessed the capacities of a CDMA based mobile satellite system. Different sharing scenario were evaluated for performance analysis and comparison. We presented an approach to maximise the capacity of the LEO and the MEO system by band sharing. The simulation results presented in this paper are for a conventional band segmentation approach and a new overlapped band sharing approach. In a band segmentation approach, each system operates in its own allocated bandwidth. However, in the band shared case, external interference is unavoidable due to overlapping of carrier bandwidths. The results show that the total capacity can be increased by band sharing the available bandwidth between the LEO and the MEO system. This increase in total system capacity due to band sharing has enhanced system performance by utilising the available bandwidth efficiently.

REFERENCES

[1] B R Vojcic et al, "Total capacity in a shared CDMA LEOs environment", IEEE J. on Sel. Areas in Comm., Vol. 13, No. 2, pp. 232-244, February 1995.

[2] L B Milstein and J Wang, "Interference suppression for CDMA overlays of narrowband waveforms", Proc. of the IEEE 3rd Int. Sym. on Spread Spectrum Tech. and App., July 1994.

[3] B Vucetic and J Du, "Channel modelling and simulation in satellite mobile communication systems" IEEE J. on Sel. Areas in comm., pp. 1209-1218, October 1992.

[4] E Lutz et al, "The land mobile satellite communication channel - recording, statistics and channel model", IEEE Trans. On Veh. Tech., pp.375-385, May 1991.

[5] G Butt et al, "Narrowband channel statistics from multiband propagation measurements applicable to high elevation angle land mobile satellite systems", IEEE J. on Sel. Areas in Comm., pp.1219-1226, 1992.

[6] E Lutz, "A markov model for correlated land mobile satellite channels", Int. J. of Satellite Communications, Vol. 4, pp. 333-339, July 1996.

[7] W J Vogel et al, "Land-mobile satellite propagation measurements in Australia using ETS-V and Inmarsat-Pacific", JHU/APL SIR89U-037, August 1989.

[8] H M Aziz et al, "Band Sharing between CDMA based Non-geostationary Satellite-PCNs", IEE 5th International Conference on Satellite for Mobile Communications and Navigation, London, UK, May 1996.

[9] H M Aziz et al, "Comparison of total system capacity for band sharing between CDMA based non-geostationary satellite-PCNs under imperfect power control conditions", IEEE 47th Veh. Tech. Conf., May 1997.

Baseband Processor for FedSat

W.G. Cowley[+], W.N. Farrell[+] and D.A. Powell[++]
Cooperative Research Centre for Satellite Systems
[+]Institute for Telecommunications Research
University of South Australia, The Levels Campus
Mawson Lakes, South Australia, 5095
[++]DSpace Pty Ltd,
Mawson Lakes, South Australia, 5095
Bill.Cowley@unisa.edu.au

ABSTRACT

An Australian microsatellite called "FedSat" is planned for launch in 2000 [1]. This LEO satellite will be used for scientific research and as a technology demonstrator. FedSat will include a Communications Payload (CP) which provides links in the UHF and Ka bands, and includes an on-board modem called the Baseband Processor (BBP).

The BBP will be used during CP modes of operation which involve on-board demodulation and modulation. The paper gives an overview of the FedSat CP and a brief description of the two-way UHF band messaging and the multimedia Ka band packet-switching applications. It includes an overview of modem processing, architecture, implementation and current status.

1. INTRODUCTION

The Cooperative Research Centre for Satellite Systems (CRCSS) was established in January 1998. The core participants of the CRCSS are drawn from Australian Industry, Australian Universities, and the CSIRO (Commonwealth Scientific and Industrial Research Organisation).

The mission of the CRCSS is to deliver sustainable advantage for Australian industries and government agencies based on the applications of small satellites. The participants aim to undertake a targeted research and development program in communication, space science, remote sensing, and space engineering. The Centre's first major space mission will be to develop an innovative scientific satellite, FedSat, and install it in orbit in late 2000 for the Centenary of Federation.

FedSat will be the first Australian scientific satellite mission for more than thirty years. It will be a low cost microsatellite, conducting communications, space science, remote sensing and engineering experiments. The FedSat mission will give Australian scientists and engineers valuable data about the space environment, as well as experience in space engineering and in practical applications of space technologies. FedSat is expected to have a mass of approximately 50kg and reside in a Low Earth Orbit (LEO). Space Innovations Limited (SIL) has been selected to provide the FedSat satellite bus.

The aims of the FedSat CP are to provide both space and terrestrial facilities for the communications research program, to supply communication services for other FedSat research experiments, to provide backup for satellite Telemetry, Tracking, and Control System (TT&C), and to demonstrate LEO satellite communications capabilities.

The CP will provide a two-way communications links in the UHF and Ka bands for applications described in the next section. The Baseband Processor (BBP) provides advanced on-board communications processing for the CP. The BBP is a packet modem with low power and flexible rate operation. DSpace Pty Ltd and the Institute for Telecommunications Research (ITR) at the University of South Australia are responsible for the design, development and testing of the BBP.

Section 2 of this paper describes BBP requirements for the UHF and Ka band applications. In Section 3, selected aspects of modem processing are described, while Section 4 discusses the BBP implementation.

2. BBP REQUIREMENTS

As mentioned in Section 1, the BBP supports two main modes of operation. We now outline the modem requirements for the Ka and UHF band applications.

Ka Band BBP Requirements

For Ka band operation the CP supports both "bent pipe" and on-board processing modes. This paper considers the latter. The aim is to demonstrate flexible packet communications between two or more Ka band earth stations (ESs), tailored to support "multi-media"

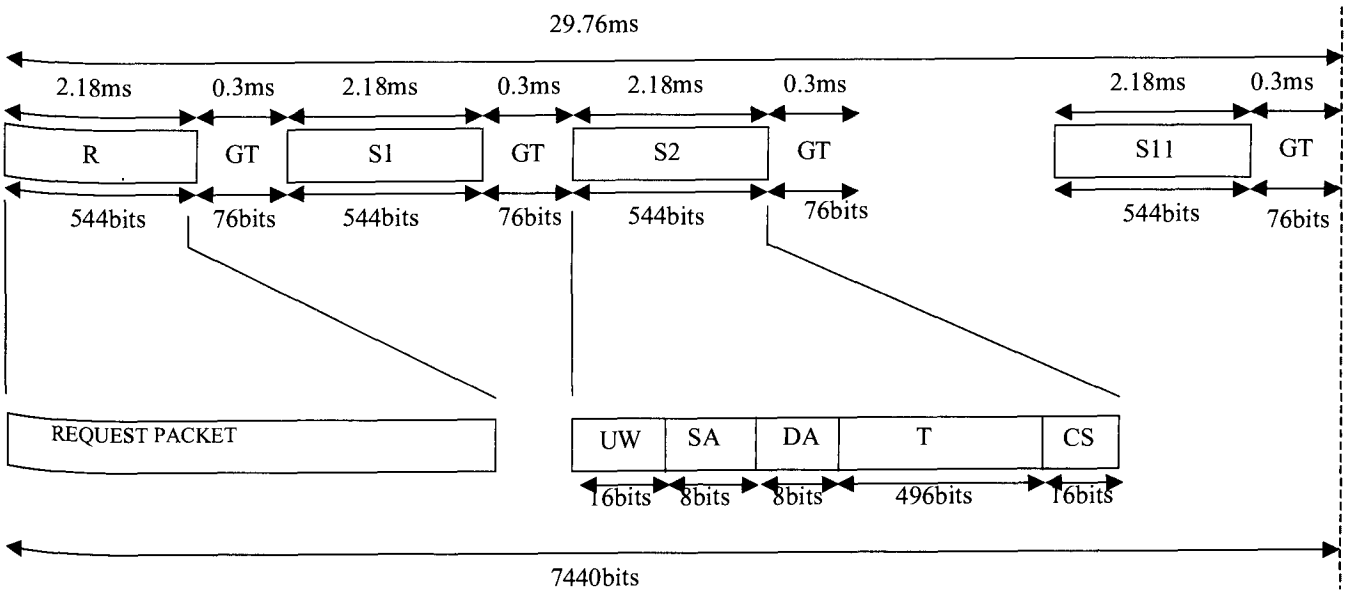


Figure 1 Ka Band Uplink Frame Format

applications using compressed video and audio signals. Bit rate assignment must be flexible enough to cope with a wide range of variable traffic rates from each ES.

To satisfy these requirements the FedSat BBP employs a TDMA uplink and TDM downlink. The frame rate is selected to suit multimedia applications. Figure 1 shows the uplink frame structure for Ka band packet switching mode. Each of the 11 traffic slots may be assigned to a specific ES, generally for a period of about 1 second. These assignments are made in the BBP as a result of reservation requests in an Aloha-style request slot.

Uplink and downlink symbol rates are the same, although there is slightly more downlink capacity due to the uplink overheads such as guard bands. The downlink is a continuous TDM signal since, when low-gain iso-flux antennas are employed on the satellite, there is no advantage to be gained with a TDMA downlink. Each ES derives its frame timing from the TDM downlink The first downlink slot is used for reservation acknowledgments and status information. The remaining slots may be used to forward uplink slots, or for "broadcast mode" applications.

A sample link budget for the Ka band uplink is shown in Table 1. This assumes a tracking ~1m parabolic antenna in the ES. The nominal bit rate is 250kbit/s with rate-half coded QPSK. LEO orbit at Ka band gives rise to doppler frequency offsets up to approximately 700 kHz. This is difficult to handle in a burst demodulator where the bit rate is limited. In order to simplify the processing on-board, the Ka ESs adjust their transmit frequency so that frequency offset is very small at the satellite (eg < 1 kHz). This strategy assumes that each ES knows the position of the satellite sufficiently accurately to predict the doppler trajectory during the pass. Obviously an acquisition procedure is required to determine the initial frequency offset due to LO drift.

Carrier frequ	30000 MHz
Link distance	800 km
Required Eb/N0	10 dB
Tx EIRP	48.5 dBW
Rx G/T	-26.7 dB/K
Link Margin	3 dB
Atmos loss	1 dB
Modem loss	1 dB
Free space loss	180.0 dB
Rx C/N0	65.3 dB-Hz
Info bit rate	340.3 kbps

Table 1: Sample Ka Band Link Budget for FedSat Uplink

In a similar fashion the Ka ES adjusts its slot transmission time relative to the derived frame clock in order that their transmission arrives at the satellite on time. This strategy allows minimal guard bands in the frame structure shown in Figure 1.

UHF Band BBP Requirements

The objective for the UHF band is to implement a two-way messaging system which provides robust data collection and store-and-forward services to low cost mobile terminals (MTs). Efficient use of the narrowband UHF frequency allocations is required and several MTs must be able to access the system at any time.

To satisfy these requirements a TDMA and TDM access technique is also used, with a reservation-based allocation system. Frame periods are much longer, for reasons explained later. Link budgets constrain the transmission rates to about 4 kbit/s.

A number of frame formats are envisaged in this service. For example, the mode primarily suited to remote data collection employs a 4 kbit/s, half-rate coded, QPSK uplink and an uncoded, 1kbit/s, BPSK or FSK downlink. In this case the downlink is mainly used for acknowledgments, status information and short messages. A lower rate downlink allows more link margin suitable for uncoded low-cost MTs, for example for oceanographic data collection applications. Another mode of operation is better suited to messaging between MTs and employs equal uplink and downlink transmission rates of 4 kbit/s.

The on-board SIL data handling system (DHS) computer will be used to store messages from, or to, MTs. Messages received from the "data collection" MTs will be stored in the DHS until they can be returned to the TT&C station in Adelaide via the S band telemetry system.

In the UHF MTs, doppler and slot timing compensation is not possible since the MTs are low cost and don't know their own locations. Consequently the MT must be able to estimate the frequency offset and remove it. Also the TDMA guard bands must be large enough to accommodate variations in packet transmission time.

OVERVIEW OF BBP PROCESSING

This section gives an overview of some aspects of BBP signal processing and simulations. In general, the BBP employs feedforward synchronisation approaches to achieve burst-mode packet processing, plus IF sampling and synthesis (e.g. see [3], [4], [5]).

Figure 2 shows a block diagram of the demodulator processing for the Ka band rate. As previously explained, doppler offsets will be removed in this situation, so the NCO #1 represents a very simple fixed-frequency down-conversion operation.

The feedforward phase estimator shown in Figure 2 accepts one-sample-per-symbol-period samples and estimates the phase offset over a limited number of symbols. Although most of the frequency offset has been removed at this stage, the residual offset still proves difficult to handle due to the significant rate of doppler change.

Figure 3 shows simulation results of one approach to this problem. The phase ambiguity has been resolved from the unique word (UW), and then a sequence of phase estimates have been calculated (via [5]) during the slot duration. An observation interval (N_o) of 12 symbols per phase estimate was employed and the phase estimates were "unwrapped" progressively from the UW phase. The simulation used ideal symbol timing in an AWGN channel. It can be seen that this technique can handle reasonable frequency offset. (The frequency offsets in the figure have been normalised by the symbol rate.) In addition it avoids the penalty of differential coding (eg see [2]).

UHF processing involves a lower symbol rate but is complicated by the need to estimate and remove doppler frequency offsets. A rate-half turbo coding scheme is envisaged on the UHF uplink, with interleaver sizes selected to match the slot durations.

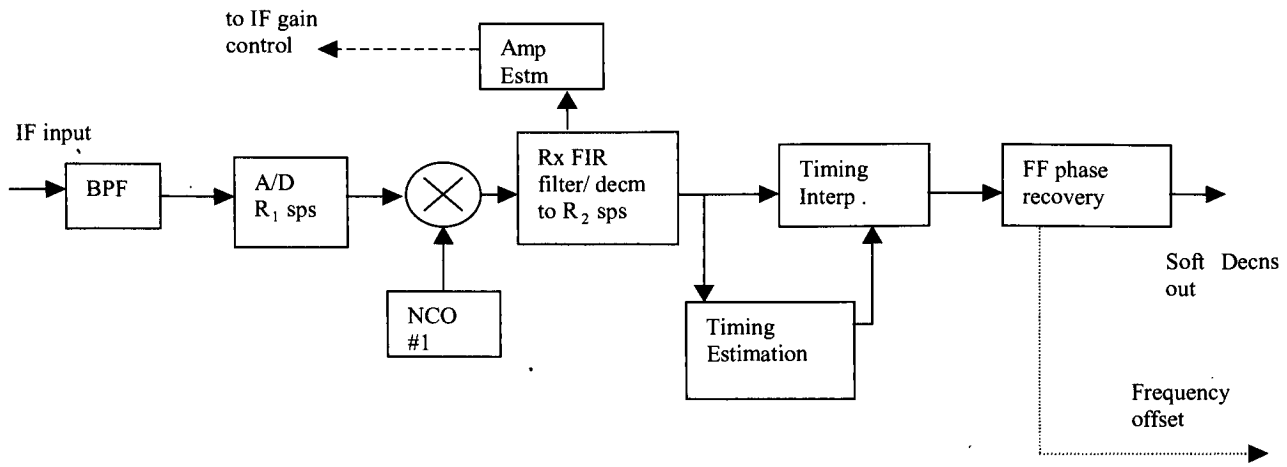


Figure 2 Ka Band Demodulator Processing

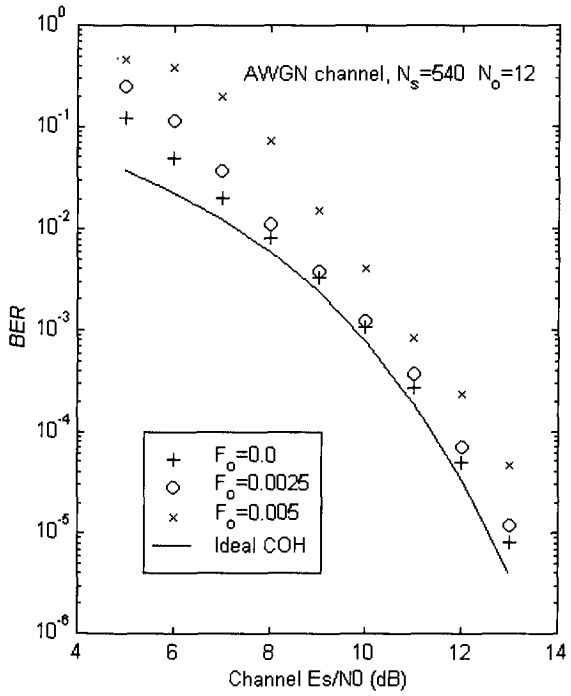


Figure 3 Detector Simulation showing Effects of Frequency Offset

Modeling of the UHF communications channel is being carried out for oceanographic data collection from floating buoys. The UHF maritime channel is modeled as a shadowed Rician channel where the shadowing statistics depend on the ocean state, satellite elevation and pass bearing relative to wave direction. The reservation and ARQ schemes have been simulated using Opnet in order to estimate the overall link capacity. An example of a high elevation pass is shown in Figure 4. In this case the minimum shadowing during the middle of the pass was quite small (about 10%). As might be expected, the arrival rate of valid packets is much lower at the edges of the pass. We are currently trying to determine realistic parameter values for the ocean surface statistics to complete this model.

BBP IMPLEMENTATION

The BBP design is constrained by power, mass, area and radiation specifications, however it is the radiation and power limitations which impact the most on the system architecture and limit the selection of suitable components. In addition to these constraints, the BBP is designed to operate in two, very different modes of operation, thus requiring some flexibility in design. Due to the power budget of 4W, the majority of the BBP will be implemented using FPGAs. There are a number of one time programmable (OTP) and reprogrammable FPGAs which are suitable for space applications, solving some of the constraints.

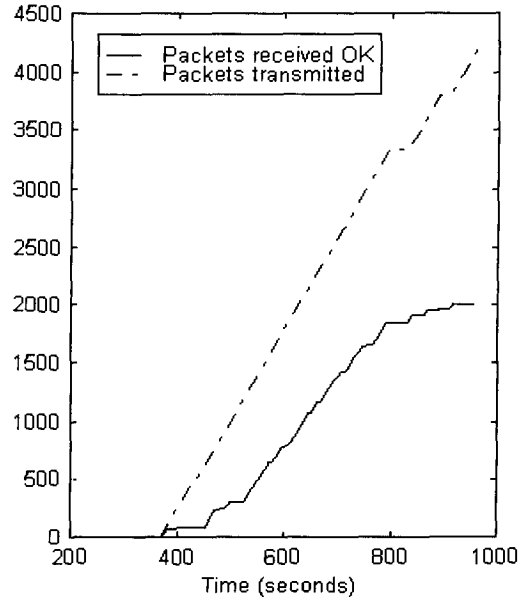


Figure 4 Simulation of ARQ Protocol for UHF MT

The BBP architecture can be divided into three main modules: modulator, demodulator and a data routing processor (DRP). Both modulator and demodulators are to be implemented solely using FPGAs, in order to meet the speed requirements, whilst minimising power usage. To implement the demodulator using a processor, it is expected that a high power DSP would be required which would exceed the power budget.

The DRP is to be implemented using a simple microcontroller as it's main task is to shift data packets, interfacing between the modulator, demodulator and the central storage system onboard the satellite. In addition, the DRP is required to implement simple channel allocation and UHF ARQ protocols. The DRP interfaces to the modulator and demodulator via a single bus, with a DMA facility included to allow for high throughput.

Figure 5 provides a high level architecture of the BBP, highlighting the three modules and indicating the main components. It can be seen that the demodulator will be implemented using 2 FPGAs; Actel (OTP) for prefiltering and Xilinx for the main acquisition and synchronisation tasks. An additional Xilinx FPGA is required for decoding tasks.

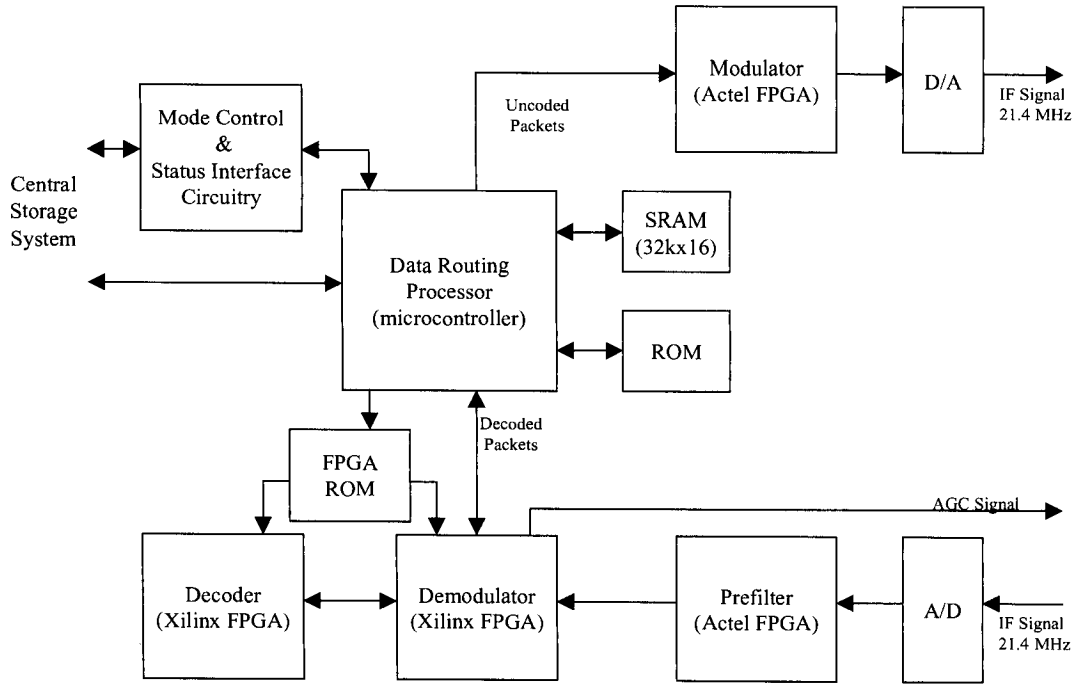


Figure 5 – BBP Architecture

Implementing two modes within the one modem represents a difficult process due to their different requirements. Therefore some flexibility in design is required with different programs expected for the microcontroller and the Xilinx FPGAs. However as the Actel OTP FPGAs will be used for the modulator and prefiltering, it is important to maximise the reusability of code where possible.

The radiation environment is not severe, with the expected total dose of < 10 krads for the 3 year mission. However radiation tolerant components are still required to provide a high degree of reliability and provide some reduction in the number of single event upsets. Despite using radiation tolerant components single event upsets are still likely to occur and thus their effects on operation must be considered during the design phase. This means fault handling of bit flips inside the microcontroller and the FPGAs should be implemented where possible.

CONCLUSIONS

This paper has given an overview of the baseband processor planned for the FedSat microsatellite. The payload implements a flexible packet modem for LEO satellite applications.

The BBP project is currently in the detailed design phase. FedSat is due to be launched in late 2000.

ACKNOWLEDGMENTS

This work was carried out with financial support from the Commonwealth of Australia through the Cooperative Research Centres Program. The South Australian Government has also provided financial support to DSpace for this project.

REFERENCES

[1] CRCSS www: http://www.crcss.csiro.au/fedsat1.htm

[2] W.G. Cowley, "Reference Symbols can improve performance over differential coding in ML and near-ML detectors", Signal Proc 71 (1998) pp. 95-99

[3] M.J. Miller, B. Vucetic, L. Berry, "Satellite Communications: Mobile and Fixed" Kluwer, '93; Chapter 5

[4] Oerder, M. and Meyr, "Digital Filter and Square Timing Recovery", IEEE Trans. on Comms, Vol. 36, No. 5, May 1988.

[5] Viterbi, A.J. and Viterbi, A.M., "Nonlinear Estimation of PSK-Modulated Carrier Phase with Application to Burst Digital Transmission", IEEE Trans. on IT, Vol. IT-29, No. 4, July 1983

Demodulation and Discrimination in Mobile Satellite Systems

Bassel F. Beidas, Ludong Wang
Hughes Network Systems
11717 Exploration Lane, Germantown, MD 20876
Email: bbeidas@hns.com

ABSTRACT

We address a demodulation receiver that could be implemented in mobile satellite digital communications systems. The advantage of implementing this proposed receiver is demonstrated in terms of the bit error rate of the overall demodulation performance. It is tested under varying channel conditions including mildly to severely frequency selective fading, and the analytical and simulated results are reported.

INTRODUCTION

In this paper, we introduce a joint bit timing synchronization and carrier frequency offset estimation algorithm that is based on the unique words inserted in bursts that employ constant envelope modulation formats. It is derived from statistical decision theory and is shown to be robust under various adverse channel conditions. In addition, we introduce channel estimation technique to extract the time-varying random fading parameters experienced in typical mobile satellite transmission. This procedure is necessary to successfully aid in coherent bit retrieval.

In the physical layer of a typical system there exists a set of channels. The first is a traffic channel used for transmission of encoded speech or data between the user and the network. The other is a control channel dedicated for conveying signaling functions. It is assumed here that the multiplexing of the channels is achieved via a Time Division Multiple Access (TDMA) scheme. On the uplink, the modulation technique employed could be a Gaussian Minimum Shift Keying (GMSK) for its constant-envelope property. The constant envelope property is highly desirable as it allows the handheld terminals to use cost-effective non-linear power amplifiers that operate in full saturation. This property however introduces memory into the modulation which could increase the complexity depending on the bandwidth efficiency required. Further, the traffic and control bursts could contain groups of unique-word (UW) symbols that are evenly distributed across the burst.

The communications link model considered varies from a

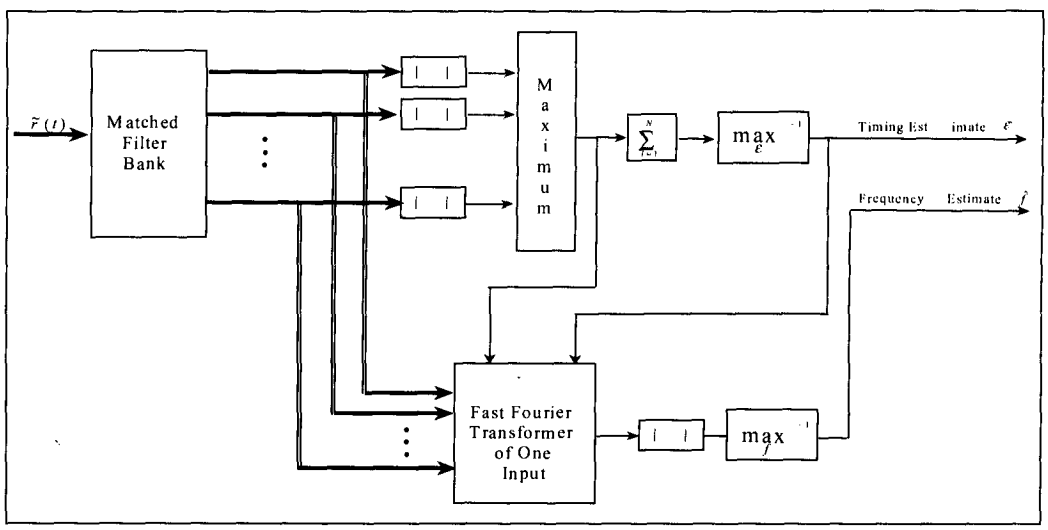


Figure 1. Architecture of timing and frequency synchronization circuitry.

static additive white Gaussian noise (AWGN) to a frequency-selective multipath one that follows a Rician model. In the latter, the Rician K factor (the ratio of direct path power to total multipath power) and the fading bandwidth are varied to cover the situations of a user terminal (UT) held by a walking user to one that is aboard a travelling vehicle.

algorithm to simultaneously identify the burst type such as control bursts that could preempt the normal traffic.

Noteworthy in this respect is that the derivation of such an algorithm is justified from a statistical decision-theoretic viewpoint. In other words, it results from the application of the average likelihood-ratio function (ALF) when the optimal non-linearities are simplified using mathematical series approximations.

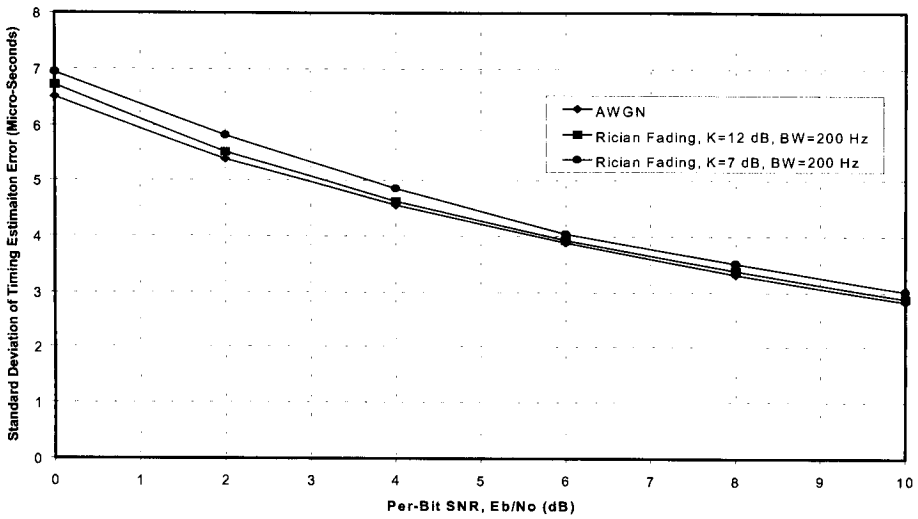


Figure 2. Timing estimation performance.

TIMING SYNCHRONIZATION

The timing synchronizer circuitry is depicted in the upper portion of Figure 1. In this figure, the first operation involved is a matched filter bank to correlate against the UW intervals. A bank of filters is needed in this case because of the variation in the waveform appearance due to the memory inherent in the constant-envelope GMSK modulation. These outputs are then processed through an envelope detector, a maximum operation and then an accumulator. This is done to combine the contributions of the distributed UW's in a non-coherent manner. Non-coherent combining is important as it produces performance that is robust with respect to the channel fading conditions. It also allows for reducing the computational complexity as the timing estimation can precede the frequency estimation procedure. This consequently reduces the joint search in the time-frequency plane to two disjoint searches in one-dimensional planes. The symbol timing estimate is then the value at which the accumulator output is maximized. The resolution of the timing estimate will depend on the number of samples per symbol afforded by the hardware. It can however be enhanced using interpolation techniques such as the Lagrange method. By correlating against different UW patterns, it is possible for the above

The standard deviation of the timing estimation error is illustrated in Figure 2 under AWGN and two Rician-faded channel conditions.

FREQUENCY OFFSET ESTIMATION

The frequency offset estimator uses the matched filter bank output as depicted in Figure 1. It can be seen that the correlation operation against the distributed UW's in principle transforms the GMSK burst into a set of discrete-time samples of a single-tone. For this, it is essential that the individual UW's be short enough in duration so that the phase rotation due to the frequency uncertainty over a UW interval is small. The matched filter output along with the timing synchronizer output is then processed through a fast Fourier transformer (FFT). The frequency estimate is based on the value at which the magnitude of the FFT is maximized. The resolution of the frequency offset estimate depends on the duration of the GMSK burst, while the search range depends on the separation between the UW's within the burst. The FFT of the single-tone is the optimal method for estimating the frequency offset.

The performance of the proposed algorithm is shown in Figure 3, measured in terms of the standard deviation of the frequency estimation error under different channel conditions.

It is well known that non-linear estimation suffers from drastic degradation in performance below some critical SNR value. This performance degradation can be attributed to the presence of large subsidiary noise spikes which cause the estimator output to peak at values that have no relation to the true value associated with the transmitted signal. When that occurs, bounding techniques such as the Cramer-Rao rule fail to be accurately representative of performance, and effort should be expended in trying to incorporate the large noise spikes associated with the strong noise condition. The modified performance bound is illustrated in Figure 3. To remedy the situation and improve performance, one can employ a recursive loop that tracks the frequency variation across multiple bursts and utilize this knowledge to progressively limit the FFT search range.

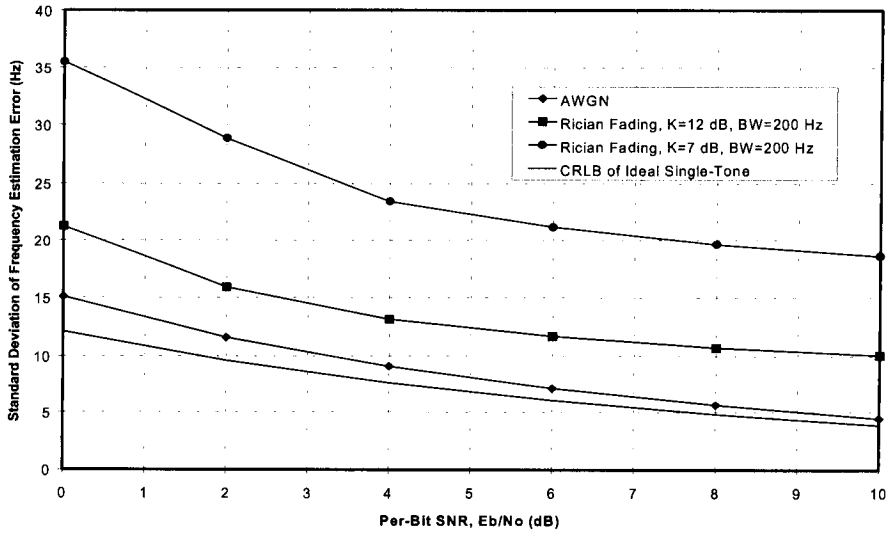


Figure 3. Frequency estimation performance.

FADING CHANNEL ESTIMATION

Rician fading is a typical channel condition in mobile satellite transmission systems. With random variation in both amplitude and phase, fading incurs performance degradation to the extent that it can not be ignored in most digital receiver applications. This is especially true when coherent demodulation is employed. Estimation and compensation of such random channel variation is indispensable.

Although channel fading is reflected by the received signal, it is not practical to perform the estimation over all symbols of each burst due to the stringent need of removing constant-envelope modulation. Data-aided estimation procedure is implemented instead.

As described previously, multiple UW's are transmitted with each burst. After matched filtering, the channel fading is theoretically retrievable from received UW symbols. Since variation of fading during each UW can actually be ignored, the filtered UW's are virtually discrete fading samples located at the multiple UW's. Based on the deviation between the filtered UW's and their presumably known patterns, the fading estimation is thus obtained as the least-square solution.

On the other hand, the above estimation is actually a noisy observation of the channel fading. To improve the estimation, the direct sample estimation is filtered with moving-average. By linear interpolation, the unknown fading variation over the whole burst can be inferred from the newly obtained finer estimation. Although various curve-fitting approaches are readily available for interpolation, simulation demonstrates that the optimal estimation is obtained with our proposed linear interpolation.

Depending on the receiver structure in application, the estimated fading patterns are applied to the received signal for compensation of either phase or amplitude, or phase only.

Figure 4 depicts the simulated demodulation performance in terms of the bit error rate including the timing and frequency errors due to the estimation algorithms. It is shown that fairly good performance can be achieved under

Rician fading channel conditions given the hardware implementation constraints.

CONCLUSIONS

An algorithm for providing timing synchronization and carrier frequency offset estimation was proposed. Its performance was tested in different channel conditions, representative of ones encountered in mobile satellite scenarios. The modulation scheme considered contains inherent memory, for example due to reasons of maintaining constant-envelope property such as GMSK. The overall demodulation results were shown when implementing the proposed scheme. The application of the algorithm can be extended to many other digital communications platforms.

REFERENCES

[1] B. F. Beidas, "Joint Burst Classification and Symbol Timing Estimation for Mobile Satellite Applications," *Hughes Network Systems Internal Memorandum*, August 13, 1997.

[2] B. F. Beidas, "Frequency Estimation for Mobile Satellite Applications," *Hughes Network Systems Internal Memorandum*, August 27, 1997.

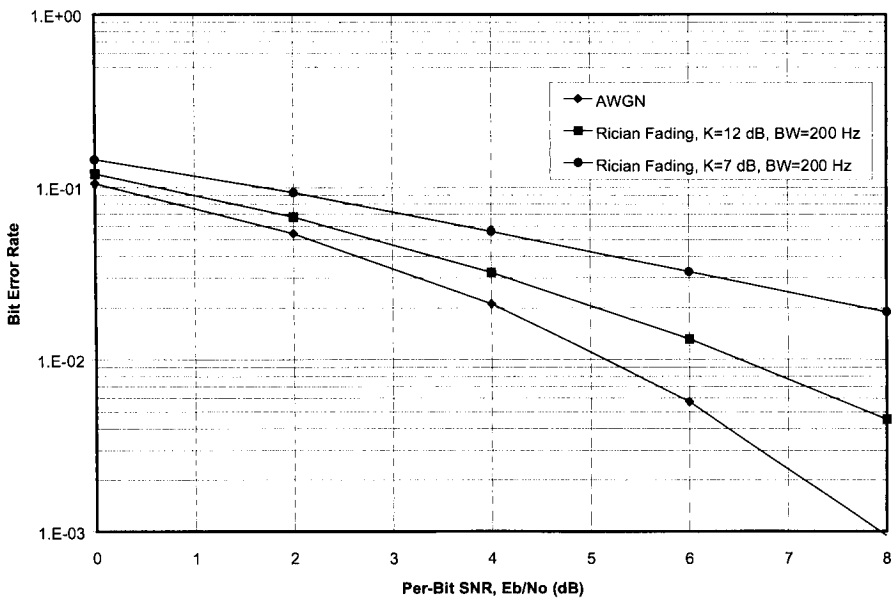


Figure 4. Bit Error Rate (BER) Performance

BCCH Processing for Mobile Satellite Communications

Zhen-Liang Shi
Hughes Network Systems
11717 Exploration Lane, Germantown, MD, 20876, USA
Email: lshi@hns.com

ABSTRACT

In this paper, we describe the basic signal processing elements for the broadcast control channel (BCCH) of mobile satellite systems. Emphases will be given on the detection, estimation and demodulation processes. In addition, some theoretical and simulation results are presented.

1. INTRODUCTION

In mobile satellite communication, the BCCH, which is a unidirectional channel from a satellite gateway station to user terminals, broadcasts general information on a beam basis. In addition, it also provides complementary information for user terminals' synchronization.

BCCH bursts for a beam are transmitted periodically on a pre-assigned frequency carrier with a much lower transmission rate compared to the traffic channel (TCH) bursts. Therefore with the same modulation rate, there must be a long idle period between any two consecutive BCCH bursts. Figure 1 reveals a transmission format of BCCH. From this figure, the BCCH bursts are transmitted at the period of one BCCH frame consisting of 24 time-slots. Each BCCH burst occupies one slot which is made of a number of modulation symbols. In the Hughes Geo Mobile (GEM^{TM}) satellite system [1], some slots in between two consecturive BCCH bursts may be used to transmit a synchronization burst named frequency correction channel (FCCH). The FCCH is a chirp modulated burst in GEM. CW burst can be also used for the same purpose. In this paper, we do not use any synchronization burst. Instead, detection, frequency and time estimation are all based on the BCCH bursts. In this paper, we assume that BCCH is BPSK modulated.

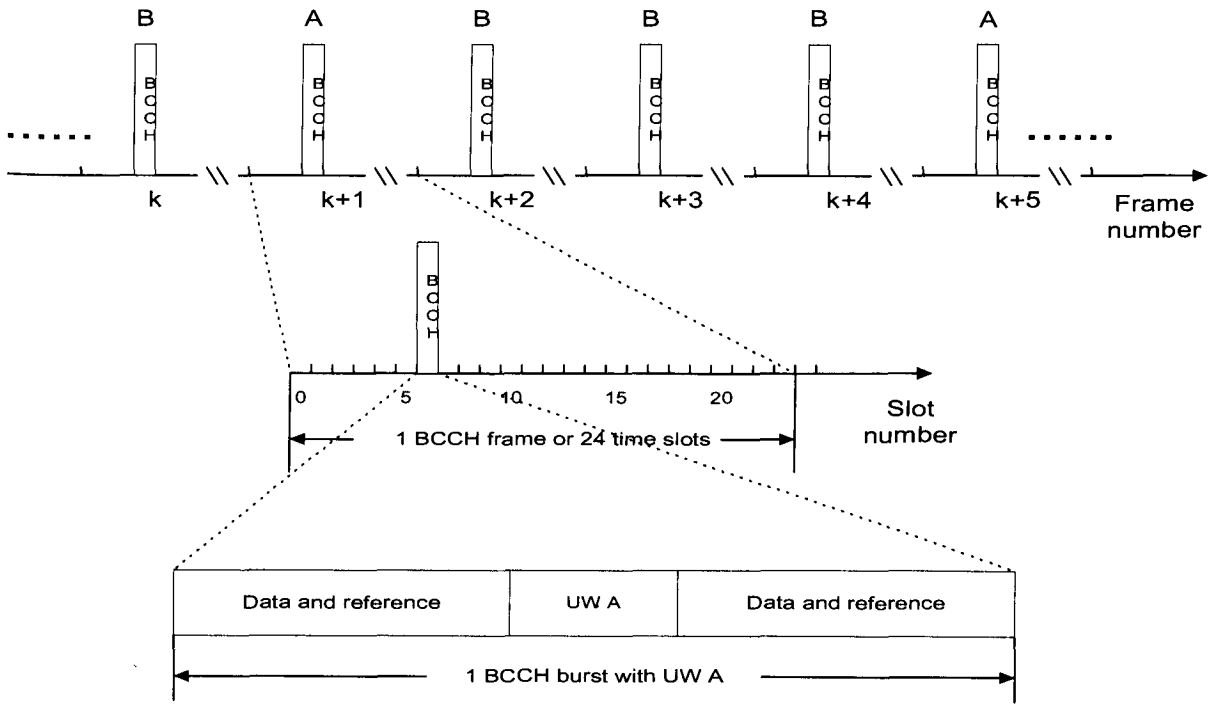


Figure 1. BCCH transmission format

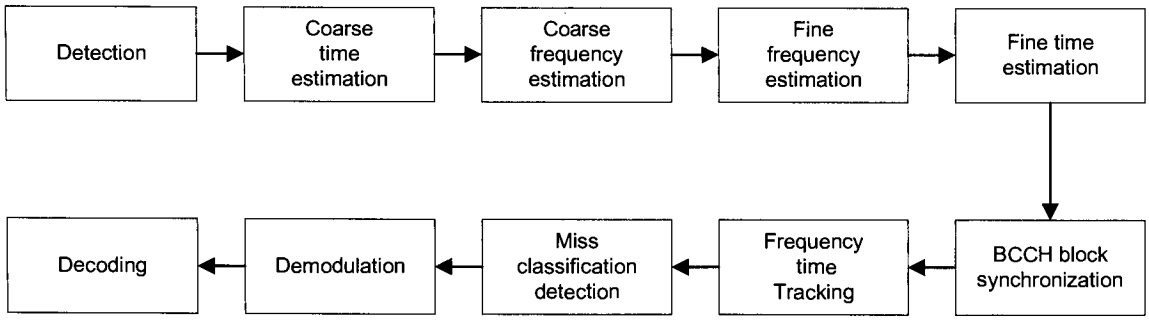


Figure 2. Block diagram of BCCH processing

A complete BCCH data information block may consists of several BCCH bursts. These bursts carry their unique words (UW) in certain pattern, such as ABBB for a block made of 4 bursts, where the burst carrying A is considered to be the first burst of the block. A and B are two different unique words with a small cross-correlation. It is clear from diversity point of view that transmitting a BCCH block over a few frames on which the fading profiles are independent is better than transmitting the block in one frame.

The BCCH block synchronization can be achieved by checking the UW pattern of at least 3 consecutive BCCH bursts. One pair of UWs may be constructed by two combined Barker sequences, i.e., A=[BK BK] and B=[BK –BK], where BK is a Barker sequence. The cross-correlation of the two UWs is zero.

A user terminal (UT) has a non-volatile storage of BCCH carrier information. When the UT powers on, it initially limits its search to the BCCH carriers included in the storage list, and then locks to the strongest one, which is most likely from the beam that the UT is in. The selected BCCH carrier is further processed for frequency and time information. With the precise frequency and time reference, the UT is able to conduct demodulation and decoding.

The block diagram in Figure 2 shows the BCCH processing elements. In the following sections, we describe these blocks in more detail.

2. DETECTION

The initial time and frequency uncertainties can be very large during the detection. The time uncertainty comes from the long idle period of the BCCH transmission, and

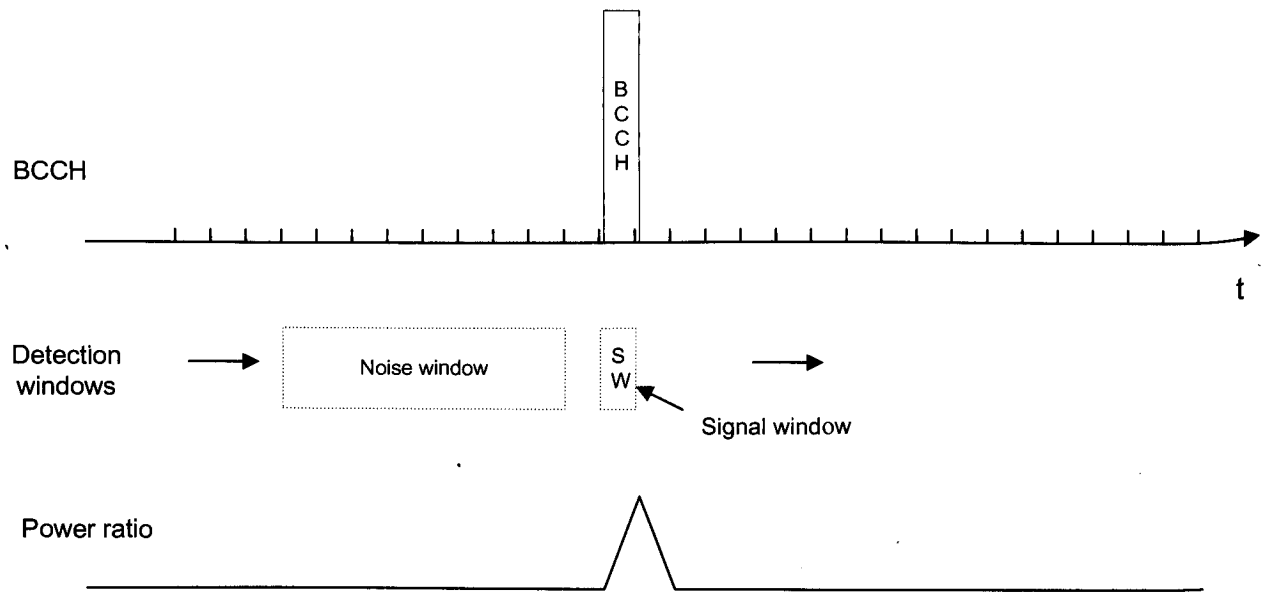


Figure 3. Power ratio detection for BCCH

the large frequency uncertainty is mainly caused by the instability of UTs' oscillators. The detection filter has to be large enough to accommodate the frequency offset.

Burst envelope detector or power detector may be used to acquire a BCCH burst. The problem is how to determine the threshold of the detector. In reality, the signal as well as the noise floor may all vary. Though an optimal threshold can be found when the noise floor is known, the threshold optimized may not be optimal all the time since the noise floor could change significantly. Besides, the estimate on the noise may be far from accurate if it takes samples from the BCCH signal. This is true since timing is not available during the detection. Figure 3 illustrates a sliding power ratio detector proposed for BCCH signal detection. In this figure, the signal window tries to measure the power of the BCCH burst and the noise window tries to measure the noise floor in the idle period. The ratio of the two measurements is compared with a threshold. If the threshold is exceeded, a tentative BCCH carrier is found and its power level is stored. Then the UT will test other carriers using the same method until all the carriers on the list have been tested. The carrier corresponding to the largest power level is chosen for further signal processing. If the test ratio does not exceed the threshold, the UT just slides its two windows by one sample and repeats the ratio test. The new signal power and noise power can be easily computed recursively. If the test has been repeated for one whole frame, but no BCCH signal is found, the UT shall test a new carrier. Using this method, the threshold can be easily optimized for the minimum operating SNR, and the threshold does not need to change as the noise floor varies.

The detection performance may be measured by the probability of missing a packet when the signal window aligns with the BCCH burst and the probability of false detection when the signal window contains noise only. It is easy to show that the test variable has a singly-non-central F distribution [2]. For the proposed ratio detector, these probabilities are plotted in Figure 4. In this figure, we see that the longer the noise window, the better the detection performance.

3. TIME AND FREQUENCY ESTIMATION

Time and frequency estimations for BCCH are conducted in the order as described in Figure 2. First we obtain a coarse timing from the detected burst. With this timing, we can estimate the large part of the frequency offset. This frequency offset can be removed for the new bursts to pass a much tighter detection filter. Then the fine frequency and time estimations are conducted on the filtered signal.

3.1 COARSE TIMING AND FREQUENCY

Upon a detection of a BCCH carrier, the UT can roughly measure the burst timing by looking at the power profile of the power ratio in time domain. The time instant corresponding to the maximum power output is used as coarse burst timing. In practice, the timing thus obtained is

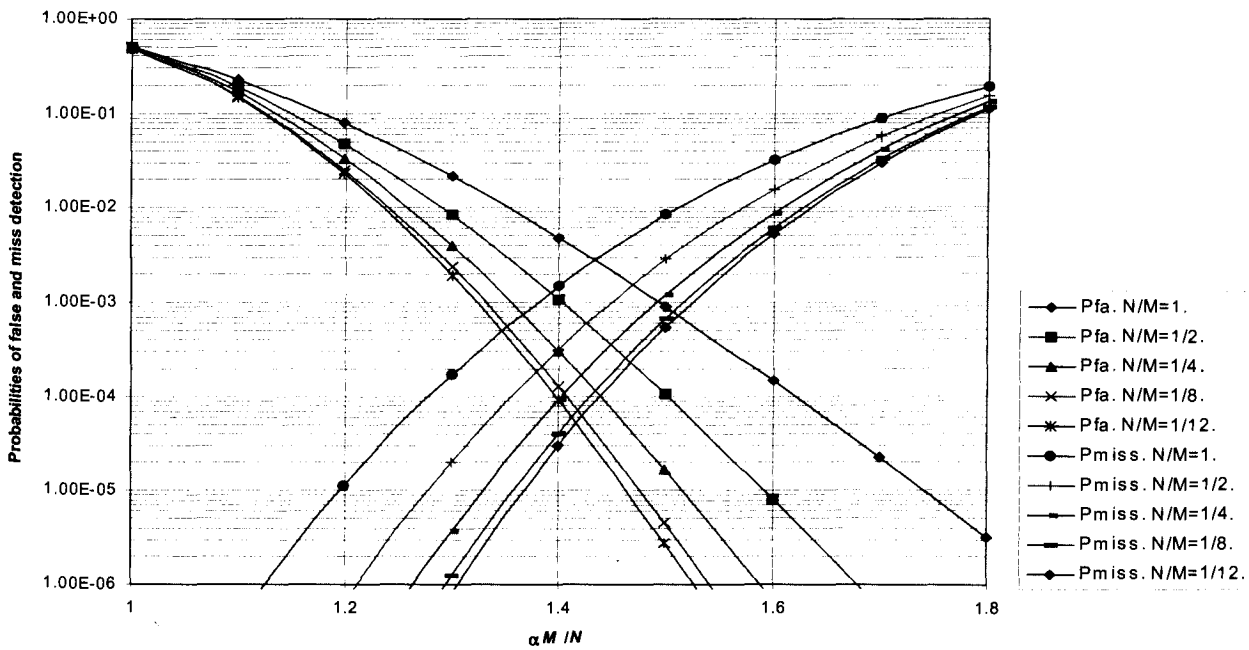


Figure 4. Probabilities of miss and false detection for power ratio detector. *Es/No*=0 dB. *N/M* is the ratio of the signal window length to the noise window length.

far from accurate at low SNR. Multiple bursts are used to improve the time estimation.

The coarse frequency estimation also requires a multi-burst processing. It simply uses the same bursts that the coarse timing has used. Since we use a wide detection filter, both the estimators are not optimal and they can produce some very large estimation errors if the SNR is below a certain level. This is noticed as the threshold effect. Multi-burst processing is used to combat the effect. The number of bursts used for this purpose can be adjusted according to the channel condition. If the channel is really poor, we use more bursts; otherwise, fewer bursts.

3.2 FINE TIMING AND FREQUENCY

With the coarse frequency estimation, a large portion of the frequency offset is removed before a new burst is passed to the tight filter. The output of the filter is squared to remove its BPSK modulation. Then its residual frequency offset is estimated by zero-padding DFT with much finer frequency bins [3].

The fine time estimation is based on the correlation measurement of the unique words. Two UW matched filters are operated all the time. Again multi-burst estimation is used for the fine timing in order to improve the timing performance in fading channel.

4. BCCH BLOCK SYNCHRONIZATION

The block synchronization is a process for recognizing the UW pattern of the received BCCH bursts. If 4 consecutive bursts carry a UW pattern of ABBB, they are recognized as a complete BCCH block. The BCCH decoder always takes the soft decision values of a whole block.

BCCH block synchronization is conducted together with the multi-burst fine time estimation. There are always 4 pattern candidates in the sync process, such as ABBB..., BBBA..., BBAB... and BABB.... After the multi-burst correlation, we select the pattern which yields the maximum correlation value. The time estimation is measured based on the detected UW pattern.

Table 1 lists the simulation results for BCCH detection, estimation and the block synchronization. Different initial frequency offsets are considered in the simulation. Because the detection filter cannot have a flat frequency response up to very high freuqncy, the number of miss detection is higher for the burst with large frequency offset.

5. DEMODULATION

BCCH demodulation includes frequency and time tracking, miss block sync detection, channel estimation and bit demodulation. The frequency and time tracking can be done in block basis or multi-block basis. For frequency tracking, single burst frequency estimation is good enough for the tracking loop input. For the timing, multi-burst estimation is necessary for the loop input.

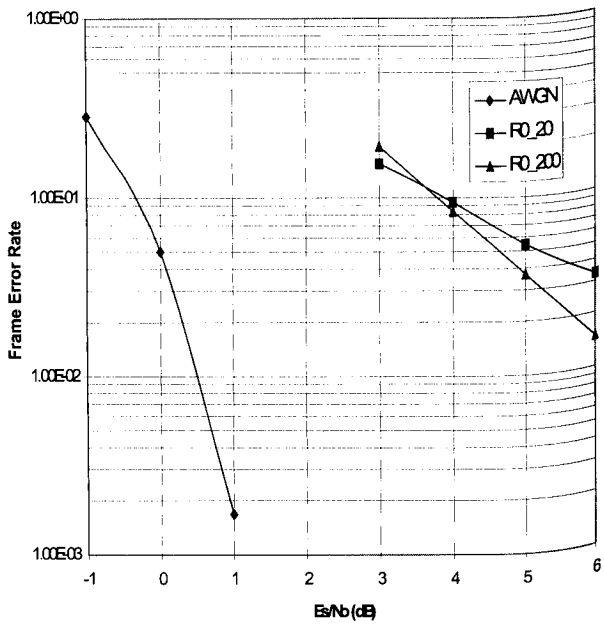


Figure 5. Simulated frame error rates

As shown in Table 1, the block synchronization can be wrong. The miss sync detection is employed during the demodulation process. This is done by monitoring the correlation of A and B matched filters. If the number of

Table 1. Simulation results for BCCH detection, estimation and block synchronization

# of bursts tested	Es/No (dB) in AWGN	Initial frequency offset (Hz)	# of misses	# of wrong block sync.	Standard deviation of frequency estimation (Hz)	Standard deviation of time estimation (μs)
5000	0	-53	59	1	7.69	3.34
5000	0	4035	135	1	7.63	3.34
5000	0	7933	800	0	7.42	3.31

miss matches for a number of bursts exceeds a threshold, the miss sync is detected; otherwise, the block sync is considered to be right.

Channel estimation has to consider the two extreme channel conditions. One of them is the AWGN or static channel with very low SNR. The other is the Rician fading channel with K=0 dB and fading bandwidth of 200 Hz. Block phase estimation is recommended for the channel estimation. Figure 5 shows the simulated frame (block) error rate.

6. CONCLUSION

In this paper, we present the detailed signal processing procedures for BCCH channel of mobile satellite systems. We proposed a power ratio detector for burst acquisition, which eliminates the requirement for computing the optimal threshold from time to time. We also proposed the multi-burst estimation process for frequency and time. And finally we recommend the block phase channel estimation, which yields satisfactory FER performance.

7. ACKNOWLEDGMENT

The author would like to thank Y. Vasavada, M. Tseytlin and Y. Antia for helpful discussions. Special thanks are due to W. Lu for firmware implementation and test.

REFERENCES

[1] GEM 05.02 (HNS): GEM™ *Satellite Telecommunication System (Phase 2)*; Multiplexing and Multiple Access.

[2] R. Price,"Some non-central F-distribution expressed in closed form," Biometrika, Vol. 51, pp. 107-122, 1964.

[3] A. V. Oppenheim, R. W. Schafer, Discrete-Time Signal Processing, Prentice Hall, Englewood Cliffs, New Jersey, 1989.

Theory and Design of an Advanced Multi-User Bandwidth-On-Demand Mobile Communication System Using Tree-Structured and Polyphase Filter Banks

M. Sablatash, J. Lodge
Communications Research Centre
3701 Carling Avenue
P.O. Box 11490, Station H
Ottawa, Ontario, Canada K2H 8S2
Email: mike.sablatash@crc.ca

ABSTRACT

The main problem addressed is the spectrally efficient multiplexing of offset quadrature phase shift keyed (OQPSK) or digital vestigial sideband (VSB) signals for mobile satellite and personal communications applications. A set of requirements are set down for an advanced mobile out-bound communication system. A system is described that has its genesis in wavelet packet-based tree-structured synthesis quadrature mirror filter (QMF) banks for multiplexing signals at the transmitter, and corresponding analysis QMF banks for demultiplexing signals at the receiver. The underlying theoretical principles and the progress that has been made, including a novel synchronization scheme, in realizing a system design based on such a multiplexer–demultiplexer pair constructed of tree-structured QMF bank pairs that meets these requirements for out-bound transmission are described with the aid of computational examples and computer simulation results. Future work on error control using turbo codes and design of the system for in-bound communication is proposed.

INTRODUCTION

This paper is on communications signal processing applied to the theory and design of multi-user systems for mobile satellite, personal and multi-media communications which are spectrally efficient, minimally interfering, have significant bandwidth-on-demand capability, and yet are reasonably efficiently and economically implementable. It focuses on a concise exposition of the design of the digital signal processing for a wavelet packet-based synthesis quadrature mirror filter (QMF) bank tree [1, 10] at the transmitter for multiplexing digital signals. The receiver consists of the corresponding matching analysis QMF bank tree. We describe studies that have been undertaken at CRC to provide new results on how to design a multi-user communication system that significantly improve the spectral efficiency of future transmission systems with an FDM/FDMA component (perhaps in combination with time division and/or code division techniques). This is achieved, in the out-bound direction, along with other important benefits, by allowing significant spectral overlap between adjacent channels and applying orthogonal multiplexing (based on the orthogonality of signal paths through wavelet packet trees in this case). A number of potentially important applications for such a multi-user system have been identified, so the search for improvements of such designs is important for current and future system developments.

An Example System Configuration for an Application [15]

As one application among the considered systems that could advantageously use the features of the system design described herein, terrestrial transmission facilities are connected by digital trunks to a satellite ground station as in Fig. 1. In the ground station, a digital to FDM converter creates a composite FDM signal in which the sidebands for individual channels overlap, creating spectral efficiency. This composite FDM signal is sent to a satellite that relays it to mobile stations. This direction of transmission is called the out-bound direction. In the mobile station a receiver demodulates the FDM signal associated with its carrier. The channels and bandwidths are assigned using a dedicated channel.

Such a system is attractive because it requires no on-board satellite processing, permits a simple per channel mobile station transmitter-receiver, and is spectrally efficient. There is ground station processing, and the multicarrier system probably requires backoff of the satellite amplifiers to avoid saturation.

REQUIREMENTS ON MULTIPLEXER-DEMULTIPLEXER

Our studies of properties of multiplexer-demultiplexer pairs based on the wavelet packet synthesis and analysis

trees of QMF pairs [1-7, 10] for multiplexing (as in Fig. 2) and demultiplexing binary signals, respectively, and matching wideband VSB filters at the roots of the synthesis and analysis filter banks, made it possible to envision a set of requirements that a multiplexer in the

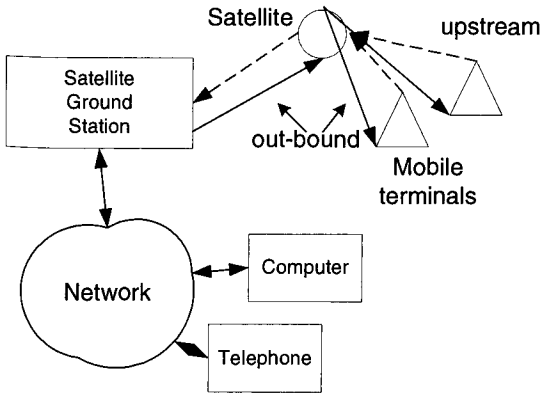


Fig. 1: A typical communication application is shown in this figure to illustrate how this fits into an overall network.

transmitter and a matching demultiplexer in the receiver must satisfy, and to realize typical design examples which satisfy many of these requirements [8-16]. In the frequency division multiplexing of streams of digital signals for personal communication systems such as mobile satellite communication systems it has become a high priority to use spectrum-efficient schemes which also meet a number of other performance and implementation requirements. These requirements include (with references to publications on our studies in which each of the requirements have been addressed or illustrated):

1. The theoretically perfect orthogonality of the signals that are received at the outputs of the demultiplexer when there are separate input signal streams into the multiplexer. The orthogonality means, in communication terms, that the receiver is matched to the transmitter so there is no intersymbol interference and no interference from other channels along a path from the input to the multiplexer to the output of the receiver. In this paper the signal inputs into the multiplexer are assumed to be real. The equivalence of digital Vestigial Sideband (VSB) to Offset Quadrature Phase Shift Keying (OQPSK) enables the studies of VSB to apply to OQPSK [8].
2. The magnitude frequency responses of each of the multiplexer channels from the leaves to the root of the wavelet packet-based filter bank tree used here as the multiplexer (as in Fig. 2), consist of two channels, symmetrically situated with respect the origin of frequency, one at positive frequencies and one at negative frequencies. When these channels are modulated to radio frequencies the positive and negative frequencies around the unit circle are mapped to opposite sides of the carrier frequency, so the signals in the negative and positive parts of each multiplexer channel will generally experience different channel degradations. It is clearly desirable to keep the information for each multiplexer channel in either a positive or negative frequency band [8].
3. The paths taken by the input signals through the multiplexer, to the multiplexer output are such that the magnitude frequency responses along these paths have a stopband attenuation that is greater or equal to a specified minimum value. These multiplexer paths are referred to herein as multiplexer channels. This requirement on stopband attenuation is necessary to keep crosstalk from other channels and out-of-band interference from other channels below a specified value [9].
4. The whole of the frequency axis around the unit circle must be utilized and the packing of the magnitude frequency responses of the multiplexer channels for both positive and negative frequencies must be as efficient as possible to conserve the spectrum [8-16].
5. Because of the growth of multimedia services of widely differing bandwidths, as much bandwidth-on-demand capability as possible must be provided [8-16].
6. The minimization of overall system delay by using a concatenation of wavelet packet-based filter bank trees and DFT polyphase filter banks (as in Fig. 3, and related Figs. 4 - 6) to trade off number of levels of bandwidth on demand and overall filter length from the input of the transmitter to the output of the receiver [12,13].
7. Design of a simplified receiver (as in Fig. 6) for the case in which the filter designs have non-linear phase [12-14], and for which a further simplification in the receiver is effected when all the filter designs have linear phase, as described in [15], and briefly in this paper.
8. The system must have sensitivities to carrier phase (see Fig. 7) and symbol timing offsets (see Fig. 8) that enable design of a synchronization scheme without overly difficult specifications [12, 14].
9. The system should have low sensitivities of carrier phase and symbol timing offsets to channel rolloffs [14].
10. For reasonable errors in phase and symbol timing offsets, the system must have BER performance that can be compensated by feasible error control schemes [14].
11. A synchronization scheme with suitable performance can be designed [15, 16].

GENESIS OF CONCEPTS AND DESCRIPTION OF SYSTEM DESIGN

We were motivated originally by the orthogonality and other properties of wavelet packets [1, 10] that seemed to have potential applications to communications, further inspired by insights into such applications from an early work by Learned [2], followed by [3], and our later studies of [4-7], to design multiplexer-demultiplexer pairs of trees of quadrature mirror filter (QMF) pairs with many desirable properties [8-10, 11]. The relationship of these filter bank trees to wavelet packet trees is derived and discussed in [10]. In [10] survey of the theory and applications of scaling functions, wavelets and wavelet packets in communications is also given.

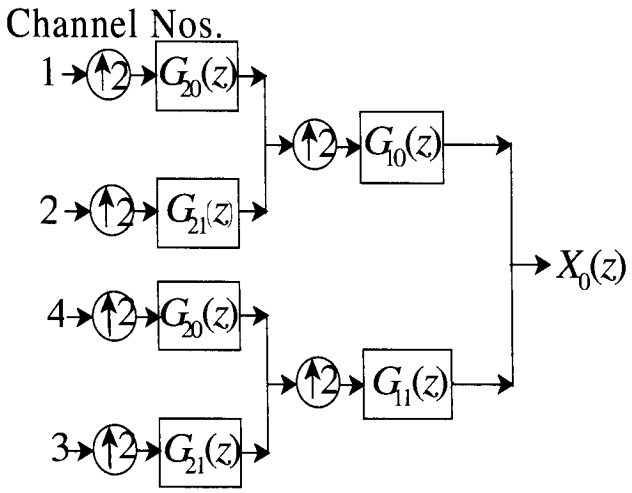


Fig. 2: The first basic 4-input wavelet packet-based synthesis filter bank repeated 3 times, with different channel numbering each time with successive output signals at the roots as the input signals $X_1(z)$, $X_2(z)$ and $X_3(z)$ to a DFT polyphase synthesis filter bank shown in Fig. 3.

In [8] the equivalence of Offset Quadrature Phase Shift Keying (OQPSK) and digital Vestigial Sideband (VSB) is proven, enabling the use of real inputs in all our subsequent studies, and descriptions and discussions in terms of VSB concepts.

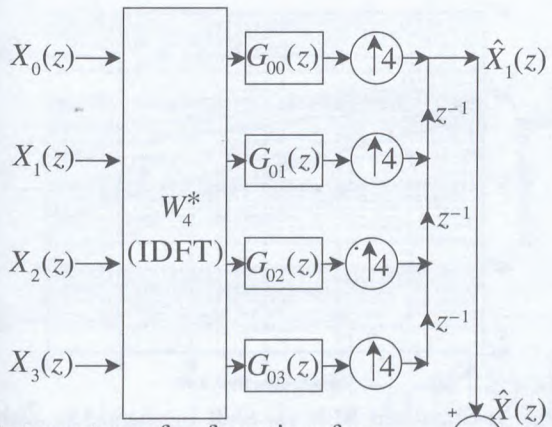
Two designs for the filters in the QMF pairs that are used to construct the filter bank trees are described in [9]. In both of these designs the minimum stopband attenuation in multiplexer channels from the leaves, or from an intermediate node of the tree, to the root of the tree, can be specified at a prescribed value for which the required filter design or designs can be found. In one of these designs the same QMF pair is repeated throughout the tree, as required for the standard definition of the wavelet packet tree. In this case the multiplexer channel magnitude frequency responses are not symmetric about the centres of their passbands. In the other case, the QMF pair designs are replicated at each level of the tree, but the designs at each level are different. The total number of coefficients in the path corresponding to a multiplexer channel is substantially lower, by about one-half, than in the previous case. Since the 3-dB bandwidth is the same in the two cases, and it is convenient to use this as the measure of essential bandwidth, the tiles in the time-frequency plane remain the same in height, but differ in width. Thus, this second design no longer gives the wavelet packet tree that is standard, but an extension thereof which we have employed in most of our subsequent studies. The main reasons for our preference is that this design yields fewer total filter coefficients in the tree, and the multiplexer channel magnitude frequency characteristics are symmetrical about the centres of their passbands.

In [8-10] we have also shown how to design a wideband VSB filter ($G_{10}(-jz)$ in Fig. 2) following the root of the tree to eliminate (attenuate to below the minimum specified multiplexer channel attenuation) signals in all the multiplexer channels at negative frequencies and to pass the signals in all the multiplexer channels at positive frequencies. The multiplexer channels at positive frequencies are thus "pure" VSB channels, in contrast to the QAM channels that exist at the roots of the trees. A related wideband VSB filter design ($G_{10}(jz)$ in Fig. 2) after the root of an identical tree can be used to pass all the multiplexer channels at negative frequencies and eliminate all those at positive frequencies. The outputs of the two wideband VSB filters can be multiplexed to obtain a covering of the positive and negative and frequencies with single channel multiplexer VSB communication channels. Adjacent channels overlap at the 3-dB down points, so the spectrum coverage is efficient.

By feeding signals into the multiplexer at different levels of the tree bandwidth on demand by factors of two is realized [8].

Another flexibility in design requirements we have explored is the tradeoff between number of levels of bandwidth on demand and the total number of taps along a multiplexer channel (or total delay). This has been done, as in Fig. 3, by using the outputs at the roots of identical filterbank trees as inputs to two related DFT polyphase synthesis filter banks, whose outputs are multiplexed, and implementing a receiver, as in Fig. 6, with the matched filtering realizing the filtering by the matching DFT polyphase analysis filterbank and then by the correct filters in the matching analysis filter bank tree, as described and discussed in [12, 13]. For example, with the outputs at the roots of 4 trees with 4 leaves as inputs into each of two closely related DFT polyphase synthesis filter banks (see Fig. 3), and a matching receiver (Fig. 6) simplified to receive one channel at a time, there are 32 multiplexer channels, of which 16 are shown in Fig. 5, and the other 16 fill the gaps. The delay and number of coefficients is about one quarter of those required if the outputs of two identical 16-input trees, each followed by a wideband VSB filter (one passing positive and the other negative frequencies), were multiplexed to obtain 32 multiplexer channels, but there are now only two levels of bandwidth on demand instead of four.

A study of the BER performance of the concatenation of filter bank trees and DFT polyphase filter banks with different phase offsets (e.g., Fig. 7), timing offsets (e.g., Fig.8) and combinations of both phase and timing offsets at the receiver has been documented in [12, 14], and indicated that the design of a synchronization scheme of reasonable complexity should be possible without introducing errors which cannot be corrected by currently available forward error correction schemes such as turbo codes and hypercodes.



Prototype transfer function for DFT polyphase synthesis filter bank above is $A(z)$

Prototype transfer function for 4-input DFT polyphase synthesis filter bank with output $\hat{X}_2(z)$ is $jz^{-D}A(1/z)$

Fig. 3: Arrangement of DFT polyphase synthesis filter banks.

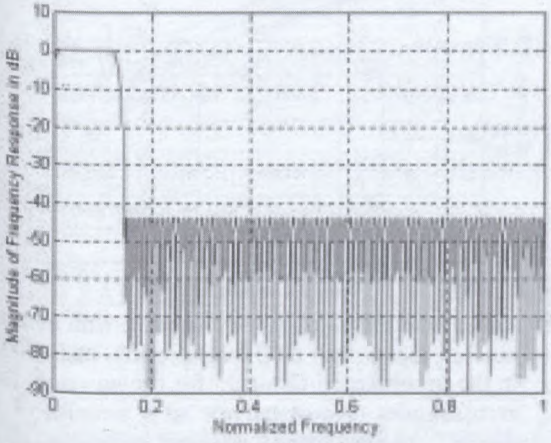


Fig. 4: The magnitude frequency response of the 8th-band equiripple stopband VSB filter used for calculating the polyphase components for the DFT polyphase filter banks in the example.

All the designs for filters described above resulted in perfect reconstruction at the output of the receiver. This results in the theoretically perfect orthogonality of the signals that are received at the output of the demultiplexer from each of the separate input signal streams into the multiplexer. Note that the filter designs used in the filter banks described and used in our publications cited above all had nonlinear phase. We have not been able to obtain a further simplification of the receiver structure in these cases.

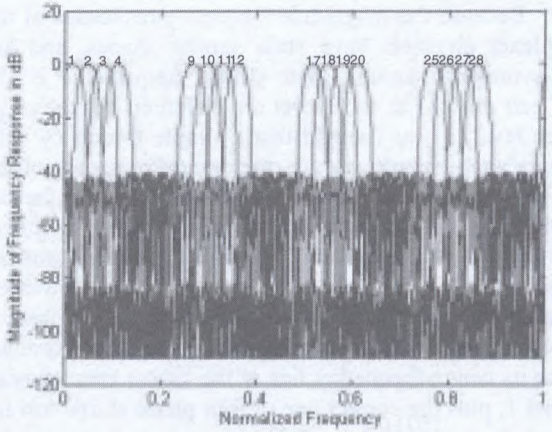
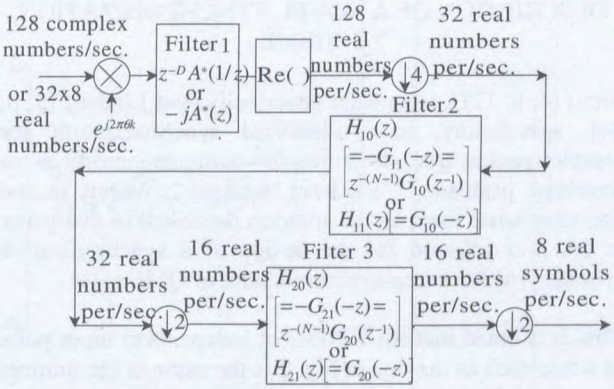


Fig.5: The composite magnitude frequency response $\hat{X}_1(z)$.



Note 1: The $G_{10}(z)$ and $G_{11}(z)$ are the lowpass and highpass transfer functions of the QMF pair next to the root of the synthesis filter bank trees, and the $G_{20}(z)$ and $G_{21}(z)$ are the transfer functions of the QMF pairs at the leaves of the trees.

Fig. 6: Block diagram of the receiver architecture, and relative symbol rates at various points.

Using the linear phase filter designs with equiripple stopband attenuation with a minimum of 40 dB, provided to us by W.F. McGee [15], resulted in negligible (for example, sidelobes of about 4% of the unit pulse at the output of the receiver, as response to a unit pulse at an input port at the leaf of the synthesis filter bank tree in the transmitter) intersymbol interference (ISI) due to non-perfect reconstruction, but provided other, overriding advantages, one of which was the further simplification of the receiver.

Fig. 9 shows an example that yields 8 multiplexer channels at positive frequencies and 8 multiplexer channels at negative frequencies, for a total of 16 multiplexer channels around the unit circle. (Fig. 9 also shows the proposed synchronization scheme at the transmitter, which will be succinctly described in the following section). Fig. 10 shows the matching receiver to the transmitter of Fig. 9.

The filters, in general, could be non-linear phase or linear phase. Because the magnitude frequency responses of the multiplexer channels have such similar shapes, and are even symmetric around their centre frequencies if the QMF pair designs at each level are different, but the same at each level, it was thought that a simple frequency shift to centre each channel at, say, the centre frequency of the first multiplexer channel would enable a further simplification of the receiver. However, in the nonlinear phase case it was not found possible to obtain a further simplification. In the linear phase case, however, assuming that the filtering by the receiver is matched to receive channel 1, the frequency shift of any other channel so that its centre frequency lies at the centre frequency of channel 1, plus the correct one of four phase shifts- $\pi/4$ for channels 2, 6, 10 and 14, $\pi/2$ for channels 3, 7, 11 and 15, $3\pi/4$ for channels 4, 8, 12 and 16, and 2π for channels 5, 9 and 13-will enable reception of the other channels using the same filtering as for channel 1.

DESCRIPTION OF A NOVEL SYNCHRONIZATION SCHEME

Jones [4, p. 121], somewhat generically, and Lindsey [5, p. 84], specifically, have identified synchronization for wavelet packet tree structures for communications as an unsolved problem. We have seen no solution in the literature since then, so the solution described in this paper is the first solution for the design of a synchronization scheme [16] for a tree-structured bank of QMF pairs.

First it is noted that the number of independent input ports in a tree such as the one in Fig. 9 is the same as the number of independent wavelet packet basis functions for such a tree of QMF pairs [10], and is given by the recursion

$$T_n = T_{n-1}^2 + 1.$$

The recursion begins with $n=1$, for which $T_{n-1}=0$. The rapid growth in number of possible independent input ports is itself intriguing in its possibilities for innovative communication applications. For the tree with 8 leaves and 3 levels shown in Fig. 9 there are 26 sets of independent input ports-25 if direct transmission into the root of the tree is considered to be a degenerate case.

The main idea of the synchronization scheme at the transmitter is to insert a sequence of 32 synchronization bits into the data stream after the root of the tree into a window as shown in Fig. 9 that is correctly located, as described in [16], and to effect this insertion in spite of the fact that the data inputs into input ports generally cause samples of responses to fall inside the window. This is accomplished by "zeroing out" the effects of the data by finding the real inputs at pre-calculated locations of synchronization words at each of the input ports in a set which will result in the negative of the values appearing in the window due to the data outside the locations of these input synchronization words at the input ports in each set.

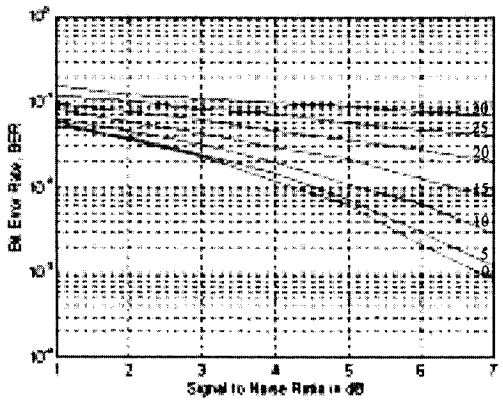


Fig. 7: Simulated BER vs. SNR in channel 6, with non-zero random inputs into all 32 input ports, and AWGN in the transmission channel, for phase errors from zero degrees (lowest curve), by 5-degree intervals, to 30 degrees (highest curve), in all 32 communication channels, in the receiver demodulator, for 25% rolloff in the multiplexer-demultiplexer channel magnitude frequency characteristics.

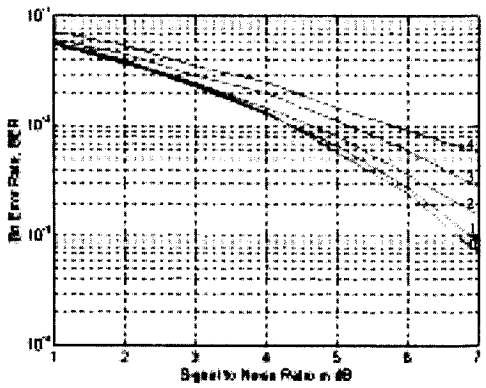


Fig. 8: Simulated BER vs. SNR for channel 6 with non-zero random inputs into all 32 input ports, and AWGN in the transmission channel, for timing errors from zero samples (lowest curve) to 4 samples of 16, which is (4/16)T baud, (highest curve), in all 32 communication channels, in the receiver demodulator, for the phase offset yielding the best BER vs. SNR plot, for 25% rolloff in the multiplexer-demultiplexer channel magnitude frequency characteristics.

The values appearing in the window due to unit pulses at the locations in the synchronization words are related by a transfer matrix, A, whose columns are the unit pulse responses in the window. This is just the path through all the high pass filters, which are maximum phase, in the nonlinear phase case. For the linear phase case all of the paths from each level of the tree have the same delay, so the locations of the synchronization words at input ports at the same level are all the same. For the nonlinear case,

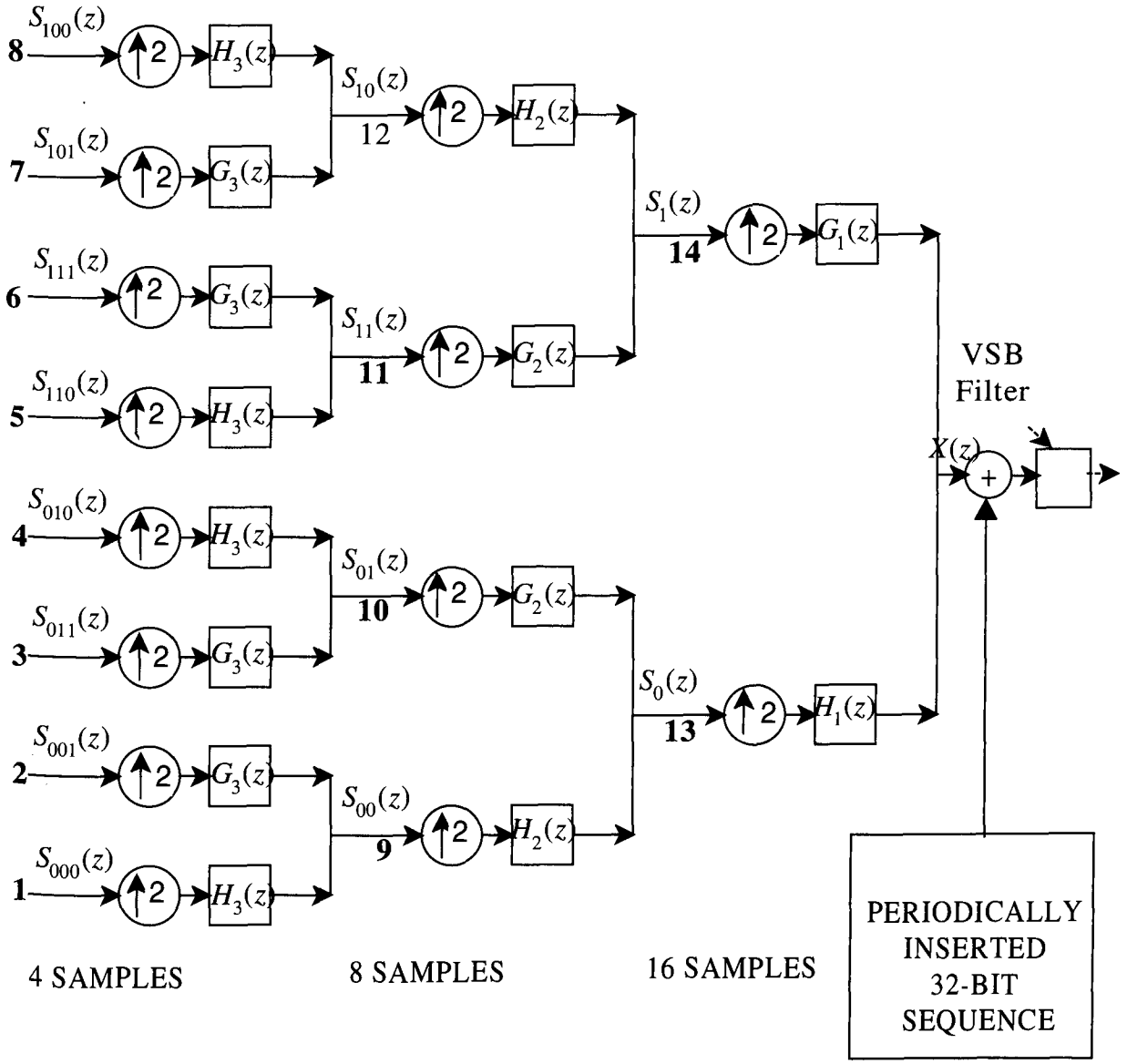


Fig. 9: Example tree with 8 leaves and different designs of the quadrature mirror filter pairs at each level used to illustrate the proposed synchronization scheme method and computations.

they are generally different, even at the same level. This is one attractive feature of the linear phase case for calculations and simulations. Another is that the window and input sync word locations can all be easily calculated analytically. As seen from Fig. 9, for an example with independent input ports 6, 5, 12 and 13 there are 4 successive inputs into each of ports 6 and 5, 8 successive inputs into port 12, and 16 inputs into port 13. Since the path from port 6 is all through highpass filters it is the longest delay path determining the window location in the nonlinear phase case. All paths from the leaves of the tree are the same lengths in the linear phase case. The input unit pulse in the first position of the sync word at port 6 results in a response which determines the location of the window to be such that the peak of that response and several of the largest peaks on each side of it must fall within the leftmost quarter of the window. This response is shifted to the right by 8 positions for each of the following successive 3 inputs at port 6, and the last one must, like the first one, have its maximum energy inside the window, which is achieved by as symmetrical a location of the four responses within the window as possible. The input sync word for port 6 is located in positions 1-4. The beginning of the sync words at ports 5, 12 and 13 must be delayed for the maximum energy to fall in the window. Responses to successive unit pulses at ports 12 and 13 are shifted by 4 and 2 samples, respectively, in the window. The 32x32 matrix A is formed from the 32 columns of unit pulse responses in the window due to unit pulse inputs in the sync words at the

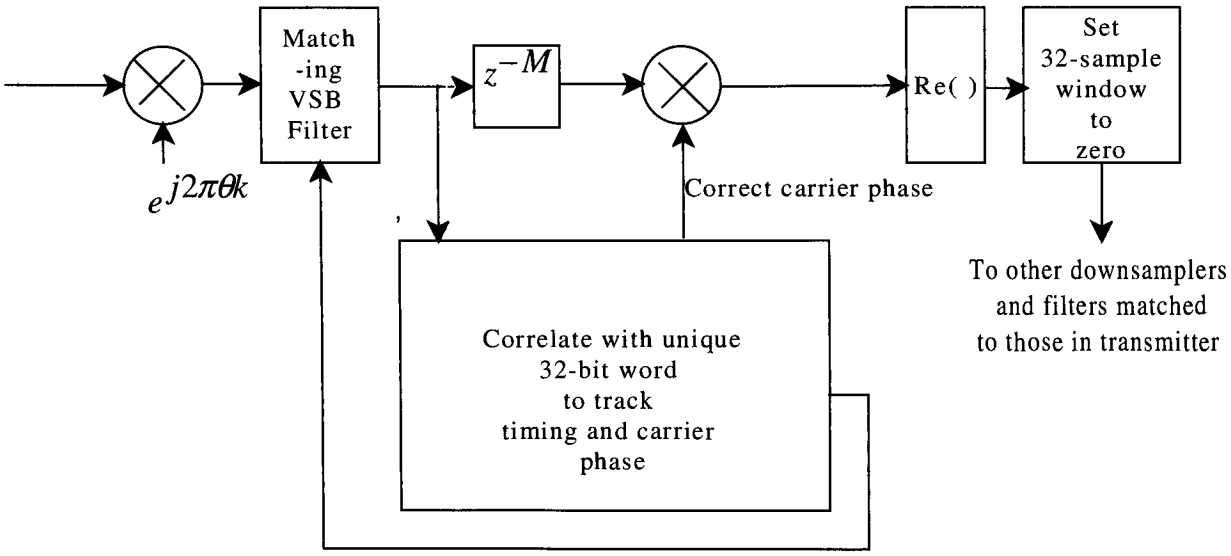


Fig. 10: Matched receiver to the transmitter of Fig. 1. The exponential multiplier at the input frequency shifts a selected channel to the centre frequency of a fixed baseband channel to simplify the receiver further in the linear phase case.

input ports in each set. This matrix has interesting properties, and the ratios of the maximum to minimum eigenvalues of $A\,A^T$ are very close to and greater than unity, showing that the A is well-conditioned, so we can use A^{-1} as shown in [21] to zero out the window, after which a 32-bit synchronization sequence is inserted into the window location. A dedicated channel communicates information such as the configuration of input ports to the receiver. Analysis, calculations and simulations for Rician fading channels verify that the receiver shown in Fig. 10 will work with the synchronization scheme shown in Fig. 9 at the transmitter. The effects of a variety of Doppler shifts have been evaluated for the linear phase case in [15].

FUTURE RESEARCH

This includes turbo code or hypercode implementation in the out-bound direction, further studies and improvements of the synchronization scheme, performance studies and possible redesign for a variety of channel models such as frequency selective fading, study of the backoff of the satellite amplifiers needed, design of a method for achieving this backoff, and and design of the upstream transmission and receiving schemes.

For the upstream direction the transmission is asynchronous so multi-user detection methods using soft-decision inputs and outputs in an iterative method as done in turbo decoding offer a promising new approach [17].

ACKNOWLEDGEMENTS

We are very grateful to Dr. W.F. McGee for many discussions and contributions to the theoretical, design and simulation studies.

REFERENCES

[1] M.V. Wickerhauser, Adapted wavelet analysis from theory to software, Wellesley, MA: A.K. Peters,1994.

[2] R.E. Learned, H. Krim, B. Claus, A.S. Willsky and W.C. Karl, Wavelet-packet-based multiple access communication; in Proceedings of SPIE Intl. Symposium on Optics, Imaging and Instrumentation-Mathematical Imaging: Wavelet Applications in Signal and Image Processing, No. 2303, San Diego, CA, 1994, pp. 246-259.

[3] J. Wu, K.M. Wong and Q. Jin, Multiplexing based on wavelet packets; in Proceedings of SPIE Intl. Symposium on AEROSENS, No. 2491, Orlando, Florida, 1995, pp. 315-326.

[4] W. W. Jones, A unified approach to orthogonally multiplexed communication using wavelet bases and digital filter banks, Ph.D. Dissertation, Faculty of the Russ College of Engineering and Technology, Ohio University, 1994.

[5] A. R. Lindsey, Generalized orthogonally multiplexed communication via wavelet packet basis, Ph.D. Dissertation, Faculty of the Russ College of Engineering and Technology, Ohio University, 1995.

[6] A. R. Lindsey, Wavelet packet modulation for orthogonally multiplexed communications, IEEE Trans. Signal Processing. vol. 45, pp. 1336-1339, May 1997.

[7] K. M. Wong, J. Wu, T. N. Davidson and Q. Jin, Wavelet packet division multiplexing and wavelet packet

design under timing error effects, IEEE Trans. Signal Processing, vol. 45, pp. 2877-2890, Dec. 1997.
[8] M. Sablatash, J.H. Lodge and W.F. McGee, Equivalence between vestigial sideband (VSB) and offset quadrature phase shift (OQPSK) modulations and relationships to wavelet packet-based multiplexing; in Proceedings of the 18th Biennial Symposium on Communications, Queen's University, Kingston, Ontario, Canada, 1996, pp. 339-342.
[9] M. Sablatash, J.H. Lodge and W.F. McGee, The design of filter banks with specified minimum stopband attenuation for wavelet packet-based multiple access communications; in Proceedings of the 18th Biennial Symposium on Communications, Queen's University, Kingston, Ontario, Canada, 1996, pp. 53-56.
[10] M. Sablatash, J. H. Lodge and C.J. Zarowski
Theory and design of communication systems based on scaling functions, wavelets, wavelet packets and filter banks; in Proceedings of Wireless '96, the 8th International Conference on Wireless Communications, vol.2, Coast Plaza Hotel, Calgary, Alberta, Canada, 1996, pp. 640-659.
[11] M. Sablatash, T. Cooklev and J. Lodge, Design and implementation of wavelet packet-based filter bank trees for multiple access communications; in Proceedings of the IEEE International Conference on Communications (ICC '97), Montreal, Quebec, Canada, vol. 1, 1997, pp. 176-180.
[12] M. Sablatash, W.F. McGee and J. Lodge, Designs of prototype filters for calculation of polyphase components of DFT filter banks and simulation studies to evaluate the bit error rate performance of a multiplexer-demultiplexer filter bank transmitter-receiver for out-bound transmission with phase and timing errors at the receiver; in Proceedings of Wireless '97, the 9th International Conference on Wireless Communications, vol. 1, Coast Plaza Hotel, Calgary, Alberta, Canada, 1997, pp. 222-241.
[13]. M. Sablatash and J. Lodge, Theory and design of spectrum-efficient bandwidth-on-demand multiplexer-demultiplexer pairs based on wavelet packet tree and polyphase filter banks; in Proceedings of the 1998 International Conference on Acoustics, Speech and Signal Processing (ICASSP '98), vol. 3, Seattle, Washington, USA, 1998, pp. 1797-1800.
[14] M. Sablatash, W.F. McGee and J. Lodge, The performance of a spectrum-efficient multicarrier transmission system with phase and timing offsets; in Proceedings of Wireless '98, The 10th International Conference on Wireless Communications, vol. 1, Coast Plaza Hotel, Calgary, Alberta, Canada, 1998, pp. 418-434.
[15] W.F. McGee, Study into synchronization scheme for improved spectrum efficiency for FDMA/TDMA transmission in mobile and mobile satellite environments, Final Rep. on CRC Contract U6800-9-0526, Jan. 16, 1999.
[16] M. Sablatash and J. Lodge, Design of a synchronization scheme for a bandwidth-on-demand multiplexer-demultiplexer pair based on wavelet packet tree filter banks; accepted for presentation and inclusion in the Proceedings of the 1999 International Conference on Acoustics, Speech and Signal Processing, ICASSP '99, vol. 4, Civic Plaza, Hyatt Regency, Phoenix, Arizona, 1999, pp. 2211-2214.
[17] M. Moher and P. Guinand, An iterative algorithm for asynchronous coded multiuser detection, IEEE Communications Letters, vol. 2, pp. 229-231, Aug. 1998.

Investigation of Satellite Diversity and Handover Strategies in Land Mobile Satellite Systems based on a Ray Tracing Propagation Model

Martin Döttling, Thomas Zwick, Werner Wiesbeck

University of Karlsruhe
Institut für Höchstfrequenztechnik und Elektronik (IHE)
D-76128 Karlsruhe, Germany
E-Mail: Martin.Doettling@etec.uni-karlsruhe.de

ABSTRACT

Land Mobile Satellite (LMS) Systems require fade countermeasure techniques since they suffer from limited link margins and severe channel degradations. In this paper, satellite diversity and satellite handover strategies are evaluated and compared by means of case studies in non-urban areas. A novel way of post-processing ray tracing results is outlined; quality criteria for handover and diversity are defined, calculated and compared. Focus is on typical scenarios and the correlation between parameters that influence system performance.

INTRODUCTION

The majority of the available LMS propagation models are derived from measurements and of statistical nature. Their aim is to support general system design (orbit type, link margin, etc.) and to provide estimates of channel statistics. However, due to their empirical background, it is virtually impossible to consider the complex interaction and correlation between different propagation factors, like azimuth correlation of shadowing, blockage, signaling delay, channel state, operational scenario, multipath effects, mobile and satellite motion and Doppler shift. Moreover, with the venue of third generation systems, the relevance of wideband effects has to be investigated and realistic time series of the signal-to-noise ratio (SNR) are required for efficient receiver design.

The new simulation tool, which is introduced in this paper, is designed to fill this gap and to complement conventional channel models. It inherently considers correlation of the above mentioned propagation effects using a ray tracing propagation model in conjunction with an orbit generator. The mobile surroundings are simulated by high resolution topographical and land use data, as well as by stochastically generated roadside objects [1]. At each time step, all relevant signal paths are calculated for each satellite, including amplitude, polarization, delay time, Doppler frequency and direction of arrival. Hence, wideband time series of power delay profiles are available. Comparisons with LMS measurements show the validity and capability of this approach [2, 3].

The propagation data is used to investigate different handover, diversity and combining schemes for the satellite-to-mobile link. First, time series of bit energy per noise spectral density (Eb/N0) are calculated from the power delay profiles, including the influence of system noise, intersymbol interference, multiple access interference and imperfect power control due to signaling delay. Next, probability density functions (PDF) and cumulative distribution functions (CDF) of Eb/N0 are calculated and compared. Selection, equal gain and maximum ratio combining are considered for two- and three-branch diversity.

Based on the channel time series, the effect of different handover initialization thresholds is evaluated. Quality measures, like handover rate, handover delay and outage time, are calculated. A comparison between diversity and handover is performed. Results are given for different satellite constellations, like Iridium, Globalstar and ICO, and different mobile operational scenarios (pedestrian and vehicular).

THE CHANNEL MODEL

The surroundings of the mobile terminal (MT) are characterized by topographical and land use data containing geographical height and a classification of the surface properties, respectively. As an example, Fig. 1 depicts an area of 20km x 20km in the Rhine Valley, southern Germany. The z-axis is topographical height, whereas the land use classes are shown in different colors. The landscape shows a typical non-urban mixture of terrain and land use elements. The mobile trajectory follows the German highway no. 8 westbound and no. 5 southbound, respectively. The 50m-grid is sufficient to resolve all relevant land use and terrain features. However, further

input data is necessary to compensate for the missing roadside objects, which influence propagation notably [4]. Therefore roadside trees and buildings are generated stochastically, with varying statistics for density, height and location according to the land use class of the mobile position [3].

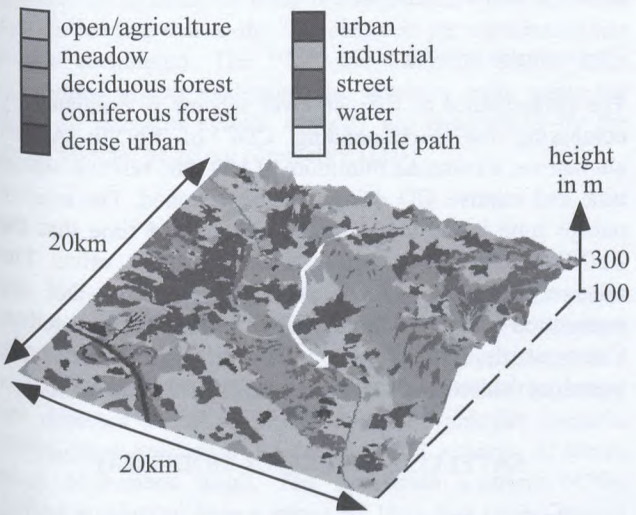


Fig. 1: Topography, land use and mobile trajectory in the Rhine Valley, southern Germany

An orbit generator provides the satellites' positions for each time step of the simulation. Topography, land use, mobile trajectory, roadside obstacles and orbit data are fed into the ray tracer. The ray tracing algorithm is divided into a 2D Vertical Plane Model and a 3D Scattering Model. The Vertical Plane Model contains the most important signal paths in a vertical plane that embodies satellite and mobile. Free space propagation, losses due to diffraction or transmission through vegetation are considered as well as ground reflection and backreflection. Single scattering contributions are evaluated in the 3D Scattering Model [2].

For each satellite and every time step t the complex polarimetric transmission matrix $T_i(t)$, delay time $\tau_i(t)$ and direction of arrival (ϑ_i, ψ_i) are calculated for all $N(t)$ relevant rays. After weighting each ray with the antenna pattern C_R of the MT the complex transmission factor $A_i(t)$ is obtained:

$$A_i(t) = C_R^T(\vartheta_i, \psi_i) \cdot T_i(t), \tag{1}$$

where the superscript T denotes the transpose operation. The superposition of all contributions yields the power delay profile, which is equivalent to the satellite's time-variant channel transfer function

$$h_{ch}(\tau, t) = \sum_{i=1}^{N(t)} A_i(t)\, \delta(\tau - \tau_i(t), t). \tag{2}$$

CALCULATION OF EFFECTIVE SNR TIME SERIES

The time series of power delay profiles are transformed into instantaneous effective SNR values by

$$SNR(t) = \frac{S(t)}{E\{N_{sys}\} + N_{isi}(t) + E\{N_{mai}\}}, \tag{3}$$

where $S(t)$ denotes the signal, N_{sys} the system noise, $N_{isi}(t)$ the equivalent noise due to intersymbol interference and N_{mai} the noise due to multiple access interference in code division multiple access (CDMA) systems. $E\{x\}$ denotes the expectation of variable x. All noise contributions are modeled as white Gaussian noise.

The received signal in the n-th symbol period T_s can be expressed as

$$\begin{aligned} y_R(\tau, t) &= \sum_{l=-\infty}^{\infty} a_l h_{tot}(\lfloor n - l \rfloor T_S t) \\ &= a_n h_{tot}(\Omega t) + \sum_{\substack{l=-\infty \\ l \neq n}}^{\infty} a_l h_{tot}(\lfloor n - l \rfloor T_S t), \end{aligned} \tag{4}$$

where the first term represents the contribution of the wanted symbol a_n and the second term the influence of all other symbols [5]. The total transfer function h_{tot} is given by

$$h_{tot}(\tau, t) = h_{sys}(\tau) * h_{ch}(\tau, t), \tag{5}$$

where the channel transfer function h_{ch} is computed by the ray tracer. In the sequel, raised-cosine filtering is assumed for the system transfer function h_{sys}. By evaluation of (5), (4) and (2) the wanted signal power yields:

$$S(t) = \left(\frac{\lambda_0}{4\pi}\right)^2 P_T G_T P_R \left| \sum_{i=1}^{N(t)} A_i(t) \cdot h_{sys}(-\tau_i(t), t) \right|^2 \tag{6}$$

and the corresponding intersymbol interference power is

$$N_{isi}(t) = \left(\frac{\lambda_0}{4\pi}\right)^2 P_T G_T P_R \cdot \left| \sum_{\substack{l=-\infty \\ l \neq n}}^{\infty} \sum_{i=1}^{N(t)} A_i(t)\, h_{sys}(\lfloor n - l \rfloor T_S - \tau_i(t), t) \right|^2. \tag{7}$$

Although N_{isi} is negligible for second generation systems, it may influence future wideband LMS systems.

For CDMA systems multiple access interference is the limiting factor for SNR. For the downlink the MAI is a function of the received power $P_s = S_s + N_{isi,s}$ from each visible satellite s and the number of equivalent channels C_{eq}:

$$E\{N_{mai}(t)\} = \sum_{s=s_0}^{S(t)} P_s(t) \cdot E\{C_{eq}(t)\}. \quad (8)$$

P_s is calculated by the propagation model for all $S(t)$ satellites. For synchronous orthogonal CDMA $s_0 = 2$ (i. e. no MAI from the serving satellite), else $s_0 = 1$. C_{eq} depends on the number of co-channel beams, beam overlap, channel utilization, voice activity and antenna discrimination. Comprehensive derivations of $E\{C_{eq}\}$ can be found in [6, 7].

Signaling Delay

For LMS systems propagation time has to be considered. Especially the efficiency of system control commands is impaired by the signaling delay, which can be approximated by

$$\Delta t_s = n_{st} \cdot t_{st}(h, \varepsilon_{gw}, \varepsilon_{mt}) + t_{pr}, \quad (9)$$

where n_{st} is the number of single trips between gateway and MT, t_{st} the corresponding propagation time and t_{pr} an additional delay due to signal processing in the control instances. The impact of the slant range on t_{st} is calculated based on orbit heigth h and the satellite's elevations as seen from the gateway (ε_{gw}) and from the mobile terminal (ε_{mt}).

Power Control (PC)

The system's power control is simulated by a user-defined number of power control steps, the corresponding thresholds and power correction steps. Additionally, the target SNR value, the maximum and minimum transmitted power, as well as the power control update rate are specified by the user. The power-controlled received power P_s is calculated by:

$$P_s(t) = \gamma(t - \Delta t_s) \cdot P_{s,nom}(t), \quad (10)$$

where γ is the power control factor and $P_{s,nom}$ the received power for nominal transmitted power.

HANDOVER MODELING

This study considers the feasibility and benefit of satellite handover based on SNR measurements, since this type of handover requires the highest signaling effort and charges the system resources notably. To mitigate shadowing, a fast handover scheme is necessary. It is based on a constant monitoring of the SNR values of all satellites. If the SNR of the active satellite falls below a certain margin ΔSNR_{HO} with respect to the best satellite, a HO is initiated. To prevent system congestion with signaling, a maximum HO rate can be specified. Based on these parameters, a time series of Eb/N0 is calculated, which includes the effect of handover and signaling delay Δt_s.

Performance criteria

The performance of the handover scheme is evaluated by comparing the corresponding CDF of Eb/N0. As an alternative, a common minimum of HO rate, relative outage time and relative HO delay may be searched. The relative outage time is defined as the percentage of time that the Eb/N0 time series is below a certain target value. The relative HO delay is the percentage of time that the connection is not provided by the best available satellite. Consequently, the relative HO delay includes both, signaling delay and delay due to the hysteresis ΔSNR_{HO}.

SATELLITE DIVERSITY MODELING

Another way to increase system availability is satellite diversity. The SNR values of all satellites are permanently monitored. For m-branch satellite diversity, a connection with maximum m satellites is established and power-controlled. The actual number of satellites in the active set is determined by ΔSNR_{add} and ΔSNR_{drop}. If the SNR level of an inactive satellite is less than ΔSNR_{add} below the best satellite a request for adding this satellite to the active set is issued. In a similar way, a request for dropping a satellite from the active set is transmitted if its SNR value is more than ΔSNR_{drop} below the best satellite. Both actions occurr with signaling delay.

Combining

The superposition of the active satellites' signals is performed in the combiner. Each branch i is weighted with factor b_i. For selection combining (SC) $b_i = 1$ for the best satellite and $b_i = 0$ else. The equal gain combiner (EGC) uses the same factors in each branch, while the weights for maximum ration combining (MRC) are:

$$b_i(t) = \sqrt{\frac{S_i(t)}{N_i(t)}}. \quad (11)$$

Note that the noise level of the satellites can be different. For coherent detection, the resulting signal power $S_c(t)$ and noise power $N_c(t)$ for $L(t)$ active satellites yields:

$$S_c(t) = \left| \sum_{i=1}^{L(t)} b_i(t) \sqrt{\frac{S_i(t)}{N_i(t)}} \right|^2, \quad (12)$$

$$N_c(t) = \sum_{i=1}^{L(t)} b_i^2(t) \cdot N_i(t) \qquad (13)$$

Performance criteria

In diversity systems the trade-off between diversity gain and system loading due to the $L(t)$ channels per connection has to be considered. The PDF and CDF of Eb/N0 after combining allows evaluating the diversity gain, while the costs in terms of system capacity are determined by the mean number of active channels $E\{L(t)\}$.

RESULTS AND DISCUSSION

The interaction between mobile velocity, imperfect power control, signaling delay, different HO and diversity schemes are investigated by means of exemplary simulations in a non-urban area. Topography, land use and mobile trajectory are depicted in Fig. 1. The high-speed vehicular scenario (190s, 50ms resolution) uses a mean MT velocity of 60m/s (e. g. high-speed train). The pedestrian scenario (470s, 100ms resolution) uses a speed of 1m/s and covers only a part of the mobile path of Fig. 1. The signaling delay parameters are set to $n_{sl} = 4$ and $t_{pr} = 50$ms. Table 1 lists the start, climax and end elevation of all visible satellites for the LMS systems under consideration. Since not all required system parameters for Iridium, ICO and Globalstar are available in open literature the absolute values of the results are subject to change. Therefore the following comparisons should only be regarded as typical for the type of system these implementations represent.

Table 1: Start, climax and end elevation of satellites

System	pedestrian scenario elevation in °	high-speed scenario elevation in °
Iridium	<8 – 40 – 33	28 – 46 – 40
	19 – 23 – <8	15 – 15 – <8
	22 – 22 – <8	
ICO	40 – 40 – 32	64 – 66 – 66
	28 – 36 – 36	10 – 12 – 12
Globalstar	<10 – 60 – 60	44 – 44 – 24
	52 – 54 – <10	18 – 40 – 40
	33 – 33 – 10	28 – 28 – 10
	<10 – 25 –25	<10 – 15 – 15

The influence of mobile speed on power control efficiency is illustrated by comparing the time series of Eb/N0 of both scenarios for the Globalstar simulation. The resulting Eb/N0 is calculated assuming three-branch satellite diversity with MRC. A 3-bit power control is used with parameters as in [6]. The dynamic range is ±10dB and the target SNR corresponds to Eb/N0 = 7.5dB after combining.

Fig. 2 shows the PDF of Eb/N0 with and without power control for both scenarios. It is evident, that for low earth orbit (LEO) systems slow fading and light shadowing at low mobile speeds can be controlled, whereas the power control becomes more and more ineffective as the mobile speed (and as a result the fading bandwidth) increases. Similar observations are made in [6].

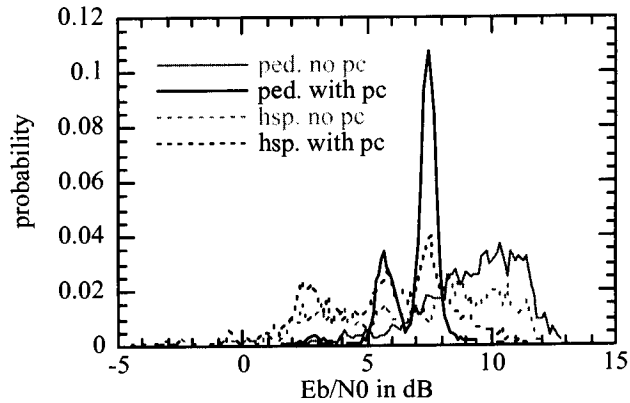


Fig. 2: PDF of Eb/N0 for Globalstar, with and without power control (pc) for high-speed (hsp) and pedestrian (ped) scenario

The impact of signaling delay and HO initialization criteria on system performance is investigated by comparing a medium earth orbit (MEO) system like ICO to a LEO system (Iridium). For the pedestrian scenario the HO performance criteria are displayed as a function of ΔSNR_{HO}. For ICO the minimum of the relative HO delay occurs at $\Delta SNR_{HO} = 5$dB (Fig. 3).

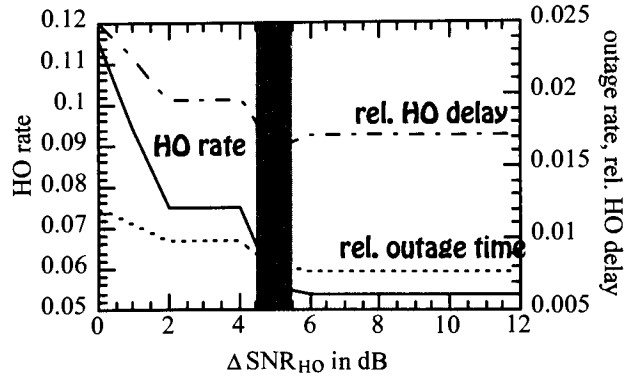


Fig. 3: HO performance for ICO, pedestrian scenario

The relative outage time and HO delay decreases with increasing ΔSNR_{HO} due to the considerable signaling delay for MEO systems: for a small hysteresis ΔSNR_{HO} many HO lead from the short-time shadowed "best" satellite to a non-optimal one. At the completion time of the HO, the channel states of the satellites may already have changed, so that the HO leads to deterioration. Thus, ΔSNR_{HO} based HO initialization seems not to be appropriate for MEO systems.

Sophisticated HO schemes should include e. g. satellite elevation or a history of SNR values to prevent such erroneous HO and be channel adaptive [8]. Fig. 4 shows the results for Iridium. Due to the smaller signaling delay, less erroneous HO occur, and the relative HO delay and outage time increase with increasing ΔSNR_{HO}. The optimum value of ΔSNR_H is 3dB. However, it is observed that for both systems the differences in HO delay and outage time are not significant.

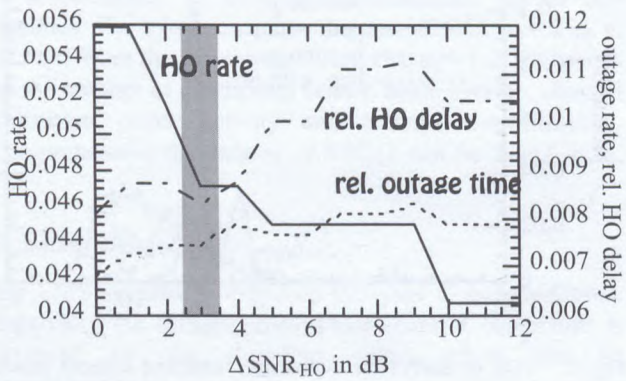


Fig. 4: HO performance for Iridium, pedestrian scenario

The following comparisons investigate, whether satellite diversity yields considerable performance gain with respect to a fast satellite handover approach. Fig. 5 and Fig. 6 oppose a HO system approach (m=1) to two- and three-branch satellite diversity with MRC.

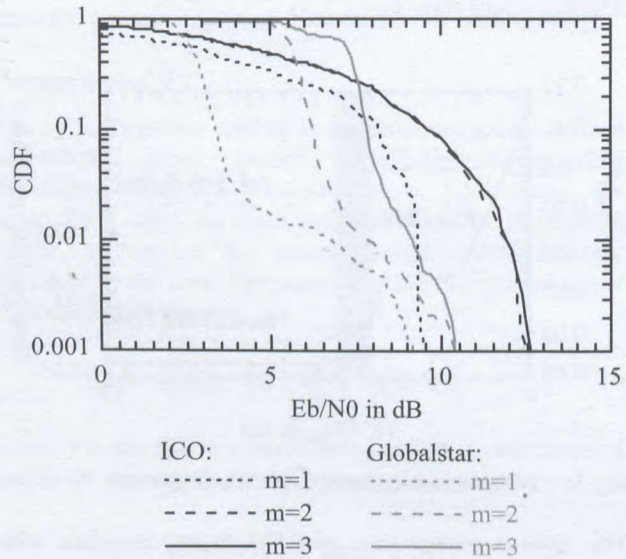


Fig. 5: Comparison of HO and diversity schemes, pedestrian scenario, MRC

In the pedestrian scenario (Fig. 5) both systems benefit from satellite diversity. However, at low Eb/N0 values the effect is more distinct for Globalstar, while the diversity gain for ICO occurs at higher Eb/N0 values. For a target Eb/N0 value of 5dB, the improvement is 92% ($m = 2$) and 96% ($m = 3$) for Globalstar, while ICO shows an increase of 14% in both cases. Two-branch and three-branch diversity produce identical results for ICO, while there is an additional gain for Globalstar at Eb/N0 > 6dB. For Globalstar, the diversity gain is 3dB ($m = 2$) and 3.4dB ($m = 3$) at the Eb/N0 value that is exceeded by 90%. These values are slightly smaller than results reported in urban environment [9]. Note that [9] does not consider imperfect power control and signaling delay. Fig. 6 shows the same comparison for the high-speed scenario. While the diversity gain at Eb/N0 = 5dB has practically disappeared for ICO (5%), Globalstar still benefits from diversity gain (32% for $m = 2$ and 40% for $m = 3$).

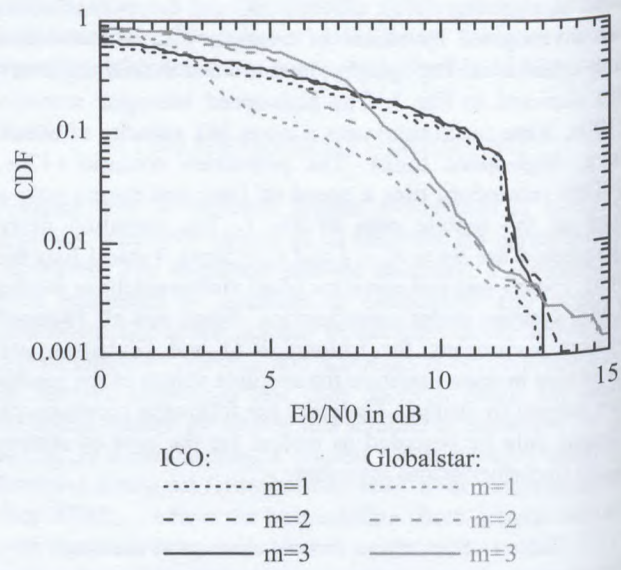


Fig. 6: Comparison of HO and diversity schemes, high-speed scenario, MRC

Different combining schemes for the Globalstar simulation with $m = 2$ are compared in Fig. 7. For both scenarios, the EGC and MRC perform nearly equal since the SNR values of the active satellites are comparable. MRC and EGC are considerably better than SC: an increase of 82% is observed for the pedestrian scenario at Eb/N0 = 5dB. For the high-speed application the increase is still 31%.

The influence of different margins ΔSNR_{drop} and ΔSNR_{add} on the Globalstar system performance ($m = 2$; MRC) is depicted in Fig. 8. Both values are set to 3dB, 6dB, 9dB and 12dB, respectively. For the high-speed scenario, no significant gain is achieved for Eb/N0 < 10dB. The mean number of active channel rises from 1.46 to 1.75. An interesting effect is visible in the pedestrian scenario: the smallest hysteresis (3dB) with lowest channel occupancy ($E\{L(t)\} = 1.90$) performs best. This can be explained by the active set updating algorithm, which aims at minimizing

of them falls short of ΔSNR_{drop}, even if a new satellite would be the better choice. As a result, large hysteresis leads to a deadlock in a non-optimal active set. Low ΔSNR_{drop} values cause more frequent changes of satellites but yield 10% better performance at Eb/N0 = 7dB.

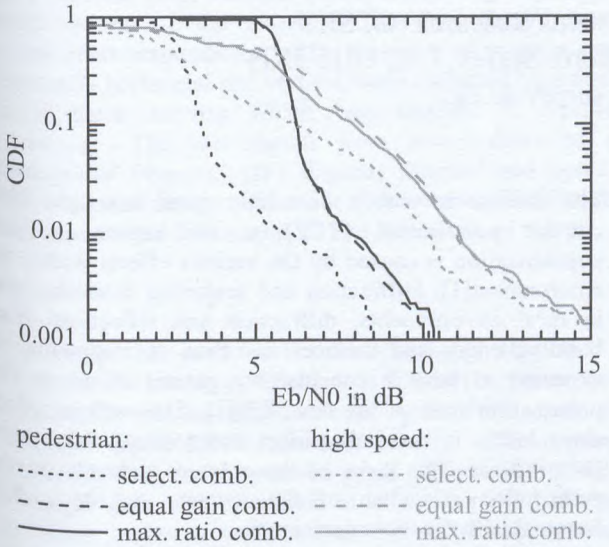


Fig. 7: Comparison of different combining schemes, Globalstar, m=2

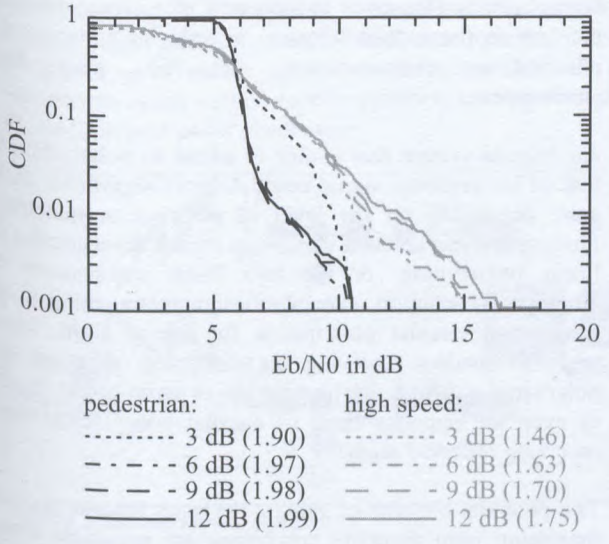


Fig. 8: Comparison of different satellite diversity schemes, Globalstar, m=2, MRC, in brackets: mean number of active channels

CONCLUSION

This paper outlines a new simulation tool to investigate satellite diversity and satellite handover based on a ray optical LMS propagation model. Due to the physical and deterministic basis, the software includes the complex interaction and correlation of several factors that play a major role in LMS system performance, like correlation of shadowing, satellite elevation, blockage, signaling delay, mobile speed, imperfect power control and system parameters for HO and diversity. Thus the new approach is seen as a necessary complement to existing statistical models. It provides important in-sight for system parameter definition and allows to compare different fade countermeasure techniques. First exemplary results are shown and discussed to stress the capability of the new approach. Further simulations are intended to provide a broad basis for general conclusions in system design issues.

REFERENCES

[1] M. Döttling, A. Jahn, H. Ernst, S. Buonomo, Land Mobile Satellite Propagation Channel - A Combined Deterministic and Statistical Modelling Approach, *Proc. Fourth European Conf. on Satellite Communications ECSC-4,* 1997, pp. 182-187.

[2] M. Döttling, A. Jahn, J. Kunisch, S. Buonomo, A Versatile Channel Simulator for Land Mobile Satellite Applications, *Proc. IEEE Vehicular Technology Conf. VTC'98*, 1998, pp. 213-217.

[3] IMST, DLR, IHE, Land Mobile Satellite Propagation Model for Non-Urban Areas, *Final Report*, European Space Agency Contract No. AO/1-3101/96NL/NB, March, 1998.

[4] J. Goldhirsh, W. Vogel, Propagation Effects for Land Mobile Satellite Systems: overview, experimental and modeling results, *NASA Ref. Publ. 1274*, 1992.

[5] M. J. Miller, B. Vucetic, L. Berry, *Satellite Communications – Mobile and Fixed Services,* Kluwer Academic Publishers, 1993.

[6] R. De Gaudenzi, F. Giannetti, DS-CDMA Satellite Diversity Reception for Personal Satellite Communication: Satellite-to-Mobile Link Performance Analysis, *IEEE Trans. on Vehicular Technology*, vol. 47, no. 2, 1998, pp. 658-672.

[7] P. Monsen, Multiple-Access Capacity in Mobile User Satellite Systems, *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 2, 1995, pp. 222-231.

[8] H. Bischl, M. Werner, Channel Adaptive Satellite Diversity for Non-Geostationary Mobile Satellite Systems, *Proc. Fifth Int. Mobile Satellite Conf. IMSC'75*, 1997, pp. 25-31.

[9] R. Akturan, W. J. Vogel, Path Diversity for LEO Satellite-PCS in the Urban Environment, *IEEE Trans. on Vehicular Technology*, vol. 45, no. 7, 1997, pp. 1107-1116.

Measurement of the Polarisation State of Satellite to Mobile Signals in Scattering Environments

S.M.Leach, A.A.Agius, S.R.Saunders
Centre for Communication Systems Research (CCSR)
CCSR, University of Surrey, Guildford, Surrey, UK, GU2 5XH
Email: s.leach@ee.surrey.ac.uk

ABSTRACT

This paper details a measurement campaign which was undertaken to investigate the effects of buildings and trees on signal depolarisation within a personal satellite communications environment. The satellite environment was simulated using a circular polarised transmit antenna elevated above the buildings, with a dual linear polarised receive antenna with separate receivers for the vertical and horizontal components of the signal. The received signal was sampled and digitally stored for processing at a later stage. Data was collected at a number of different sites, such that the effects of diffraction and reflection from both buildings and trees were observed, along with effects within urban corridor situations.

The presented results show a clear indication that there is considerable potential diversity gain available in areas where the signal is greatly diffracted or where the main signal is a reflected one. Diversity gains of around 10dB have been observed and are presented here. There is however an adverse effect of up to 3dB due to the noise from the second receiver when comparing the signal to noise ratios from maximum ratio diversity combining with a standard circular polarised antenna in areas where the two signals are not decorrelated, such as line of sight situations or areas where there is little shadowing or multipath effect..

A pre detection weighting system could be a potential solution to the performance degradation in line of sight situations by applying complex weights to the two signals and summing them prior to the receiver rather applying them after the receiver chain at baseband. There are however complexity penalties which may make this solution not practical or realisable as training sequences which are too long and time consuming may be required to optimise the output signal to noise ratio.

INTRODUCTION

The newly proposed and deployed Satellite Personal Communication Networks (SPCN) will offer seamless communication of high quality to handheld terminal users around the globe.

The satellite to mobile downlink signal has right hand circular polarisation (RCP) at the source, a little depolarisation is caused by the various effects within the environment[1]. Diffraction and scattering from the trees in rural environments, diffraction and reflections from building edges and surfaces in urban environments are expected to have a considerably greater effect on the polarisation state of the downlink[2]. This will introduce extra losses in the link budget deteriorating the system performance. The level of these losses depends on the polarisation mismatch of the antenna on the mobile terminal with the incoming signal.

The aim of the measurement campaign reported in this contribution, was to measure and then model these depolarisation phenomena. This information can then be used to either recalculate the downlink budget including the polarisation losses or to form new requirements for the antenna on the mobile terminal in order to adjust to the polarisation characteristics of the propagation environment.

An antenna system that is able to adjust its polarisation to that of the received signal could achieve significant extra gain, depending on the level of polarisation matching. Furthermore, an antenna which can switch from circular to linear polarisation, or the two linear components of elliptical polarisation, can establish communication in all situations : circular polarisation for line of sight, linear when in shadow receiving the diffracted (elliptical co-polar) or the reflected (elliptical co- or cross-polar) signal, or even the opposite hand of circular polarisation when receiving reflected signals.

The potential benefits of using a diversity scheme such as maximum ratio diversity combining are presented along with a possible disadvantages in terms of received signal to noise ratio in line of sight environments. A potential solution to these disadvantages is also presented.

THE DEPOLARISATION MEASUREMENT CAMPAIGN

The aim of the depolarisation measurement campaign was to measure and record the different polarisation

components of the received signal at a number of locations of various types in a satellite communications environment.

A transmitting antenna with right hand circular polarisation (RCP) was placed on a stationary elevated position illuminating a number of different types of environment, trees, building surfaces, building edges and combinations of the above. The two linear wave components perpendicular to the direction of propagation, nominally horizontal and vertical, were collected by a dual linear patch antenna which was attached to mobile receivers. The two signals were mixed down to a manageable frequency (IF), digitally sampled and stored. These measurements were made at a number of different locations around the campus of The University of Surrey.

Equipment

The equipment used in the measurements (Figure 1) consisted of a carrier wave source at 2.385GHz, which was transmitted via an RCP antenna with 7dBi gain at boresight and a 3dB beamwidth of 80°. The receive antenna, a dual linear polarised patch antenna with 13dBi boresight gain was orientated at a number of different elevation angles towards either the diffracting or the reflecting object. The two received signals were fed into custom built RF receivers which filtered and mixed the frequency down to an IF of 8kHz. These IFs were sampled simultaneously using parallel 16bit analogue to digital converters at a rate of 32kS/s and the resultant samples stored on a computer. As well as the received signals, trigger pulses from a 5th wheel were recorded. The 5th wheel gave pulses approximately every 10th of a wavelength as the mobile receive platform moved through the environment under observation.

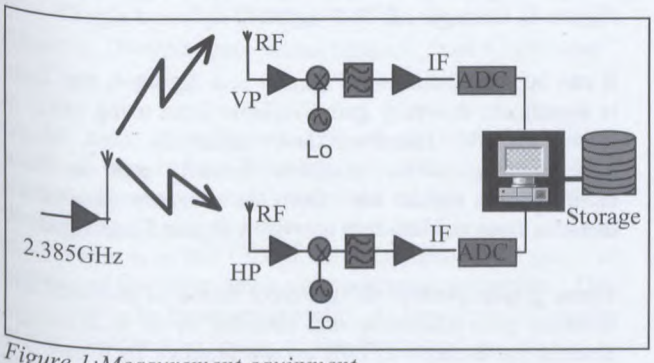


Figure 1:Measurement equipment

Subsequent off line processing was undertaken which digitally mixed the stored data down to a complex baseband signal, filtered and decimated it to a smaller and more manageable 1kS/s time based data. The stored trigger pulses from the 5th wheel were used to sample the data so that position based results which were independent of mobile speed could be obtained in addition to the time based ones. This has the advantage of removing any effect caused by the variation in speed of the receiving platform, giving further clarification during the analysis of the results.

Measurements

Measurements were made in a number of different locations which fulfilled the requirements of the campaign objectives, in terms of the position of buildings and trees, such that the measured signal variations could be attributed solely to the one diffracting or one reflecting object.

Figure 2 shows the measurement setup. The mobile receiver station, consisting of the multiple receivers, antennas, common local oscillator source, data acquisition and storage system and the 5th wheel trigger was moved along a path parallel with the diffracting or reflecting building or line of trees. The path lengths ranged from a few tens of metres up to one hundred metres depending on the surrounding area. Measurements were made at various distances from the building with the elevation of the patch set to a number of different angles.

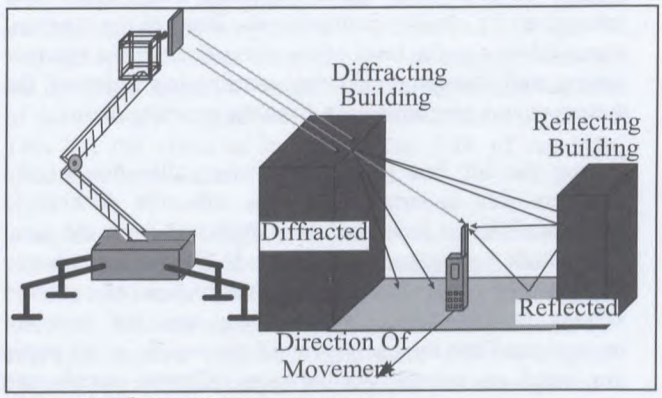


Figure 2: The measurement system

RESULTS

Analysis Methods

The results presented in this section are primarily based on comparison between cumulative distribution functions (cdfs) of the probability of receiving a signal to noise ratio less than the abscissa in the different environments and scenarios.

There are four different scenarios in which the data is analysed, the two independent and orthogonal polarised signals from the patch antenna, termed horizontal and nominally vertical are taken individually. These two polarised signals are phased with a 90° phase difference to simulate the received signal from a circular polarised antenna. The remaining two scenarios are both diversity techniques, the first is selection combining and the second, maximum ratio combining (MRC), both of which are taken calculated as a post detection system[4].

It is important to note that the results presented are a simulation for a circular polarised receive antenna and not for a post detection weighted system optimised for CP. If this were the case then the results for CP would be 3dB lower due to the extra noise from the second receiver thus showing an improvement in available diversity gain. All of the diversity gains quoted in this paper are based on the 1% probability that the signal is less than the abscissa.

Selection combining is chosen as the optimum condition for switch combining as there is prior knowledge of the signals in this case. MRC is the best diversity scheme for obtaining maximum signal to noise ratio so this indicates the absolute best available diversity gain from the two signals when employing a post detection weighting system. It is theoretically possible to obtain up to an extra 3dB of diversity gain when using a pre detection weighting system.

It is assumed in all the analysis, that the received voltages are independent of noise due to the fact that the received signals have a large signal to noise ratio. This was determined by closely examining the level of the received signal relative to the level of the noise floor of the receiver system and assuming that the contributing noise of the system comes predominantly from the receiver chains.

During the off line processing some calibration of the receivers was undertaken and the effective differences between different receivers were extracted from the data, this include both phase and amplitude differences between the receiver chains. A through calibration of received voltage to expected received power was not however incorporated into the analysis as all the results in this paper are based on comparison between different signals and diversity techniques so absolute received power and consequently signal to noise ratios are not necessary. No account for free space loss has been taken into account therefore different locations should not be directly compared in terms of power levels, thus indications of power on graphs within this paper should only be taken comparatively and not as an absolute received power from the system.

Results

Figure 3 and Figure 4 show example cdfs for received diffracted and reflected signals from one particular building, however they are indicative of findings at other locations. In the case of both diffraction and reflection, it can be seen that the result for a CP receive antenna is considerably inferior to that of the MRC case. In these regions, it can be noted that a CP antenna would perform worse than a simple vertical polarised antenna by a couple of dBs, the reason for this is that the horizontal component of the signal is attenuated significantly more by the process of diffraction and reflection. Thus the signal is no longer circular polarised, in fact it is considerably elliptical in polarisation and there is a polarisation mismatch.

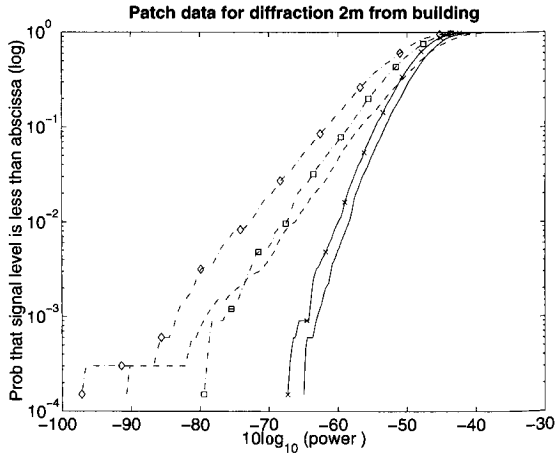


Figure 3: Example cdf in a region of diffraction

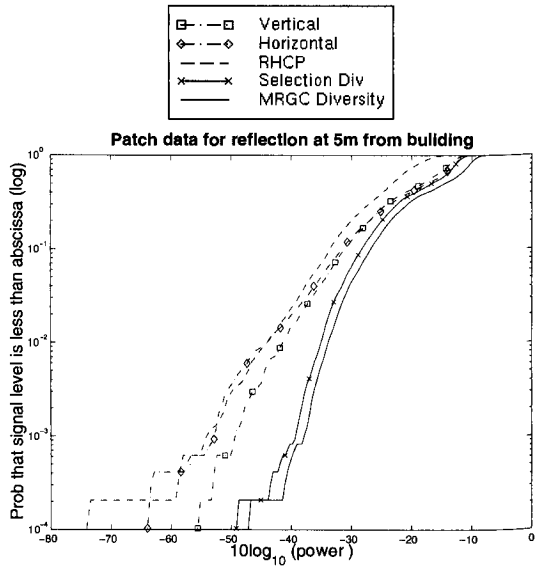


Figure 4: Example cdf in a region of reflected signals

It can be seen from both Figure 3 and Figure 4, that there is significant diversity gain available from using MRC at these specific locations and distances. A better understanding of the available diversity gain in these environments can be seen from the diversity gain versus distance from the building graphs in Figure 5 and Figure 6.

These graphs present the different forms of diversity and diversity gain relative to that obtained by an RCP receive antenna (referenced to 0dB on the graphs). In Figure 5 it can be seen that MRC does not produce diversity gain over an RCP antenna at all distances from the building. At around 14 metres a line of sight between the transmitting antenna and the receiving patch antenna was established. It is seen that diversity gain is obtained at distances less than 11m from the building. It is however within these regions of deep shadow that the signal is most attenuated by diffraction and that the diversity gain is beneficial in maintaining the link budget.

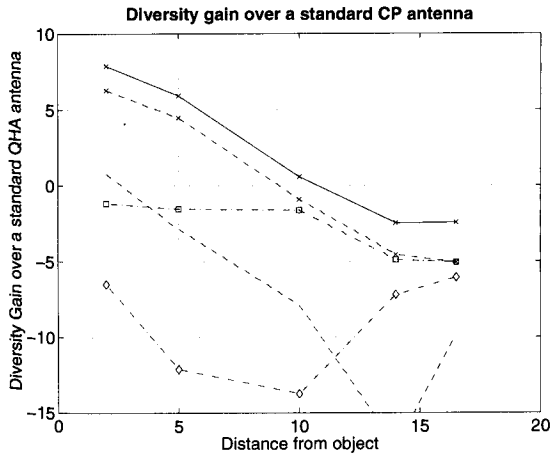


Figure 5: Diversity gain versus distance from a diffracting building

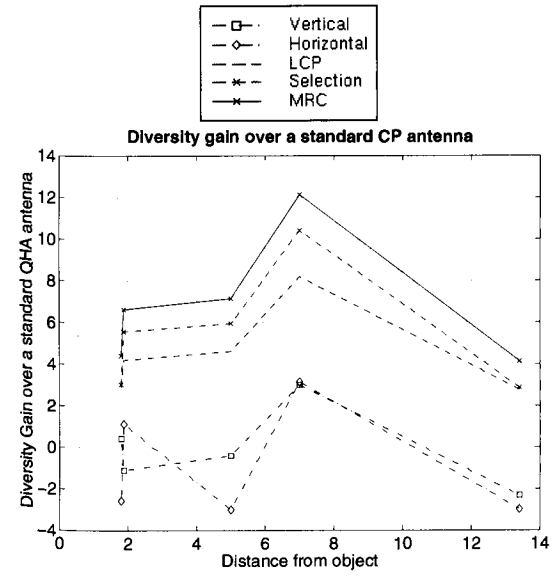


Figure 6: Diversity gain versus distance from a reflecting building

In the case of reflection (Figure 6) even the linear polarised signals have diversity gain over an RCP antenna for a small range of distances of the mobile receiver from the building. It can also be seen that the dominant circular polarisation is in fact LCP in this case, due to the sense of rotation of the wave being swapped upon reflection. This is related to the Brewster angle[3]. At low elevation angles of the satellite, the angle of incidence will be similar to that used this study. Figure 6 indicates that MRC will achieve diversity gain at all distances measured in this study when the signal is reflected by a building. MRC would still give significant diversity gain over an LCP antenna in this situation.

Discussion of Results

It is clear from the values of diversity gain obtained that the presence of a building in the path of low elevation satellite signals can have a serious impact on the polarisation state of the received signal. The fact that good diversity gain is obtained indicates that there is a significant amount of decorrelation between the horizontal and vertical components of the received signal in these situations.

The inferior performance of an MRC post detection weighting system in an area where the mobile is not in deep shadow or in open cases where there may well be in a line of sight link with the satellite is a disadvantage to this type of system. There may well be arguments that a 3dB degradation in a line of sight performance is a reasonable price to pay for diversity gain of around 10dB in regions of deep shadow such as urban areas. The answer to this may well be specific to each individual satellite communications provider based on the system and the link margin which is designed into it.

There is however another alternative which alleviates the problem of the MRC signal being inferior to that of a CP antenna in a line of sight situation. This solution is to use a pre detection weighting system. In this kind of system, the two signals are weighted at RF, combined and then fed into the receiver, thus there is only one receiver which is the dominant source of noise, thus the signal to noise ratio of the system is dependent on one source of noise not two. This has the effect of increasing the SNR of the MRC system by up to 3dB.

There are some side effects of this pre detection weighting system, namely that phase and amplitude weighting circuits need to be implemented at RF rather than a much lower IF. Optimising these weights at RF frequencies with only one receiver output is much more difficult and time consuming, and thus longer training periods are required to get the optimum SNR from the system. From a mobile terminal point of view, however there is the added advantage in terms of size and power consumption of having only one receiver.

CONCLUSIONS

The potential benefits of using a polarisation diversity technique based on maximum ratio gain combining has been presented. Benefits in areas of deep shadow or urban environments have been indicated in the region of up to 10dB of diversity gain. The possible disadvantages of a diversity system have also been highlighted in terms of performance degradation in line of sight environments and the added complexity of the mobile receiver in terms of size, weight and power consumption, all of which are very critical in design of mobile handsets nowadays.

Potential benefits of diversity without the system performance degradation and with hopefully less impact on the size, weight and battery life problem have also been alluded to. A trade off must be made between additional diversity gain available and the complexity and ease of realisation of an industrially viable product.

ACKNOWLEDGEMENTS

The authors acknowledge the support of Nokia Mobile Phones and Mobile VCE, a collaborative venture of more than 20 industrial companies and 7 UK Universities, with the financial support of the UK government.

REFERENCES

[1] R.G.Howell, J.W.Harris and M.Mehler, Satellite crosspolar measurements at BT laboratories, BT Technology Journal vol.10, no.4, pp.52-67, October 1992.

[2] A.A.Agius, S.M.Leach, P.Suvannapattana, S.R.Saunders, Intelligent handset antenna research within Mobile VCE, in Proceedings of the intelligent antenna symposium at Surrey , 1998

[3] J.D.Kraus, Electromagnetics 4th Edition, McGraw Hill, 1992 ch.13

[4] S.R.Saunders, Antennas and propagation for wireless communication systems, J Wiley & Sons, 1999, ch.15

Fade and Non-fade Statistics for Land Mobile Satellite Communication with Inclined Orbit Satellite

Tetsushi Ikegami, Ken'ichi Kaburaki and Shin'ichi Yamamoto*
Department of Electronics and Communication, Meiji University
1-1 Higashimita, Tama-ku, Kawasaki 214-8571 Japan
Tel: +81-44-934-7312, Fax: +81-44-934-7909
e-mail: ikegami@isc.meiji.ac.jp
*Kashima Space Research Center, Communications Research Laboratory,
Ministry of Posts and Telecommunications
893-1 Hirai, Kashima, Ibaraki 314 Japan

Abstract

This paper reports land mobile satellite propagation measurements at higher elevation angles using ETS-VI satellite at S-band (2.1 GHz) frequency. During the links between the ETS-VI and a ground earth station are established, the satellite is available at 50 to 67 degrees in elevation, 10 to 20 degrees higher than one with geostationary satellites. Propagation data at the same test course and different elevation angles to the satellite are taken that describe elevation dependent properties of fade and non-fade duration statistics. Measured data show that shadowing statistics depend on the elevation angle and that the use of inclined orbit satellite greatly enhances the availability of land mobile satellite communication systems. Conditional distributions of non-fade duration after fade state are analyzed for possible application to packet data communications.

Introduction

Land mobile satellite links suffer from shadowing and blockage due to road side trees and buildings. Recently, personal satellite communication systems which utilize low and medium earth orbit satellites are planned and extensive studies are performed all over the world. In these systems, the effects of shadowing will be relaxed because the satellite can be seen at relatively high elevation angles from earth stations compared with geostationary satellite systems. This will become an advantage for land mobile satellite system from the propagation stand point of view.

This paper deals with land mobile satellite propagation measurements at higher elevation angles using ETS-VI satellite at S-band (2.1 GHz) frequency. During the links between the ETS-VI and a ground earth station are established, for 3 to 4 hours, the satellite is available at 50 to 67 degrees in elevation, 10 to 20 degrees higher than one with geostationary satellites. Extensive propagation measurements are undertaken with both omni-directional and directional (15 dBi in gain) antennas at measuring van in urban, suburban and rural areas in Japan including Tokyo, Chiba, Ibaraki and Okinawa[1].

Propagation data at the same test course and different elevation angles to the satellite are taken that describe elevation dependent properties of fade and non-fade duration statistics. Non-fade duration, which is defined as a duration that received signal level exceeds a certain threshold (usually 3 to 6 dB), is a measure of link availability of mobile satellite systems. Because a fade margin of more than 20 dB is not practical in mobile satellite systems of limited available transmission power. Measured data show that shadowing statistics depend on the elevation angle and that the use of higher elevation satellite greatly enhance the availability of land mobile satellite communication systems. In order to evaluate the link availability of mobile satellite channels,

especially for packet data communications, we propose an analysis based on conditional distributions of non-fade duration after the fade state.

Measurements[1]

Outline of measurements is shown in Fig.1. Ka-band 30GHz carrier is transmitted to ETS-VI (Engineering Test Satellite six) satellite from base atation at Kashima Space Research Center. A transponder onboard the ETS-VI converts the Ka-band signal into S-band 2.1GHz signal and transmits it back to the ground in left-handed circular polarization. The test van receives the S-band signal with directional and omni-directional antennas and the measured signal strengths are recorded in DAT data recorder with running pulses of the van. Propagation losses and Doppler shift caused by satellite movement are compensated for by controlling the output frequency and the level of the signal generator at the base station. The received signal level, speed of the test van and distance pulses every 5 cm are recorded on a data recorder. Extensive propagation measurements are undertaken in urban, suburban and rural areas in Japan including Tokyo, Chiba, Ibaraki and Okinawa. In this paper, we focus on the results in suburban Kashima-city of Ibaraki obtained with a direcitonal antenna.

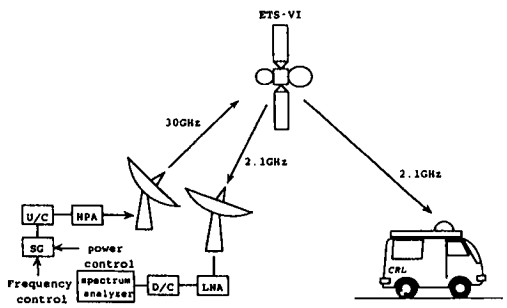


Fig. 1 Outline of measurements

Results

Fig.2 shows a cumulative distribution of the received signal level in Kashima. The abscissa shows the relative received power with respect to the line-of-sight signal power. The ordinate shows the probability that the recieved power is less than the abscissa value in Gaussian scale.

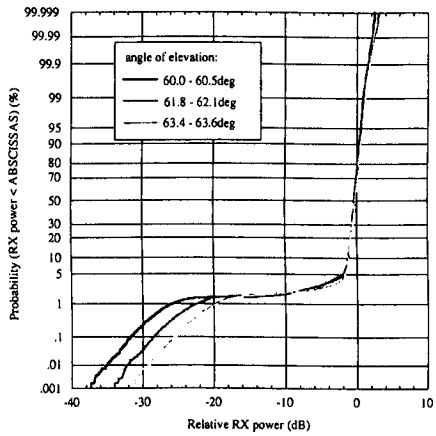


Fig. 2 Cumulative distribution of received power (Kashima)

For each run (elevation angles: 60.0-60.5, 61.8-62.1 and 63.4-63.6 degrees), the test van ran the same test course in Kashima-city. Kashima-city is a suburban small city with few tall strucute. Most of the roadside buildings are less than 3 stories. The characteristics of each elevation change at the –1 dB point, so a region above this point corresponds to line-of-sight and the region below corresponds to shadowing. Shadowing probabilities are about 5% of the total measurement duration, however becoming lower as the angle of elevation increased. Within the range of –1 to –20dB, received signal is not blocked completely and attenuated or scatterd by small structures, such as utility poles or trees. In order to evaluate the link availability of mobile satellite channels, fade and non-fade duration analysis of the received signal measurements is necessary. The fade duration(FD) means the distance at which received signal is continually below a certain threshold level. The non-fade duration(NFD) means the distance at which received level is continually above a threshold. In this paper, conditional distibutions of NFD after fade state at each elevation are analysed.

Figures 3-8 correspond to conditional distibutions of FD event after a certain non-fade state and NFD event after a certain fade state in Kashima, respectively. Each line means conditional probability that NFD or FD is greater than abscissa value, provided that previous FD or NFD is greater than the designated value.

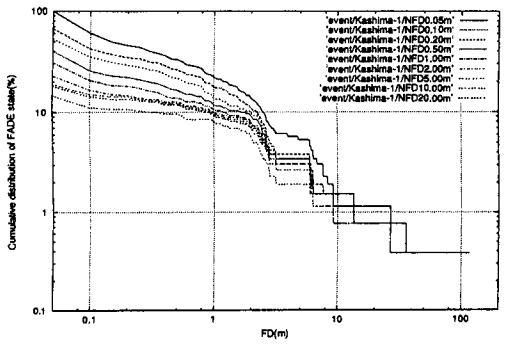


Fig. 3 Conditional cumulative distribution of FD event (Kashima, El=60.0-60.5 deg.)

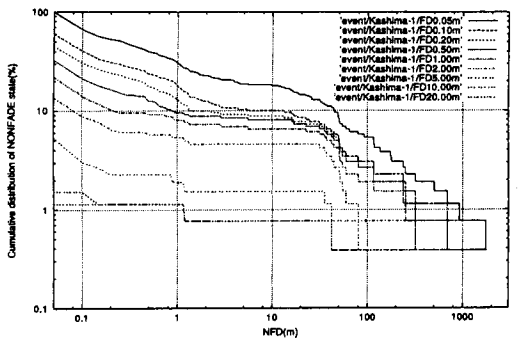


Fig. 4 Conditional cumulative distribution of NFD event (Kashima, El=60.0-60.5 deg.)

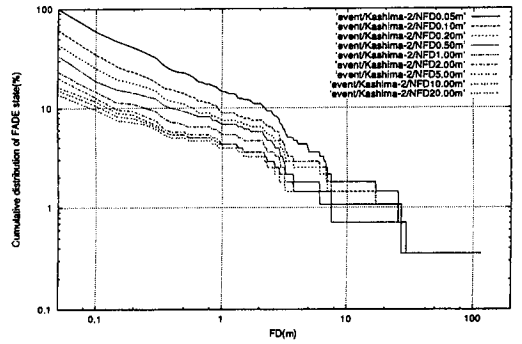


Fig. 5 Conditional cumulative distribution of FD event (Kashima, El=61.8-62.1 deg.)

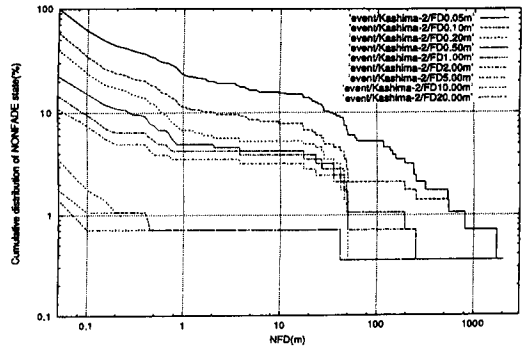


Fig. 6 Conditional cumulative distribution of NFD event (Kashima, El=61.8-62.1 deg.)

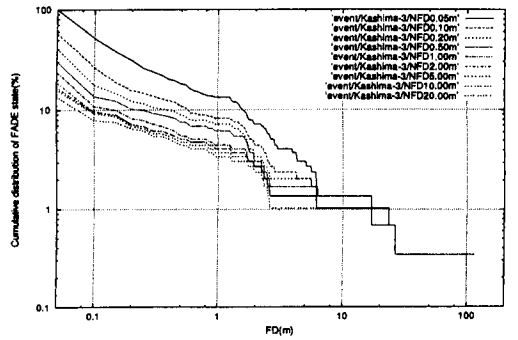


Fig. 7 Conditional cumulative distribution of FD event (Kashima, El=63.4-63.6 deg.)

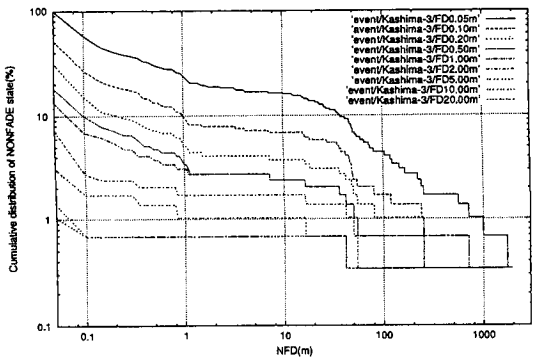


Fig. 8 Conditional cumulative distribution of NFD event (Kashima, El=63.4-63.6 deg.)

In order to clarify the link availability, cumulative distributions of Fig. 3-8 are scaled to total running distances of the test van at each run. Results are shown in Figure 9-14.

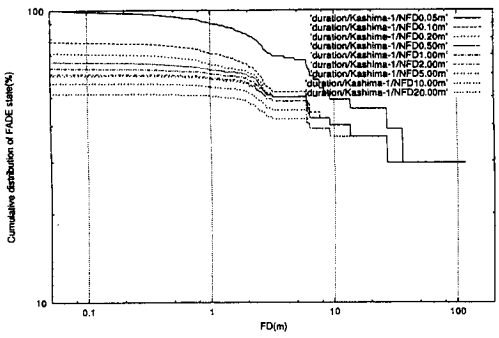


Fig. 9 Conditional cumulative distribution of FD, normalized in total run (Kashima, El=60.0-60.5 deg.)

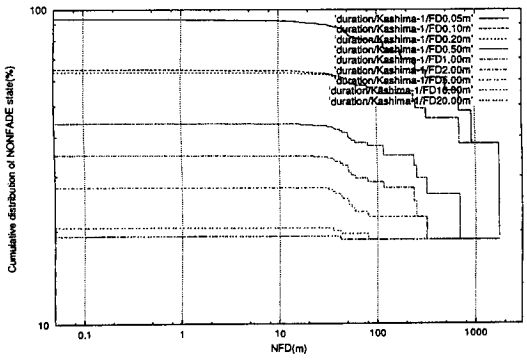


Fig. 10 Conditional cumulative distribution of NFD, normalized in total run (Kashima, El=60.0-60.5 deg.)

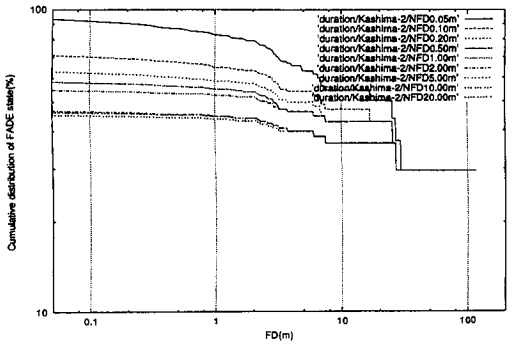


Fig. 11 Conditional cumulative distribution of FD, normalized in total run (Kashima, El=61.8-62.1 deg.)

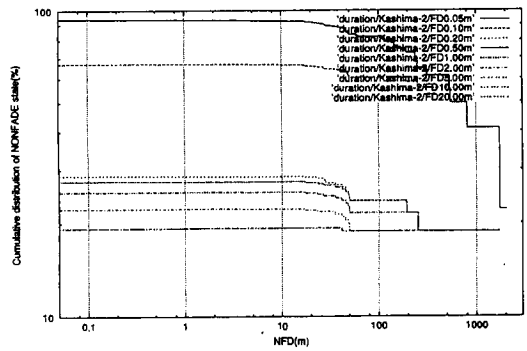


Fig. 12 Conditional cumulative distribution of NFD, normalized in total run (Kashima, El=61.8-62.1 deg.)

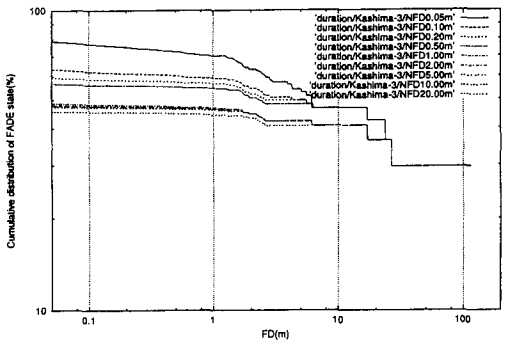


Fig. 13 Conditional cumulative distribution of FD, normalized in total run (Kashima, El=63.4-63.6 deg.)

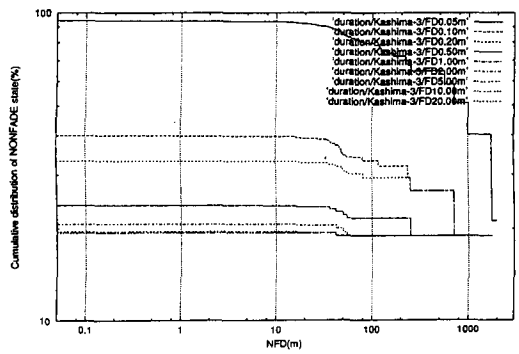


Fig. 14 Conditional cumulative distribution of NFD, normalized in total run (Kashima, El=63.4-63.6 deg.)

From these figures, following facts can be seen:

Long NFD over 500 m increases and relatively short NFD event of 0.2 to 2 m decreases as angle of elevation increases.

NFD with previous fade state of 0.05 m decreases more rapidly as angle of elevation increases than with fade state of 0.2 m.

These fact explain that long NFD of more than 1 km with short fade state, i.e. good link state, occur as elevation increases.

Conclusions

An analysis of S-band (2.1 GHz) land mobile satellite propagation measurements at higher elevation angles using ETS-VI is described. In order to evaluate the link availability of mobile satellite channels, an analysis based on

conditional distributions of non-fade duration after the fade state is proposed. Measured data show that shadowing statistics depend on the elevation angle and that the use of higher elevation satellite greatly enhance the availability of land mobile satellite communication systems. Results with different environments in urban, suburban and rural areas including Tokyo, Chiba, Ibaraki and Okinawa will be reported in the near future.

Reference

[1]S. Yamamoto, E. Okamoto, A. Ali, T. Ikegami, "Propagation Measurements for Land Mobile Communications at S-band with non-GEO ETS-VI satellite", Proc. International Mobile Sat. Conf., pp.177-182, Pasadena, June 1997.

Aeronautical Channel Measurement Trials at K Band

Matthias Holzbock, Axel Jahn, Erich Lutz
DLR, German Aerospace Center, Institute for Communications Technology
P.O.B. 1116, 82330 Oberpfaffenhofen, Germany
E-mail: Matthias.Holzbock@dlr.de

ABSTRACT

A measurement campaign is presented for high data rate satellite services at K-band using an aeronautical terminal. A mechanically steered high-gain Cassegrain antenna with a beamwidth of 4.2° is used for signal reception. Both differential GPS and avionic data based open-loop steering as well as a closed loop step tracking can be used for antenna control. ITALSAT's 18.685 GHz beacon is used for channel measurements by monitoring the received signal strength in inphase and quadrature component. The recorded data was processed in the laboratory and channel statistics have been derived.

I. INTRODUCTION

In the framework of SECOMS (Satellite EHF Communications for Mobile Multimedia Services, an European ACTS project) and ABATE (ACTS Broadband Aeronautical Terminal Experiment) several field trials and service demonstrations have been performed for future wideband multimedia satellite services at higher frequency bands (20/30 GHz and 40/50 GHz) [LLV97]. In the SECOMS project a satellite system is being developed for mobile multimedia throughout Europe. A constellation of up to five satellites operating at K/Ka and EHF bands will provide high data rate services for all kind of mobile terminals. The ABATE sub-project is focusing on the utilisation of this system for aeronautical applications; for example in-flight entertainment or telemedicine. Besides the here discussed aeronautical channel characterisation measurements, DLR has conducted trials involving high capacity bi-directional links for multimedia communications for service demonstration [HJM98].

In this paper a aeronautical satellite experiment at 18.685 GHz is highlighted. Since the behaviour of the of the mobile satellite channel has a crucial impact on the design and development of satellite-based communications networks, comprehensive field trials are necessary for system design. Moreover, the aeronautical propagation at higher frequencies is still little known. The purpose of the measurement campaign was i) to perform measurements in order to collect a channel database with a wide range of flight scenarios; and ii) to investigate antenna steering algorithms for airborne applications.

Although attenuation caused by shadowing of the wings or the tail structure can be predicted in for aeronautical applications, the effects are yet not well known especially when the satellite's elevation is in order of the bank or pitch angels of the aircraft. For example diffraction of the signal path by tail structure crossing or multipath fading effects caused by reflections are of special interest and were investigated during special flight manoeuvres.

Also flights during different weather conditions, below and above clouds, including normal in-flight manoeuvres and forced shadowing manoeuvres were performed. The link performance while the antenna exposed to the heavy vibrations during start and landing was investigated.

Fig. 1 - Aeronautical testbed DO 228II.

II. MEASUREMENT TESTBED

The receiving antenna was mounted on the two engine turboprop aircraft DO 228 of DLR as shown in Fig. 1. Important parameters of the aircraft are listed below.

- flight characteristics
  - max. altitude 12 000 ft (without oxygen)
  - 8 h flight duration, max. cruising speed ca. 250 knots
  - operable under non-freezing weather conditions
  - all autopilot and landing aids available
- volume
  - 2 pilots, one mechanic, operators or passengers, dep. on payload, max. 16 passengers
  - 2000 kg payload mass
- basic equipment
  - antenna outlet, GPS (differential, RTCM/NMEA 183), time ref. system (IRIG-A/B), inertial Gyro, compass with MagnetoFlux compensation (ARINC-429 interface), aircraft instrumentation bus (ARINC 429)

This aircraft type was chosen in order to simulate the in-flight behaviour of most of the today's aircraft types and to perform flight situations with forced shadowing from

wings and tail structure. Drawback was the cruising speed of only 250 knots and the max. flight altitude of 12 000 ft, both cruising speed and flight altitude are less than those of an airliner. But a higher cruising speed will impact the mainly the frequency shifts which can be also calculated theoretically. The flight altitude was sufficient for performing flights above the clouds. On the other hand it was possible to perform forced shadowing being of prime interest. For example, flight manoeuvres with up to 50° banked circles or ascent and decent with nose up (15° pitch) and nose down (-15° pitch) were performed as well as typical holding patterns.

III. MEASUREMENT SET-UP

The measurement set-up basically consists of an antenna rack and a rack which holds the control, monitoring and down-converter units. The receiver consists of two-stage demodulators transferring the 20 GHz signal to a 2.15 GHz and a 70 MHz intermediate frequency band. Another in phase and quadrature demodulator is mixing the received signal into baseband.

At the moment two ITALSAT satellites are transmitting a 20 GHz beacon. The orbital position of both differs by 3° and the beamwidth of the receiving antenna is 4.2°. In order to avoid interference on the measurements one of the two satellite beacons was switched off during the trials.

The receiving antenna is a Cassegrain antenna with 4.2° beamwidth (Fig. 2) produced by DLR. A side lobe suppression of min. 22 dB and a extreme light weight structure yield excellent pointing performance.

The steering platform allows a tracking of the antenna with an azimuth speed of 70°/sec and elevation speed of 40°/sec. Two control modes have been implemented for the antenna pointing, acquisition and tracking (PAT): i) open-loop tracking based on differential GPS and avionic data, and ii) closed loop step tracking algorithm. Most of the data was recorded using the open loop algorithm. A GPS reference station located at DLR premises and on-line correction of the GPS data in the aircraft via a 30 MHz differential GPS link was used in order to achieve best geographical position and flight altitude data of the aircraft. This leads to a position accuracy of 1...2 m. Three axis stabilised gyroscopes and additional compasses and angular sensors implemented in the aircraft set high precision attitude data distributed via a standardised ARINC 429 aircraft instrumentation bus to our disposal. With this knowledge of the aircraft's position, the aircraft's attitude and position of the geostationary satellite it is possible to calculate the antenna pointing angle. Geographical (latitude, longitude, altitude) and attitude (pitch, roll, magnetic heading) data was monitored during all flight trials and stored.

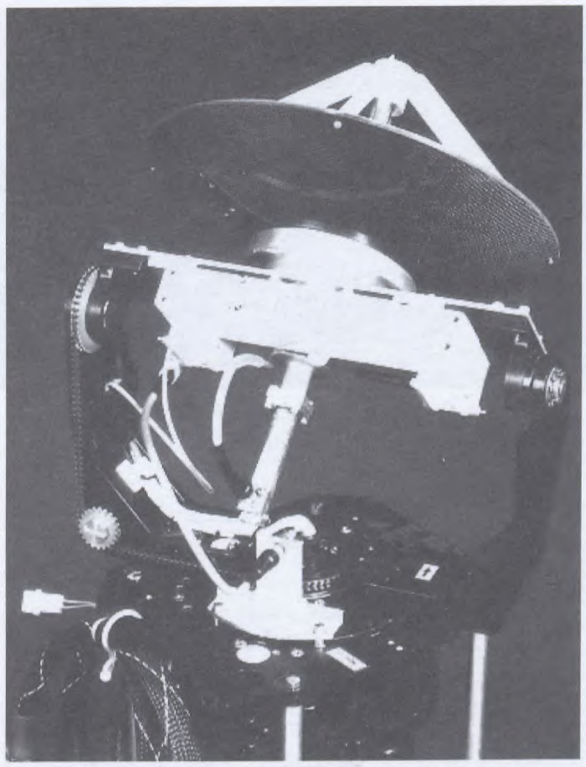


Fig. 2 - Picture of the steered platform with antenna and low noise amplifier

The pointing accuracy of this algorithm is hard to determine. Variations of the received power during a normal cruise with line-of-sight conditions give information about the pointing error, but variations have been very small and could be caused by other effects, too. The same PAT algorithm was also implemented for land-mobile trials. Here the satellite was simulated by an aircraft and the receiver was implemented in a vehicle. A video camera mounted on the steered antenna platform allowed to prove the resulting steering accuracy. The pointing error was less than 0.6° standard deviation [LJ97]. Since all equipment was identical the pointing error is assumed equal for the aeronautical experiment.

The antenna platform was top mounted on the aircraft between the wings and tail structure (Fig.3). Rotary joints enable full service coverage from -7° to 90° in elevation and several rotations in azimuth. The platform was covered by a radome. A low noise amplifier (LNA) with 50 dB gain was mounted directly at the antenna feed. It is followed by the rotary joints the two down-converters to 70 MHz (Fig. 4). Then the received signal is down-converted to baseband and recorded in inphase and quadrature component on a DAT recorder with 10kHz bandwidth. Doppler compensation was implemented to counteract frequency shifts. The actual frequency offset at 70 MHz was measured and a control signal was calculated with a PC. Based on this signal one of the synthesisers was tuned via GPIB bus to compensate for Doppler shifts.

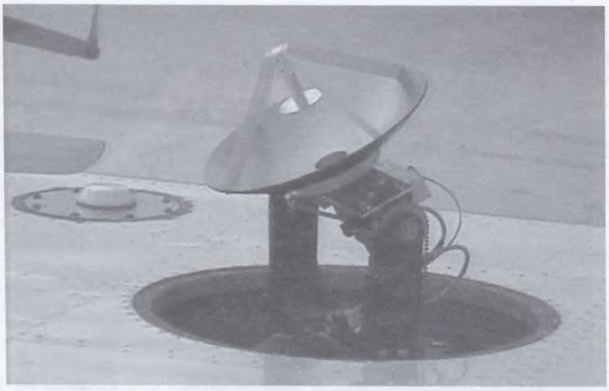


Fig. 3 - Picture of the aircraft mounted Cassegrain antenna and the GPS antenna

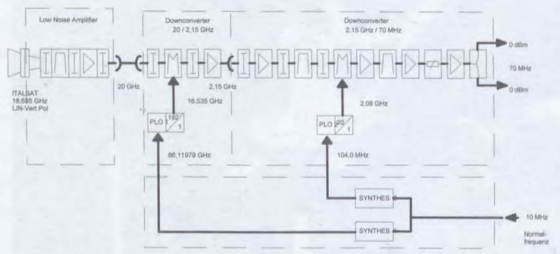


Fig. 4 - 18.685 GHz beacon reception and down-conversion to 70 MHz RF/IF section

Figure 5 shows a view inside the aircraft's cabin fitted with all the measurement equipment.

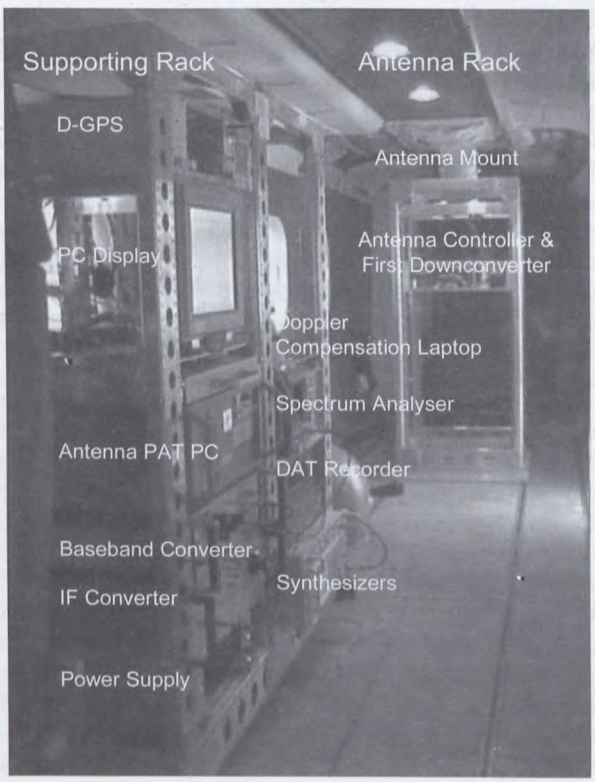


Fig. 5 - View of the measurement equipment inside the aircraft

IV. MEASUREMENT SCENARIOS

Trials during different flight scenarios and weather conditions were performed. All flights have been performed in an area of Munich, Germany.

The flight conditions comprise: Start and landing, ascent and descent, normal cruise, in flight manoeuvres, touch and go, and forced shadowing manoeuvres : for example turns with 45 degrees roll angle and curves with 20 degrees roll angle and 10 deg pitch angle.

The weather conditions comprise: rainy, cloudy, sunny, flights below, throughout and above clouds.

Before take-off, the inertial sensors of the airplane and the synthesisers involved in the reception of the satellite beacon were carefully powered up in order to reach their stability point.

For all flights, a high accuracy time handling was assured. The measurement and controlling PCs were synchronised using GPS time and programmed to store each geographical, attitude and Doppler compensation data with GPS time. Secondly, this GPS time was recorded with the I and Q components of the received beacon via IRIG-B signal to enable synchronisation and recovery of each reported scenario. Those cares enables a precise synchronisation of the different PCs and data-recorders.

Figure 6 shows a representation of the flight path of the airplane during a measurement flight. This picture was processed in the laboratory using the recorded position data (GPS).

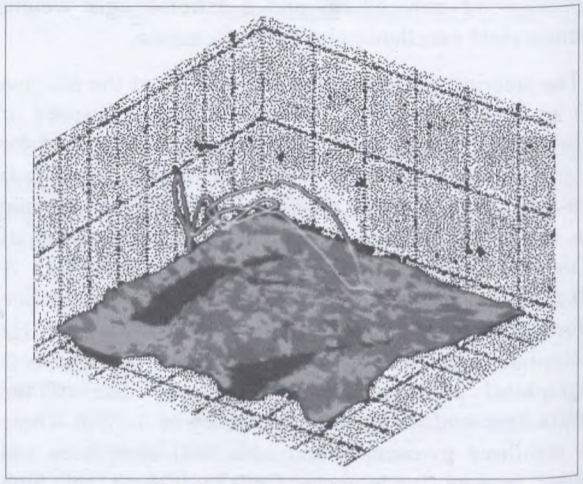


Fig. 6 - Typical flight path for a measurement scenario

V. RESULTS

Data processing in the laboratory was necessary to read the position and attitude from the Antenna PAT PC, the Doppler compensation data from the Doppler compensation laptop, and the received signal from the DAT recorder. The complete set of data is available for

each reported scenario. First, the results concerning a complete flight, then some records of the most interesting flight scenarios (normal flight, waiting loop, circle) are presented in what follows.

A. DOPPLER SHIFT MEASUREMENT

Measurements were made with and without Doppler compensation. Without Doppler compensation, the frequency of the recorded signal is more accurate but the Doppler shift can exceed the bandwidth of the measurement DAT recorder. With Doppler compensation, the Doppler shift is set to zero every second. We have recorded the Doppler correction for a complete flight. The frequency shifts of the Doppler compensated signal are about 200 Hz. The variations of the Doppler shift during the flight (fig. 7) show: i) no variation during a cruise flight, ii) maximum variation during U-turn manoeuvres. The maximum variation is about 8kHz. This implies that it would be about 12 kHz for an airliner at cruising speed.

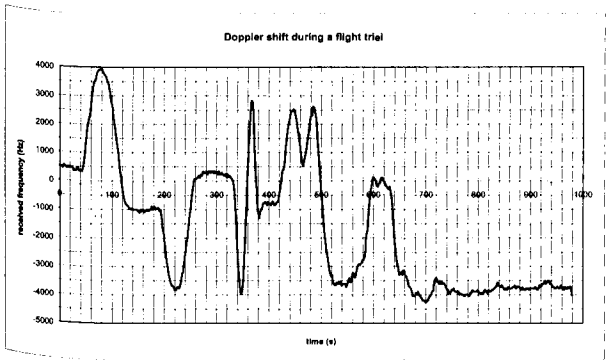


Fig. 7 - Doppler shift variations during a measurement scenario

The theoretical value of the Doppler shift between a moving airplane and a GEO satellite can be calculated by:

$$f_{Doppler} = \frac{Vd}{c} \cdot f_{Carrier}$$

with the Doppler velocity Vd

$$Vd = \frac{1}{\sqrt{rp^2 - 2\cos\alpha \cdot \cos\beta \cdot rp \cdot rs + rs^2}} \cdot \left(Vz \cdot [rp - rs \cdot \cos\beta \cdot \cos\alpha] + Vlat \cdot [rs \cdot \sin\beta] - Vlong \cdot [rs \cdot \cos\beta \cdot \sin\alpha]\right)$$

Where r_S is radius of the satellite orbit (42000km),
α is the latitude of the aircraft,
β is the longitude difference,
Vlat is the velocity in latitude direction,
Vlong is the velocity in longitude direction,
Vz is the vertical velocity of the airplane.

Using the recorded values of position and attitude, it was possible to calculate the theoretical Doppler shift. Figure 8 shows a comparison of measured and calculated Doppler shift during a U-turn. The observed behaviour is in our case a continuous variation of the frequency between +4 kHz and -4 kHz. These results confirm the analysis that have been performed before the flight trials.

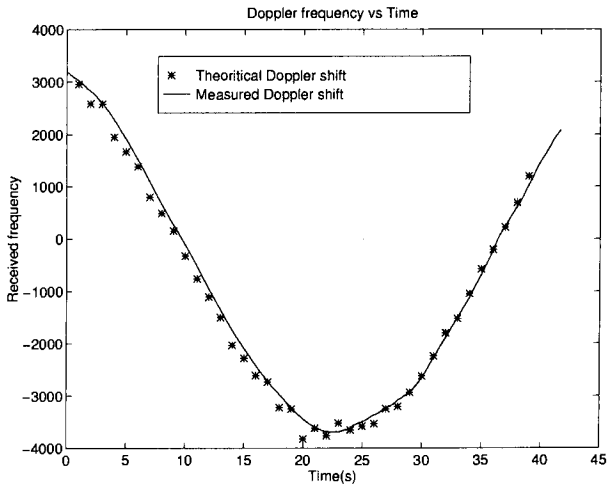


Fig. 8 - Comparison of theoretical and measured Doppler shift variation during a U-turn

B. RECEIVED POWER ANALYSIS

Besides analysing the recorded flight parameters, the received inphase and quadrature signals have been processed, yielding channel statistics. The digitised received signals were processed in an FFT-analysis. We used overlapped signal samples to increase the resolution of this analysis and we have also processed the FFT with a high resolution (4096 points). A special algorithm has been developed to correct the error introduced by frequency steps when the Doppler compensation was used. The Figure 9 shows the principles of our signal analysis.

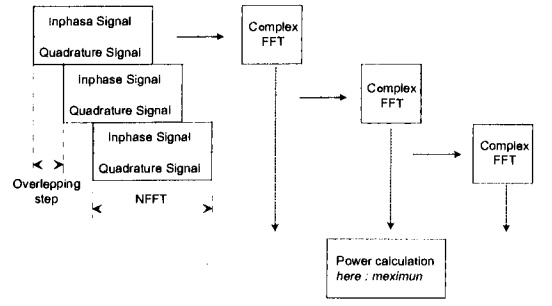


Fig. 9 - Algorithm used to process the received signals

The processed results have been carefully analysed. The results let clearly appear two states: i) when the signal was received without shadowing (this case is the most current); and (ii) the second one when the reception was attenuated because of the manoeuvres. This also confirms the Lutz Model [LCD91].

Also other effects can be observed. For example reflection due to the tail or nose structure or mean received power variation by clear or rainy weather.

C. LINE OF SIGHT CASE

The 'line of sight' case has occurred, as it was explained before, when no shadowing was present: this was the case for instance when the airplane was on the taxiway, in cruise flight and also in ascent and descent.

In this case the received power was nearly constant and the variation from the mean value can be describe using a Rican probability density function.

$$p_{RICE}(S) = c \cdot e^{-c(S+1)} I_0\left(2c\sqrt{S}\right)$$

Where S is the momentary received power (with the 0 dB reference set to mean(S)), c is the direct-to-multipath ratio (Rice factor) and I_0 is the modified Bessel function of zero'th order.

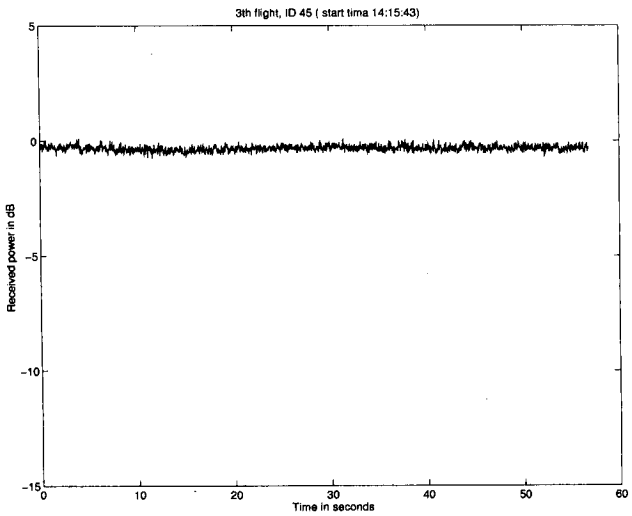


Fig. 10 - Received power for a typical line of sight case

Figure 10 shows a graph of a received power versus time for a cruise flight and figure 11 shows the corresponding probability density function.

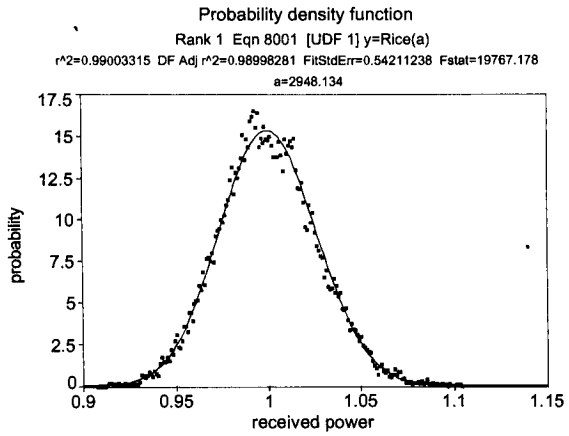


Fig. 11 - Probability density function for the same line of sight case

Using a data interpolation software it was possible to determine the value of the Rice probability density function. For example figure 11 shows a line of sight scenario. The corresponding Rice factor is c = 34 dB.

D. NON LINE OF SIGHT FLIGHT CASES

In second scenario the multipath fading or signal shadowing occurs. The reception level is changing during the flight manoeuvres. For a flight turn for instance the reception level decreases during the inclination of the body of the airplane (roll angle about 20 degree), because the wing disturbs the reception of the signal. This can be seen in figure 12. We have also tried to increase this effect by making critical turns with roll angles about 45 degrees (satellite elevation was approximately 35°). The signal shadowing of up to 15 dB can be observed when the wing crosses the line of sight as shown on figure 13.

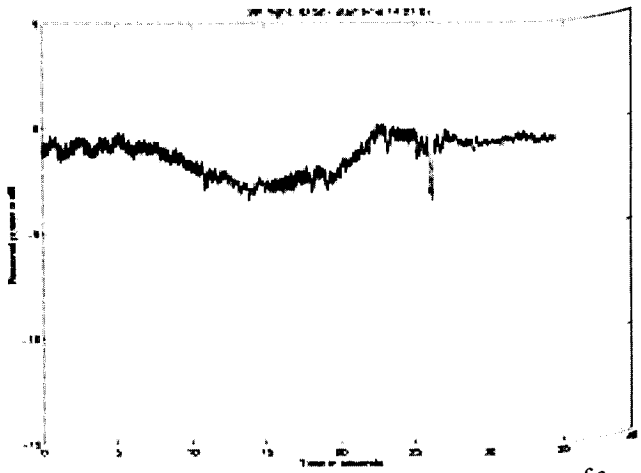


Fig. 12 - Attenuation of the received signal because of a wing.

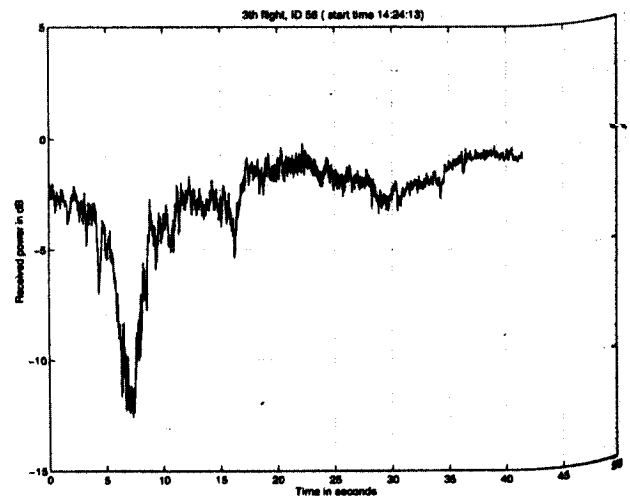


Fig. 13 - Shadowing of the signal because of the airplane structure (wing)

In figure 14 the aircraft flew in meanders like a wavy line in line with the propagation path from the satellite. In this way the aircraft tail crossed periodically the propagation path of the satellite signal, causing diffraction and shadowing. The fade depth is about 2-3 dB. These effects can be clearly identified in Fig. 14. The corresponding aircraft attitude is given in Fig. 15. The period of the aircraft movement corresponds to the signal fades

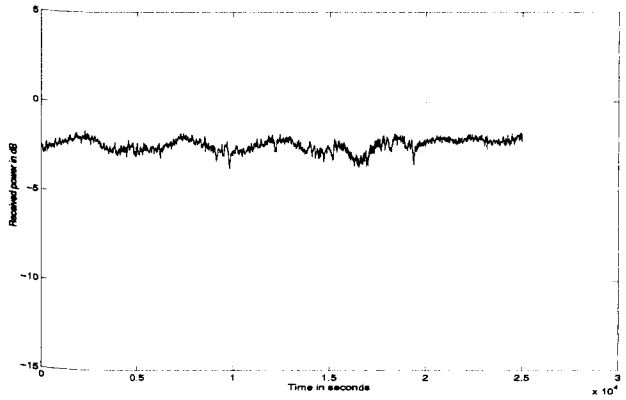


Fig. 14 - Shadowing of the signal because of the aeroplane structure (tail) during meander manoeuvres

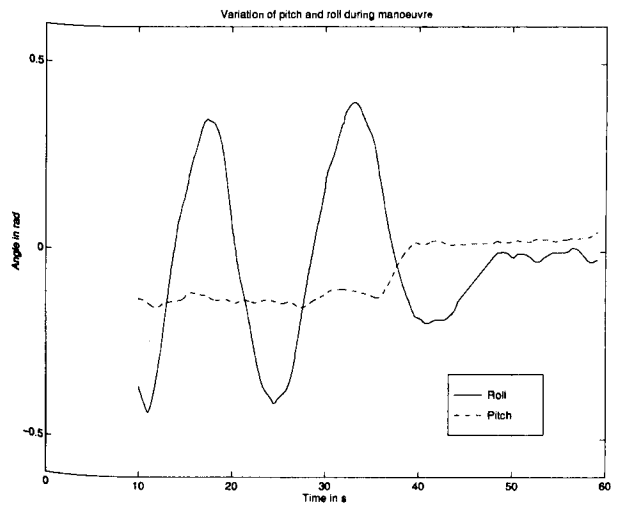


Fig. 15 - Aircraft attitude for the meander manoeuvres.

Another result is the influence of the weather condition on the received signal level. Flights under and above clouds were performed. The mean difference between the two measured received signal levels is 0.78 dB. Figure 16 shows the measured data.

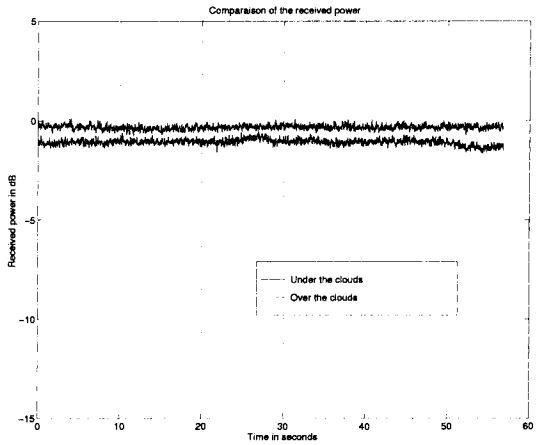


Fig.16 - Received power level comparison of different weather conditions

REFERENCES

[LLV97] G. Losquadro, M. Luglio F. Vatalaro: "A geostationary satellite system for mobile multimedia applications using portable, aeronautical and mobile terminals," *Proceedings 5th International Mobile Satellite Conference (IMSC'97)*, pp. 427--432, 1997

[HJM98] Matthias Holzbock, Axel Jahn, Massimo Barbieri, Jean-Pierre Grao: Broadband End-to-End Satellite Service Demonstrator in the Frame of ABATE, *ACTS Mobile Summit 1998,* 1998,

[LCD91] E. Lutz, D. Cygan, M. Dippold, F. Dolainsky, W. Papke: "The land mobile satellite communication channel - recording, statistics and channel model," *IEEE Trans. Vehicular Technology*, vol. 40, 1991, pp 375--386.

[JL97] A. Jahn, E. Lutz: "LMS channel measurements at EHF-band," *Proceedings 5th International Mobile Satellite Conference (IMSC'97)*, pp. 183--188, 1997.

Large Distance Site Diversity in Satellite Communication Systems: Long Term Experimental Results obtained in Italy with the Synthetic Storm Technique

Emilio Matricciani[1], Luciano Ordano[2], Luca Iorio[1]
[1]Politecnico di Milano, DEI & CNR-CSTS, Piazza L. da Vinci, 32-20133 Milano, Italy
[2]CSELT, Via G. Reiss Romoli, 274-10148 Torino, Italy
[1]matricci@elet.polimi.it; [2]ordano@cselt.it

ABSTRACT

Satellite system design for fixed or mobile communications, in bands affected by rain attenuation can benefit from knowing information on the simultaneous performance of distant stations (*large distance site diversity*) during rain. Also feeder links designed for a very high availability (of the order of 99.99% of the time) can benefit from it. To assess how distance affects site diversity performance, we have processed a large database of rain rate time series available to CSELT, for 9 years and for 12 sites, in Italy. Distance ranges from 3.3 km to 593.8 km. We have converted the rain rate time series into realistic rain *attenuation time series* in slant paths to Italsat at 39.6 GHz, circular polarization, by applying the synthetic storm technique. The results can be taken as experimental data. We have not included the attenuation due to oxygen and water vapor, because these effects do not impact on the diversity performance, as they should not significantly change within the simultaneous rain events. Cloud attenuation could be a factor to include for low power margin systems, but we have not modeled this effect. In other words, the results below concern only rain attenuation.

INTRODUCTION

Several experiments and radar simulations have provided data for small site diversity, i.e. the usually considered site diversity scheme with two (or more) stations, with distance ranging from few kilometers to some tens of kilometers. Reliable experimental information on large distance site diversity (tens to hundreds of kilometers) is very scarce and, in any case, is very lengthy and costly to obtain, so that estimates were derived from rain rate time series (e.g. [1]).

In this work we have used rain rate time series, but we have converted them into realistic rain *attenuation* time series in slant paths, by applying the synthetic storm technique as developed in [2]. We have already tested it for several applications and sites, including a case of large distance site diversity for three sites [3], with good overall results.

If we apply this technique to rain rate time series concurrently collected for many years at several distant sites, then we can simulate several receivers and transmitters, and thus estimate the site diversity (joint) probability distribution for couples of sites, for triples, etc.

To assess how distance between sites affect site diversity performance, we have processed a large rain rate database available to CSELT, for 9 years and for 12 sites in Italy, Table I. Distance ranges from 3.3 km (small distance site diversity) to 593.8 km (large distance site diversity), Table I. The rain rate time series (with data effectively averaged over about 3 minutes, subsequently sampled every minute) have been converted into rain attenuation time series at 39.6 GHz (circular polarization) to simulate slant paths to the geostationary satellite ITALSAT (13.2° E) as seen from each site (elevation angle ranges from 36° to 42°, Table II).

REVIEW OF THE SYNTHETIC STORM TECHNIQUE

We sketch a brief summary of the fundamentals of the synthetic storm technique as developed in [2], to which we refer the reader for the theory and details. A synthetic storm is obtained by converting a rain rate time series, recorded at a point by a rain gauge, and usually averaged over 1 minute, to a rain rate space series along a line, by using an estimate of the storm translation speed v, to transform time to distance, according to the uniform rectilinear motion law $s=vt$.

Evidence is good that the statistical properties of rain attenuation, derived from a large sample of rain rate time series, closely agree with those of the actual measured rain attenuation, and that the predictions are insensitive to the value of v. Rain attenuation time series can be simulated by using the average value of v, obtainable from measurements of wind speed at about the 700-mbar level, often available from meteorological or aeronautical services. In the following we have assumed v=10.6 m/s, as in [2], a value estimated from radar measurements in the Po Valley.

Table I. Distances (km) between couples of sites with rain rate time series recorded in the years 1971-1979.

Site	Ber	Bor	Cod	Don	Fuc	Lat	Lod	RCR	EUR	RMM	Son	Vit
Bereguardo	-	66.8	114.4	98.3	514.3	521.2	37.4	464.9	471.7	461.8	119.4	399.1
Borgomanero	66.8	-	97.6	78.5	580.9	587.9	91.3	531.6	538.4	528.5	118.0	465.8
Codera	114.4	97.6	-	19.2	574.2	593.8	101.9	538.7	545.9	535.8	29.3	472.6
Dongo	98.3	78.0	19.2	-	572.4	589.6	90.5	534.1	541.3	531.2	42.9	467.9
Fucino	514.3	580.9	574.2	572.4	-	76.2	493.8	88.9	88.7	91.5	551.9	130.3
Latina	521.2	587.9	593.8	589.6	76.2	-	505.5	56.4	49.8	59.5	574.1	122.3
Lodi	37.4	91.3	101.9	90.5	493.8	505.5	-	449.5	456.6	446.5	98.0	383.3
Roma C. R.	464.9	531.6	538.7	534.1	88.9	56.4	449.5	-	7.5	3.3	519.4	66.2
Roma EUR	471.7	538.4	545.9	541.3	88.7	49.8	456.6	7.5	-	10.1	526.8	73.4
Roma M. M.	461.8	528.5	535.8	531.2	91.5	59.5	446,5	3.3	10.1	-	516.6	63.3
Sondrio	119.4	118.0	29.3	42.9	551.9	574,1	98.0	519.4	526.8	516.6	-	453.7
Viterbo	399.1	465.8	472.6	467.9	130.3	122.3	383.3	66.2	73.4	63.3	453.7	-

The vertical structure of rain is modeled with two layers of precipitation of different depths. Starting from ground there is rain (hydrometeors in the form of raindrops, water temperature of 20°C), followed by a melting layer, i.e., melting hydrometeors at 0°C. The rain rate in the lower layer, R(mm/h), is assumed to be uniform and given by that measured at ground, i.e. by the rain gauge. With simple physical hypotheses, calculations show that also the precipitation rate in the melting layer, termed "apparent rain rate", can be supposed to be uniform and given by $3.134R$ [4].

The height of the precipitation (rain and melting layer) above sea level at a site is assumed to be the ITU-R 0°C isotherm height above sea level, and the depth of the melting layer to be 400 m, regardless of the latitude of the site [2]. The two layers, used together, may describe for rain attenuation prediction, on the average, and effectively, all rain events made up of the two main types of precipitation, i.e. stratiform and convective.

The synthetic storm technique is a powerful rain attenuation prediction method, which can reproduce not only the average annual rain attenuation probability distribution, *P(A)*, in a given slant path [2], but also its dynamic characteristics such as fade durations [5], power spectra [6], and worst month statistics [7]. Suitably adapted to a *2-D* space simulation it can also be used to assess the *P(A)*'s to be expected in satellite systems with mobile terminals [8], and it reproduces the probability distributions of rain rate cells dimensions [9].

The values of the constant k and α used in the specific rain attenuation formula $\gamma=kR^{\alpha}$ (dB/km) are directly given by the ITU-R [10], after the Laws and Parson raindrop size distribution, water temperature 20°C. The constants for the melting layer, i.e. drops at 0°C, are given by [11], in which there are also the 20°C parameters recommended by the ITU-R. They can be interpolated at 39.6 GHz, circular polarization, for which they are independent of slant path elevation angle.

Table II. List of sites with rain rate time series recorded in the years 1971-1979, with chronological data. The slant paths are to the geostationary satellite Italsat (13.2°E)

Site	Latitude (°N)	Longitude (°E)	Altitude a.s.l. (m)	Elevation Angle (°)
Bereguardo	45.25	9.03	98	37.7
Borgomanero	45.70	8.47	306	37.2
Codera	46.23	9.47	750	36.7
Dongo	46,12	9.28	200	36.8
Fucino Borgo 8000	41.98	12.48	650	41.5
Latina	41.50	12.90	12	42.1
Lodi	45.32	9.50	80	37.7
Roma Collegio Rom.	41.90	12.48	20	41.6
Roma EUR	41.83	12.50	32	41.7
Roma Monte Mario	41.92	12.45	139	41.6
Sondrio	46.17	9.83	298	36.8
Viterbo	42.42	12.08	327	41.0

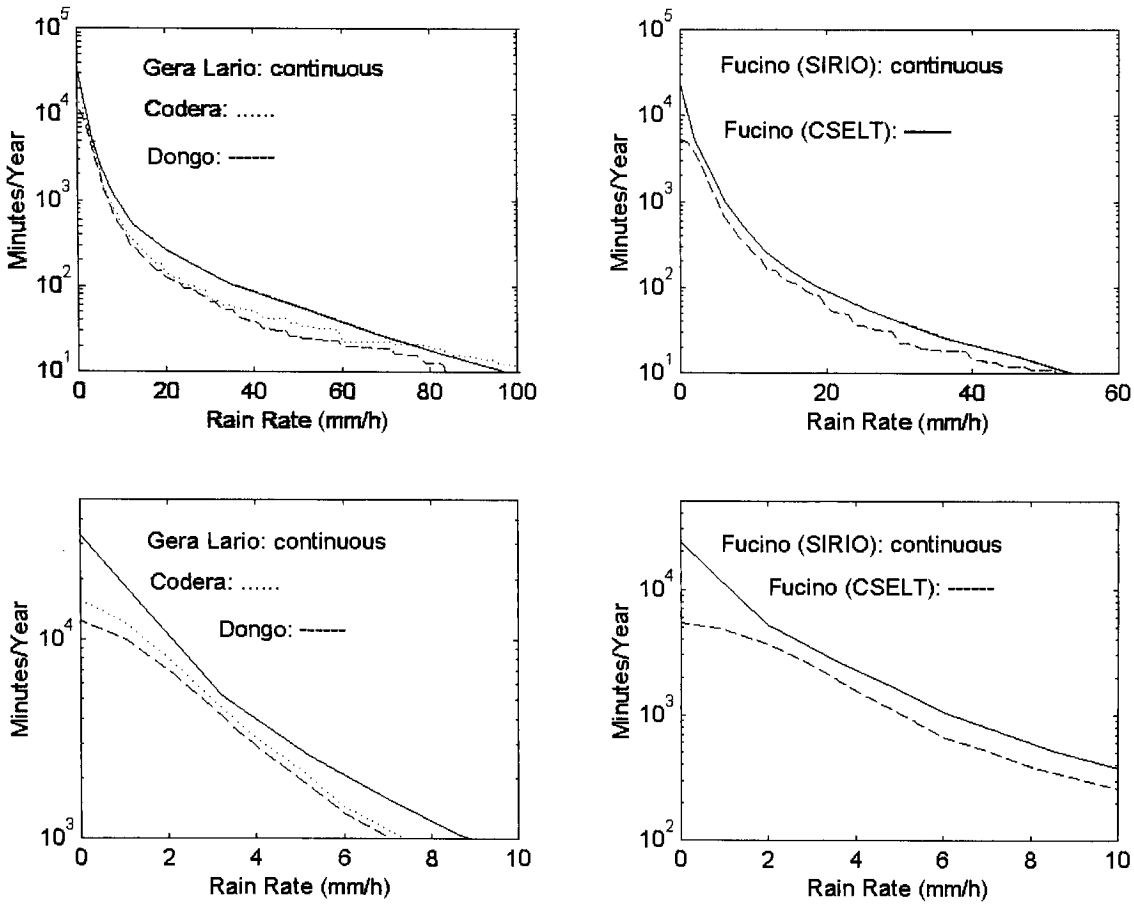


Figure 1. Cumulative total number of minutes R is exceeded in an average year in the years 1971-1979 (CSELT) and in the years 1978-1982 (SIRIO experiment, continuous line). Notice expanded abscissa scale in figures below.

RAIN RATE DATA BASE

The data base is made up of rain rate time series (averaged over about 3 minutes, subsequently sampled every minute) collected with tipping-bucket rain gauges in 12 sites in Italy for 9 years (1971-1979), originally recorded on paper and later digitalized by CSELT for radio propagation studies [12]. These data have been selected from a larger database concerning 68 sites, as they report chronological data, indispensable for site diversity assessment. The distance between couples ranges form a minimum of 3.3 km (Roma Collegio Romano, Roma Monte Mario) to a maximum of 593.8 km (Codera, Latina), Table I.

Not all rain events recorded on paper, however, were digitalized by CSELT, but only those for which: *(a)* the amount of total water collected during the rain event exceeded 4 mm, or *(b)* the maximum rain rate observed during the event exceeded 4 mm/h. In other words, the rain events with very low rain rates are not included in the data base. This is an unfortunate event because satellite systems with very low power margin are designed for high outage probabilities corresponding to these low rains.

This database limit is clearly shown in Figure 1, which reports, as an example, the cumulative total number of minutes R in abscissa was exceeded in an average year (i.e. averaged over the years 1971-1979), for three sites: Codera, Dongo in Northern Italy, and Fucino in Central Italy (once divided by the number of minutes in a year, 525600, these curves yield the probability distribution). As Codera and Dongo are only few kilometers away from Gera Lario, a site where a Telespazio satellite station is still located, and for which a reliable long term rain rate distribution is available from the times of the SIRIO experiment [13][2], 5 years (1978-1982), their distributions are compared to that of Gera Lario: we see that significant differences occur just for R<2-3 mm/h, which is consistent with the data base. Similarly the Fucino CSELT data are compared to the reliable long-term rain rate curve, measured in the same site during the SIRIO experiment in 1978-1982 [13][2]. Significant differences appear again for R<2-3 mm/h. The fact that in both cases

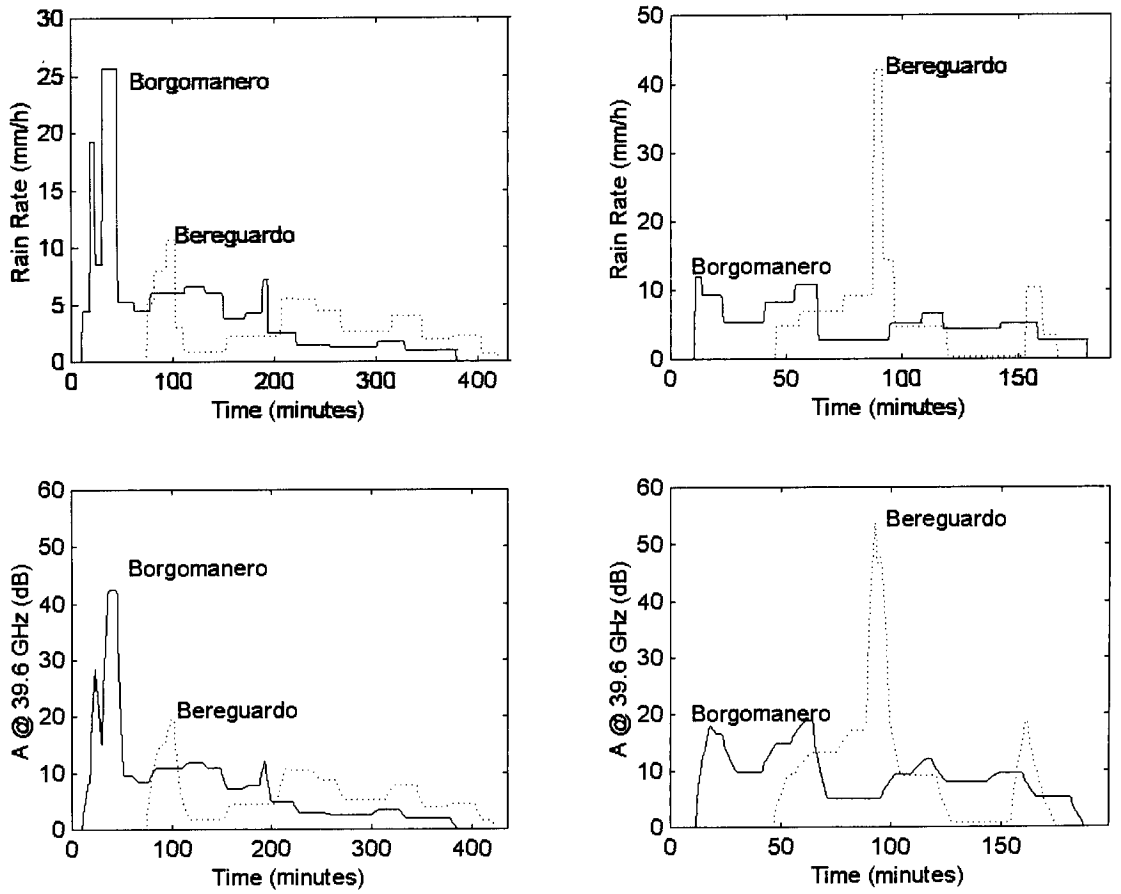


Figure 2: Examples of rain rate time series and corresponding rain attenuation time series for two sites. *Left column*: the event starts at 19:36 (local time), 28-05-1971, and ends at 1:45, 29-05-71 at Borgomanero. *Right column*: the event starts at 11:40, 20-08-1971 and ends at 14:30, 20-08-1971 at Borgomanero. Distance between sites is 66.8 km.

the CSELT distributions are below those of the SIRIO experiment could be due to the different observation periods considered (1971-1979 against 1978-1982). The same behavior is found for all other sites.

Notice, however, that the main parameter that we will be shortly using, the diversity probability factor h, is less sensitive to this data base limit, so that the results below are meaningful also for the very low attenuation.

Figure 2 shows an example of the simultaneous rain attenuation time series that can be generated form the rain rate time series. Compared to rain attenuation time series derived from real 1-minute rain rate time series (e.g., Figure 1 of [3]), the ones shown in Figure 2 are more "squared": they appear more similar to the rain rate time series than to beacon measurements, as, on the contrary, is the case with 1-minute data.

SITE DIVERSITY PROBABILITY FACTOR

Given two sites i, j, at a distance d_{ij}, and their long term probability distributions of exceeding the same rain attenuation $A_i=A_j=A$ (dB) at a given carrier frequency, $P_i(A)$, $P_j(A)$, site diversity system performance can be estimated from the (bivariate) joint probability distribution of exceeding simultaneously A, i.e. $P_{i,j}(A)$. The same information, however, can be retrieved from the probability factor $h_{ij}(A,d_{ij})$, defined as:

$$h_{ij}(A,d_{ij}) = P_{ij}(A,d_{ij})/[P_i(A)P_j(A)] \qquad (1)$$

i.e. the ratio between the joint distributions in the real case and in the case of statistical independence. This probability factor has the advantage that, from the knowledge of the marginal distributions $P_i(A)$, $P_j(A)$, it yields $P_{i,j}(A,d_{ij})$ directly. Now, since the marginal distributions $P(A)$ can be predicted with good models, such as the synthetic storm technique or others, a knowledge of h_{ij} as function of distance d_{ij}, and possibly, of attenuation A, is very useful.

Notice that h_{ij} is linked to the so-called probability improvement, defined (for site i) according to:

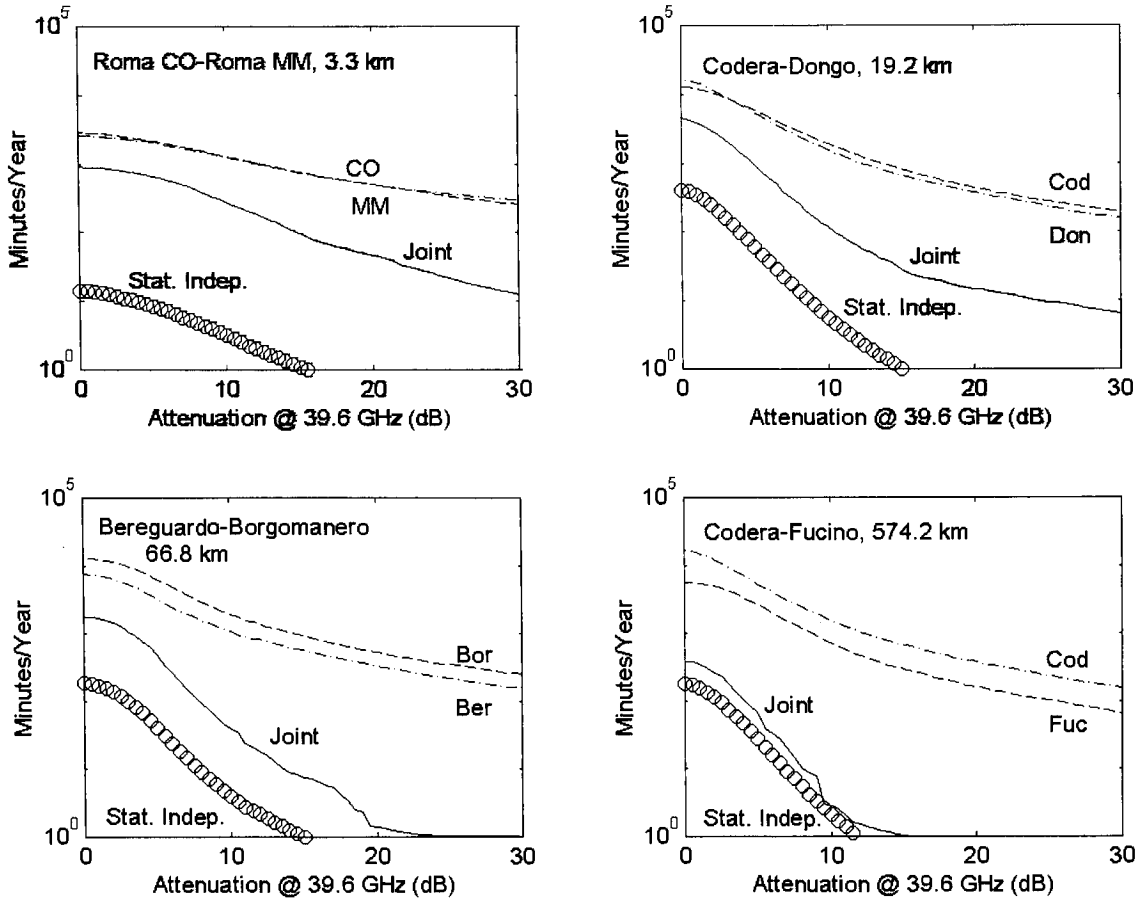


Figure 3. Cumulative total number of minutes A is exceeded in an average year, in the years 1971-1979, in the single links, in the joint link, and in case of statistical independence, at some sites.

$$I_i(A,d_{ij}) = P_i(A) / P_{ij}(A,d_{ij}) \qquad (2)$$

as

$$I_i(A,d_{ij}) = 1/[h_{ij}(A,d_{ij})P_j(A)] \qquad (3)$$

The probability factor h_{ij} ranges form $h_{ij}=0$ (mutually exclusive events), when only one radio link is faded by rain (supposedly for very distant sites) to $h_{ij}=1/P_i(A)=1/P_j(A)$ (the reciprocal of the annual probability distribution) when $d_{ij}=0$ km. The statistical independence case yields $h_{ij}=1$.

RESULTS AND CONCLUSIONS

Figure 3 shows some examples of the average cumulative times obtained for some couples of sites, in the single radio links, in the joint link, and in case of statistical independence. Several sound physical features can be noticed. The roman sites (Roma Collegio Romano, CO, and Roma Monte Mario, MM) yield, as expected, the same marginal distributions (expressed in minutes/year); also Codera (Cod) and Dongo (Don) yield about the same distributions since they are not too far. In both cases the joint distributions (here expressed in minutes/year) are far from the statistical independence, as expected.

For the other two couples, the marginal distributions are different, the joint distribution is still far from the statistical independence in the Bereguardo-Borgomanero case (66.8 km), but it is very close to it in the Codera-Fucino case (574.2 km).

The unusual downward concavity of these distributions at the low attenuations is due, of course, to the missing low rains, as mentioned above. These curves are hence reliable approximately for A>4-5 dB.

From curves like those reported in Figure 3, we can calculate the factor h_{ij}, eq.(1). Figure 4 shows some examples. We can see that for distances just below 60 km, A increases as h_{ij} gets larger. Just beyond 60 km, h_{ij} tends to be a constant and then, for larger distances, it falls off to $h_{ij}=1$ (statistical independence), or even $h_{ij}=0$ (no rain attenuation in one of the two sites). For a given A, h_{ij} is larger as the distance gets smaller, since it must approach a larger value $1/P(A)$ when d_{ij} approaches zero.

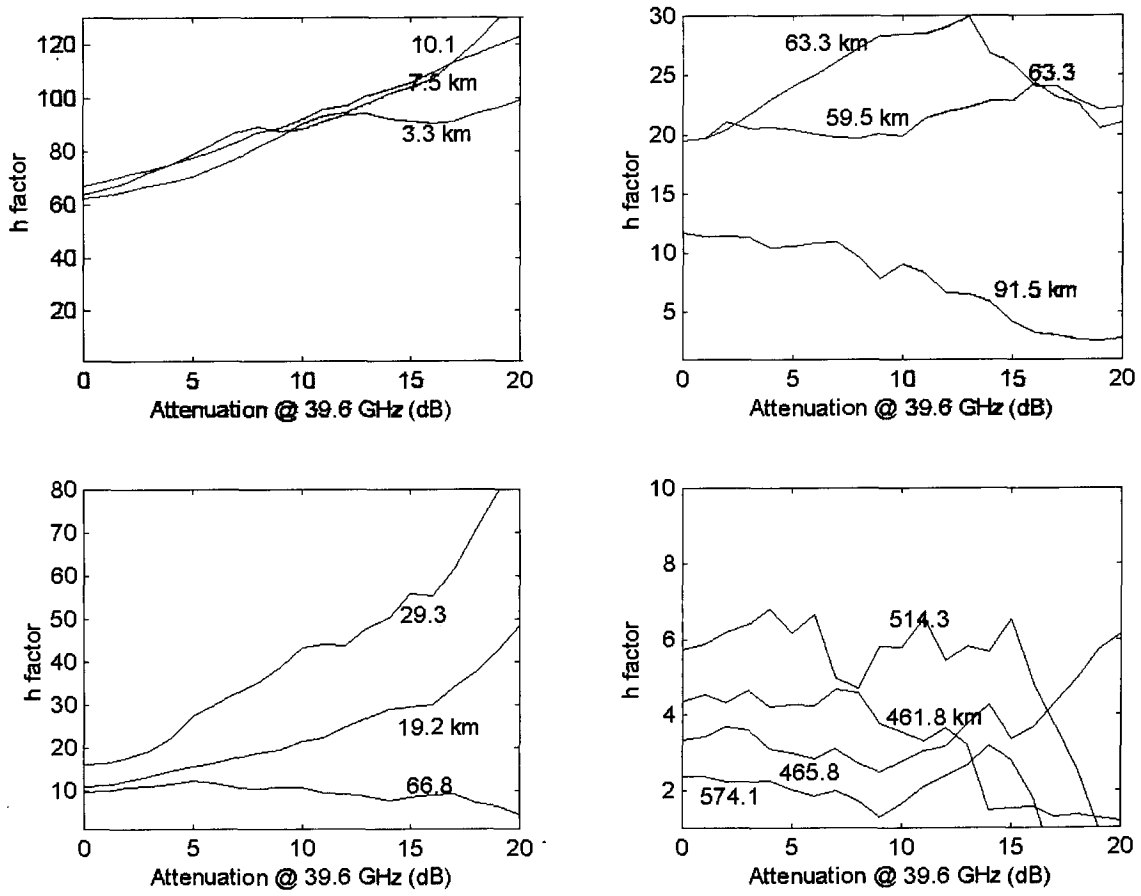


Figure 4. Examples of how the probability factor h changes as a function of A jointly exceeded, and distance between two sites.

Since h_{ij} is given by the ratio of eq.(1), we may expect that the low rains that were not considered in the analysis in $P_i(A)$, $P_j(A)$ and $P_{i,j}(A,d_{ij})$ (although with different weights), have less impact on its values. This seems to be case, according to Figure 4, which shows a regular behavior. A numerical example will help: let us consider Roma CO and Roma MM, which have the same distributions, and are only 3.3 km apart, so that we can assume $d_{ij}\approx 0$ km. In this case $h_{ij}=1/P_i(A=0dB)=1/P_j(0)=62$ according to Figure 4. This means that $P_i(0)=P_j(0)=1/62=1.6\times10^{-2}$, compared to an expected 3×10^{-2}-4×10^{-2} [13]. The distributions of Figure 3, when referred to a year, yield the less realistic value 6×10^{-3}.

Figure 5 shows $h_{ij}(A, d_{ij})$ for the entire data base (66 couples) as a function of the exceeded A. The case $A=0$ is frequency independent since it applies whenever $A=0^+$. We notice that h_{ij} is a strong function of distance, with a behavior close to the statistical independence at $d_{ij}\approx 600$ km. A range in which h is almost a constant, for any A, is below about 10 km (small site diversity). It is also interesting to notice that h is a function of rain attenuation: for a given distance d_{ij}, h increases as A becomes larger, in agreement with the higher limit it has to approach when $d_{ij}\approx 0$ km.

In conclusion, the synthetic storm technique seems to be a reliable, useful and cheap way to assess site diversity performance for a given rain attenuation A jointly exceeded. Future work will be, however, necessary to fill up the missing distances in Figure 5, especially below 10 km, between 100 and 600 km, and above 600 km. The final end would be to find a mathematical model of $h_{ij}(A, d_{ij})$.

REFERENCES

[1] F. Barbaliscia, G. Ravaioli, A. Paraboni, Characteristics of the spatial statistical dependence of rainfall rate over large areas, *IEEE Trans. on Antennas and Propagation*, 40, 1992, 8-12.

[2] E. Matricciani, Physical-mathematical model of the dynamics of rain attenuation based on rain rate time series and a two-layer vertical structure of precipitation, *Radio Science*, 31, 1996, 281-295.

[3] E. Matricciani, Wide area joint probability of rain attenuation useful to design satellite systems with a common on-board resource: experimental results obtained with the synthetic storm technique in Italy, *Fourth K_a Band Utilization Conference*, Venice, Nov.2-4, 1998, 271-

277.
[4] E. Matricciani, Rain attenuation predicted with a two-layer rain model, *European Trans. on Telecommunications and Related Technology*, 2, 1991, 715-727.
[5] E. Matricciani, Prediction of fade durations due to rain in satellite communication systems, *Radio Science*, 32, 1997, 935-941.
[6] E. Matricciani, Physical-mathematical model of
[8] E. Matricciani, S. Moretti, Rain attenuation statistics useful for the design of mobile satellite communication systems", *IEEE Trans. Vehicular Tech.*, 47, 1998, 637-648.
[9] E. Matricciani, A. Pawlina Bonati, Rain cell size statistics inferred from long term point rain rate: model and results, *Third Ka Band Utilization Conference*, Sorrento, Sept. 15-18, 1997, 299-304.
[10] ITU-R, Specific attenuation model for rain for use in prediction methods, *Propagation in Non-Ionized Media*, Recommendation 838, Geneva, 1992.
[11] D. Maggiori, Computed transmission through rain in dynamics of rain attenuation with application to power spectrum, *Electronics Letters*, 30, 1994, 522-524.
[7] E. Matricciani, Worst-month statistics of rain attenuation in a satellite link at 19.77 GHz: experimental results derived with the synthetic storm technique for the station of Gera Lario, *Fourth K_a Band Utilization Conference*, Venice, Nov.2-4, 1998, 287-292.

the 1-400 GHz frequency range for spherical and elliptical drops and any polarization, *Alta Frequenza*, 50, 1981, 262-273.
[12] G. De Renzis, G. Dellagiacoma, L. Ordano, Caratteristiche pluviometriche del territorio italiano, *Elettronica e telecomunicazioni*, 4, 1987, 177-187 (in Italian).
[13] "Special Issue on the SIRIO Program in the tenth year of satellite life", *Alta Frequenza*, 1987, No.1-2, 56.

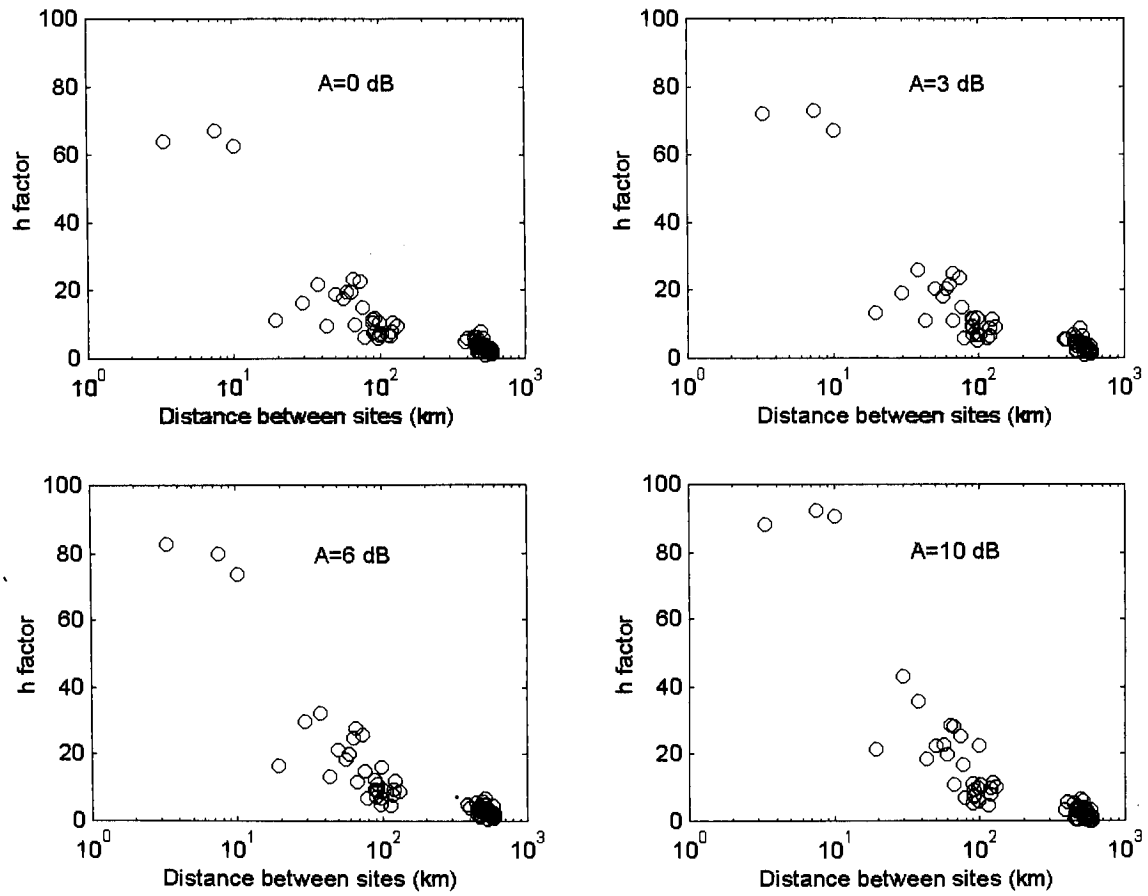


Figure 5. Probability factor h as a function of distance between two sites and rain attenuation jointly exceeded at 39.6 GHz. Notice that the case A=0 dB is frequency independent.

Satellite Constellations for Millimeter Wave Communication in the Northern Hemisphere, Using Barbaliscia's 49/22 GHz Measurements

Paul Christopher

Stanford Telecommunications
45145 Research Place
Ashburn, VA 20147
paul.christopher@acs-stel.com

ABSTRACT

This paper uses a new approach to millimeter wave satellite communication studies for the Northern Hemisphere. It starts with new and important [1,2] attenuation results in Part 1 and applies these results to high elevation angle satellite constellations in Part 2. The high elevation angles of a Teledesic system and of a Molniya system imply new attenuation maps appropriate to the constellations. Then, the 40-50 GHz and 90-100 GHz regions are examined for suitability at Cuba, New York, and Hudson Bay.

1. BARBALISCIA'S ZENITH ATTENUATION MAPS

Barbaliscia et al [1] have recognized that a large portion of satellite communications will occur during non- rainy conditions. This will be especially important for the millimeter wave communication which the Italians envision for frequencies well over 40 GHz. They have constructed topographic maps for zenith attenuation over the whole Earth[2] for 1% non-rainy conditions. This may be thought of as cloud and water vapor conditions which are almost associated with rainfall. Their results for 49.5 and 22 GHz are especially interesting because they offer the possibility of separating cloud and water vapor effects. The method for separating the effects is discussed elsewhere[3], and the result for zenith attenuation in the Northern Hemisphere as a general function of frequency is included here as the Appendix.

Attenuation of Figure 1-1, consists largely of water vapor at the 22.2 GHz resonance and some cloud attenuation. The 49.5 GHz attenuation of Figure 1-2 is assumed to be composed largely of cloud attenuation and (nearly constant) 1.2 dB oxygen attenuation. Cloud attenuation is assumed to vary as f^2. The attenuation of Figures 1 and 2 may be solved simultaneously to separate cloud and vapor.

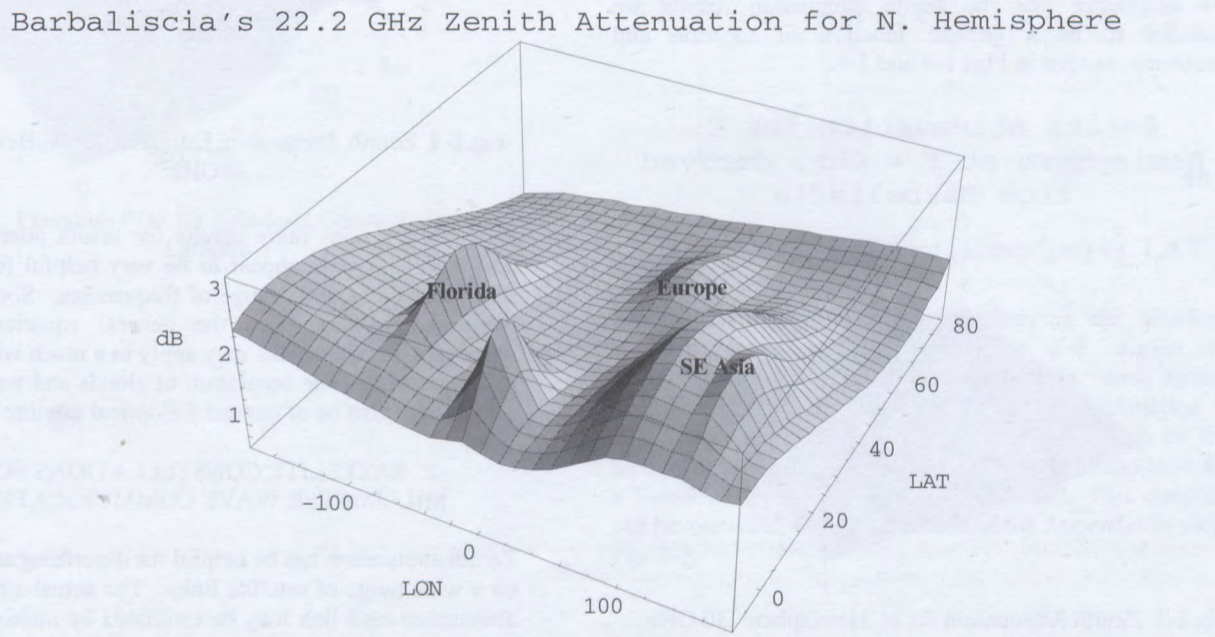


Fig. 1-1 Barbaliscia's 22.2 GHz Zenith Attenuation in Smoothed Form For Northern Hemisphere; note Longitude – 180 to +180 Deg

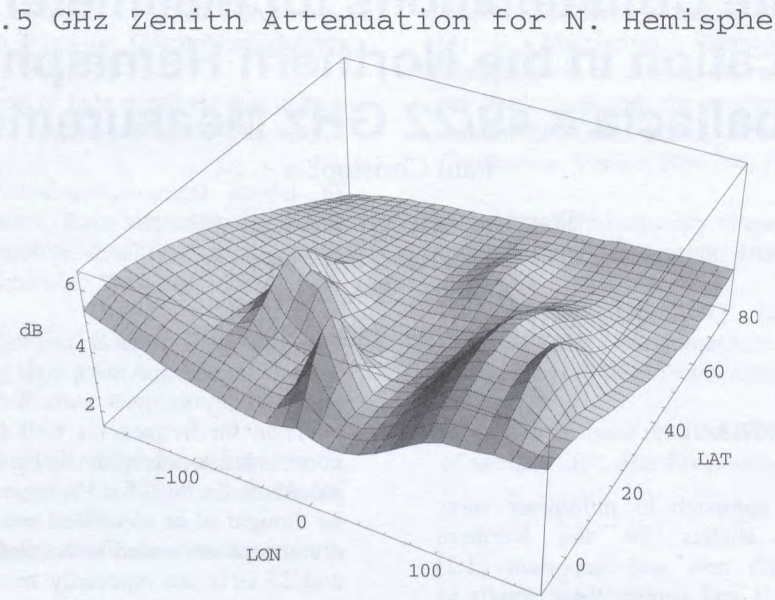


Fig.1-2 Barbaliscia's 49.5 GHz Zenith Attenuation in Smoothed Form, N. Hemisphere

For perspective, attenuation peaks at Florida, West Africa, and Southeast Asia appear in the figures from left to right, respectively.

After clouds and water vapor are separated, other interesting estimates can be found. The water vapor attenuation can be associated with a surface humidity, and the results (as function of LON, LAT, and frequency) can be substituted into an integrated gaseous attenuation model (Appendix)[4,5]. Clouds can also be added to give attenuation estimates for a wide range of frequencies.

We emphasize that the zenith attenuation results are intended to be a general function of location and frequency, as seen in Figs 1-3 and 1-4.

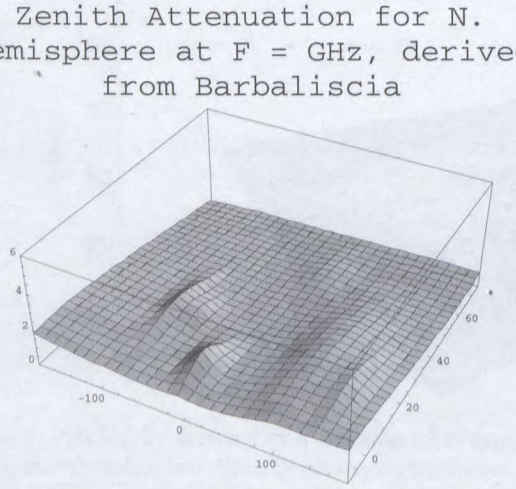


Fig.1-3 Zenith Attenuation for N. Hemisphere 30 GHz (note peaks at Florida, West Africa, SE Asia)

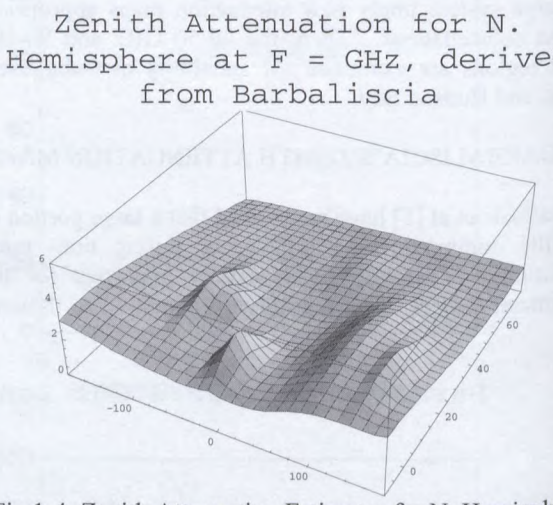


Fig.1-4 Zenith Attenuation Estimates for N. Hemisphere 40GHz

Barbaliscia's non- rainy results for zenith attenuation at 22.2 and 49.5 GHz appear to be very helpful for finding new results at a wide range of frequencies. Some results were shown here from the general equation of the Appendix, but the results may apply to a much wider range of frequencies. The separation of clouds and water vapor effects may also be of interest for optical satellite links.

2. SATELLITE CONSTELLATIONS FOR MILLIMETER WAVE COMMUNICATION

Zenith attenuation can be helpful for describing attenuation on a wide range of satellite links. The actual atmospheric attenuation on a link may be estimated by multiplying the zenith attenuation by the cosecant of the elevation angle. This corresponds closely to the path length through the

atmosphere for elevation greater than 10 degrees. We then ask: What kind of satellites would offer low attenuation for frequencies greater than 40 GHz? Geosynchronous satellites would clearly have problems because of low elevation angles at high latitudes. Fortunately, there are satellite constellations which offer high elevation angles (5) over significant parts of the Earth. The Teledesic and Molniya constellations are two representative choices for high elevation angles and promising performance at millimeter wave frequencies.

The Teledesic- Boeing constellation consists of 24 satellites per plane in 12 planes (288 satellites) at 98 degrees inclination and 1350 km altitude. Snapshots can be taken at any instant of time to find elevation angle to the highest available satellite at every point on Earth. Ninety-six snapshots in time can then be used to generate a probability density function (pdf) for elevation at each Latitude. Fig. 2-1 shows a result with the Northern Hemisphere in the front part of the plot and the Southern Hemisphere at the rear. Elevation angles at high latitudes are seen to be excellent, with elevation often greater than 60 degrees. Elevation at the Equator is not as good. The pdf may be seen to resemble a camel with scoliosis. A contour plot can also be helpful, as Fig. 2-2.

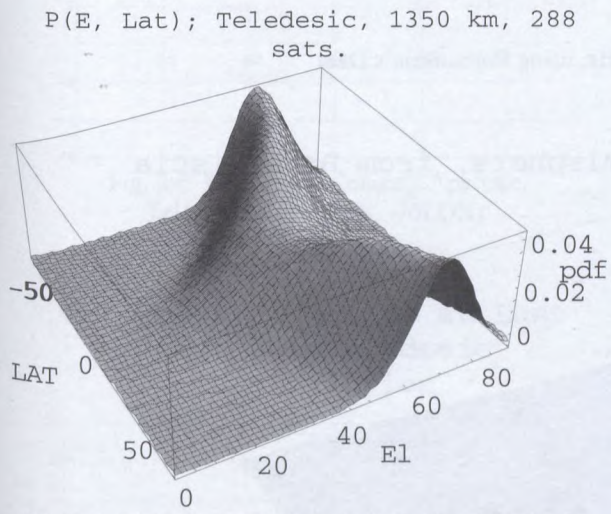


Fig. 2-1 Elevation PDF for Teledesic Constellation vs. Latitude

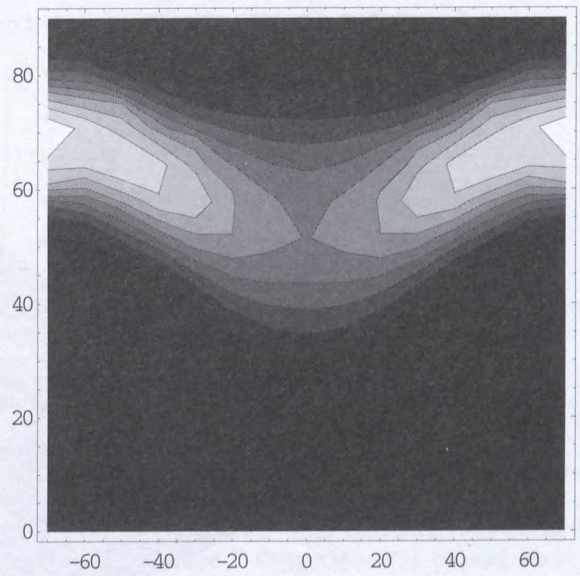


Fig. 2-2 Contour Plot of Teledesic PDF vs. LAT, Elevation

The transformation to the attenuation domain is nonlinear. The integral of cosecant(elevation) is integrated over the pdf to yield Fig. 2-3. This function corresponds to elevation approximately 5- 10 degrees lower than the mean elevation at a given Latitude.

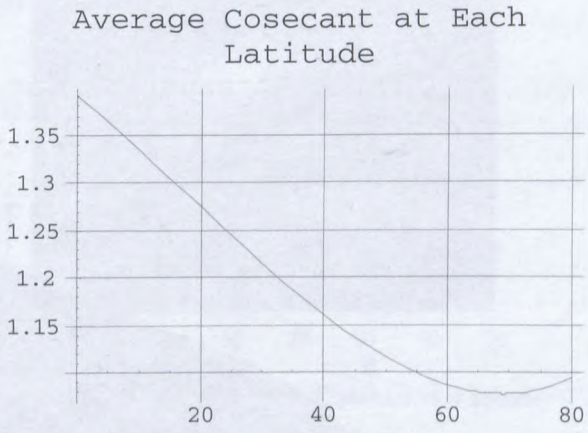


Fig. 2-3 Mean Cosecant (Elevation) vs. LAT

Attenuation at 30 GHz for Teledesic in the Northern Hemisphere may be generated as Fig. 2-4. Again, the attenuation humps are located near Florida, West Africa, and Southeast Asia, from left to right. Attenuation is higher than the earlier 30 GHz zenith attenuation by the factor shown in Fig. 2-3. How would attenuation look for a Teledesic type constellation at 40 GHz? This question can be answered with the function of the Appendix to yield Fig. 2-5.

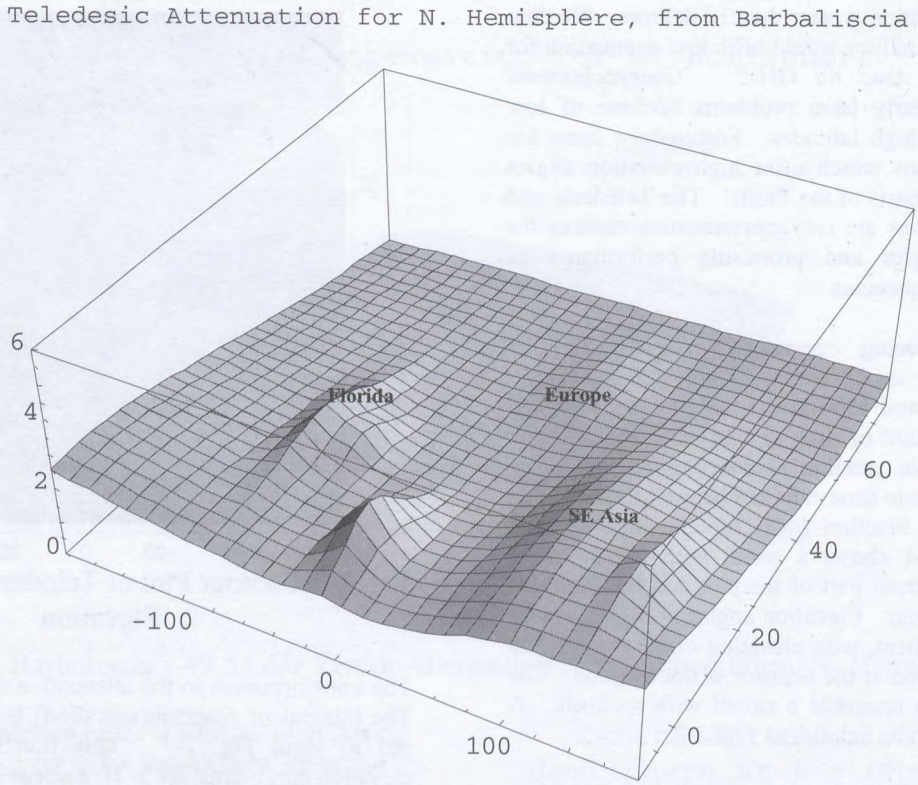


Fig. 2-4 Teledesic Attenuation at 30 GHz, using Barbaliscia's Data

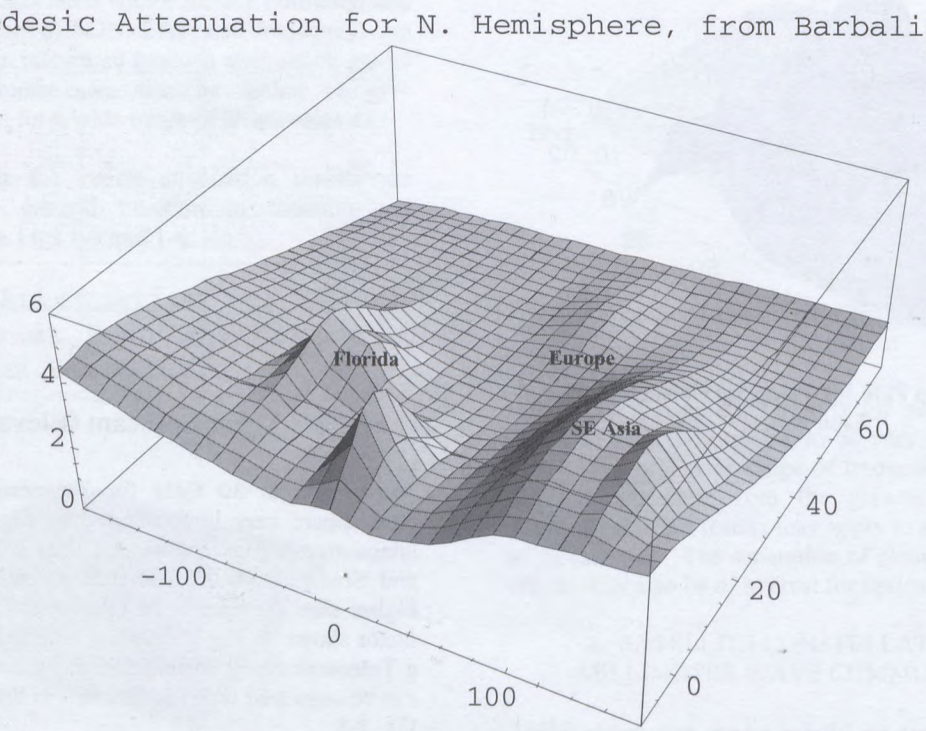


Fig. 2-5 40 GHz Attenuation for a Teledesic Type Constellation, As Implied from Barbaliscia's Data

The relative desirability of a wide range of frequencies can also be examined. Loss at constant antenna aperture can be of interest for a satellite system because system cost is strongly related to antenna size. Loss at constant aperture would be generated from [Attenuation-20Log[f]] and it might be called 'Net Loss'. One might choose a frequency to minimize Net Loss. Fig. 2-6 shows Net Loss vs frequency for a ground station near Cuba with a constellation like Teledesic. An attractive frequency (maybe even optimum) appears near 40 GHz. A ground station near New York (Fig. 2-7) could expect higher frequencies because of lower attenuation. Net Loss at 90 GHz is seen to be almost as attractive as 40 GHz for New York. An area near Hudson Bay (Fig. 2-8) has still lower attenuation and 90 GHz becomes distinctly attractive for these 1% non rainy conditions.

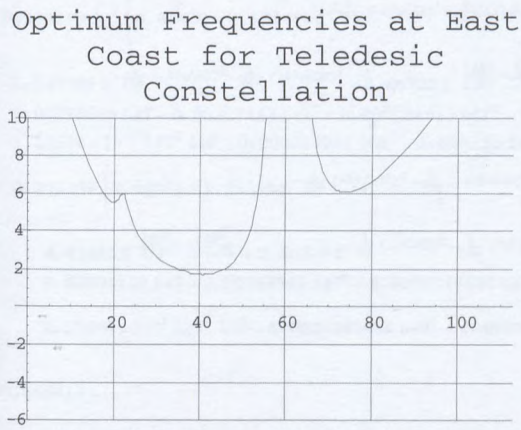


Fig. 2-6 Net Loss at Constant Aperture Teledesic, Cuba (20N, -70LON)

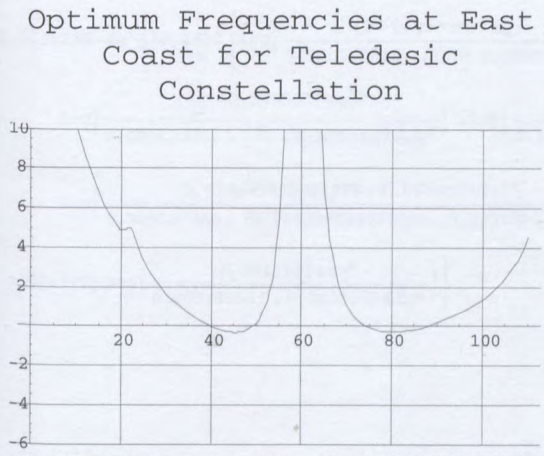


Fig. 2-7 Net Loss near New York (40N,-70LON)

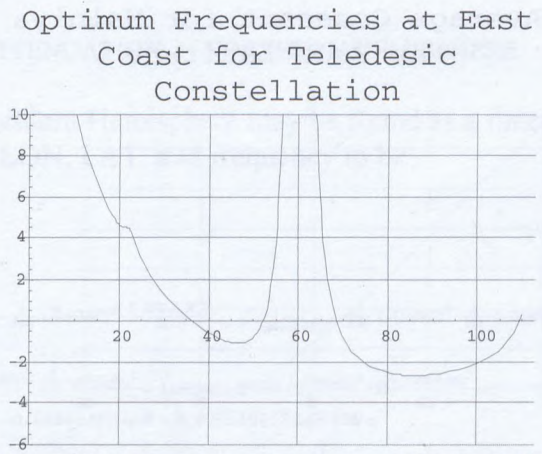


Fig. 2-8 Net Loss Near Hudson Bay (60N,-70LON)

Molniya orbits also offer high elevation angles in the Northern Hemisphere. Only three (3) phased Molniya satellites can be combined with two (2) geosynchronous satellites to deliver an elevation pdf which in some ways is as promising as the Teledesic constellation.

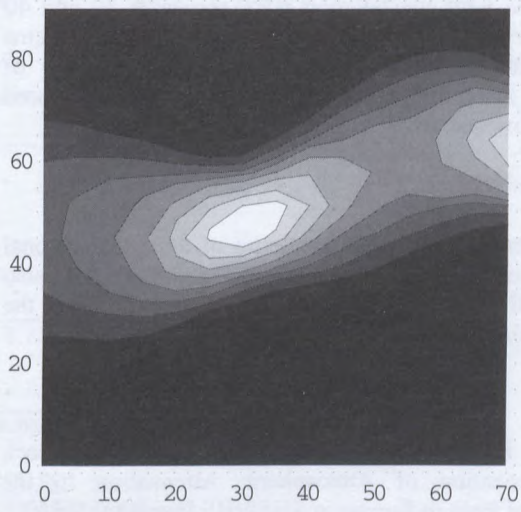


Fig. 2-9 Molniya Elevation Contours vs. Latitude

Elevation contours for the Molniya- Geo system may be found as Fig. 2-9. These elevations group around 50 to 60 degrees, and may be seen to be high throughout the Northern Hemisphere. They may also be seen to be not as favorable as the Teledesic elevation angles of Fig. 2-2. The Cosecant(elevation) may be integrated over the elevation pdf to find a representative path length multiplier for the zenith attenuation, as was done for the Teledesic example. The average Cosecant(elevation) at each latitude may be found as Fig. 2-10. A resultant attenuation for the Molniya system can be found (not shown) to be slightly higher than the Teledesic system of Fig. 2-5.

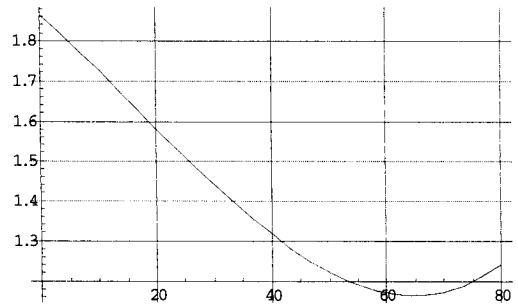


Fig 2-10 Pathlength Multiplier for Molniya, vs. Latitude

Barbaliscia's zenith attenuation plots for 1% non rainy conditions were found to be useful in many ways. We have outlined a method for converting the 49/22 GHz data to a general function for a wide range of frequencies (Appendix). The results were then applied to attenuation for attractive satellite constellations. Attenuation maps for a Teledesic type constellation were shown at 30 and 40 GHz. Attractive frequencies for constant aperture systems were found in the 40-48 GHz region for a range of latitudes. Another favorable region near 90 GHz appeared at New York and Hudson Bay.

Acknowledgement

Ivy Cooper recognized the subtleties of three-dimensional figures and adapted them to the paper. She also recognized the unusual format required for the Mathematica equation of the Appendix.

REFERENCES

[1] F. Barbaliscia, M. Boumis, A. Martellucci, "Characterization of Atmospheric Attenuation in the Absence of Rain in Europe in SHF/EHF Bands for VSAT Satcom Systems Applications," Ka Band Conference, Sorrento, Italy, Sept. 1997.

[2] F. Barbaliscia, M. Boumis, A. Martellucci, "World Wide Maps of Non Rainy Attenuation for Low-Margin Satcom Systems Operating in SHF/EHF Bands, "Ka Band Conference, Sept.1998.

[3] P.Christopher, "Zenith Attenuation in the Northern Hemisphere for 10-49 GHz, from ITALSAT 49/22 GHz Measurements," URSI, Toronto, Aug. 1999.

[4] A. K. Kamal, P. Christopher, "Communication at Millimeter Wavelengths," Proc. International Conference on Communications (ICC), Denver, June 1981.

[5] A. H. Jackson, P. Christopher, "A LEO Concept for Millimeter Wave Communication," Proc. International Mobile Satellite Conference (IMSC), Ottawa, June 1995.

[6] S. Wolfram, Mathematica a System for Doing Mathematics by Computer, 3rd Edition, 1997.

APPENDIX GENERAL EQUATION FOR NON-RAINY ZENITH ATTENUATION IN NORTHERN HEMISPHERE

Mathematica[5] offers powerful methods for multi-dimensional curve fitting. Next, it has full symbolic capabilities for integrated gaseous attenuation. The zenith attenuation for the Northern Hemisphere may be found as a function of LON, LAT, and frequency to be

$$
\begin{aligned}
azen = {} & 0.000408122\, fg^2 \left(7.04001 + 1.31368\, E^{-\frac{1}{200}(-10+LAT)^2 - \frac{1}{250}(-145+LON)^2} + 2.37081\, E^{-\frac{1}{800}(-18+LAT)^2 - \frac{(-10i+LON)^2}{1600}} + 0.766558\, E^{-\frac{1}{200}(-5i+LAT)^2 - \frac{1}{800}(-28+LON)^2} - \right. \\
& 4.47446\, E^{-\frac{LAT^2}{1600} - \frac{LON^2}{160000}} + 2.50808\, E^{-\frac{1}{72}(-i+LAT)^2 - \frac{1}{200}(7+LON)^2} + 1.50452\, E^{-\frac{1}{50}(-20+LAT)^2 - \frac{1}{200}(50+LON)^2} + 1.63483\, E^{-\frac{1}{200}(-20+LAT)^2 - \frac{1}{128}(8i+LON)^2} - \\
& 0.00902149\, LAT - 0.00392949\, LAT^2 + 0.0000616678\, LAT^3 - 2.76814 \times 10^{-7}\, LAT^4 - 0.00201488\, LON - 0.000116053\, LAT\, LON + \\
& \left. 2.19948 \times 10^{-6}\, LAT^2\, LON + 0.0000387636\, LON^2 - 4.49708 \times 10^{-7}\, LAT\, LON^2 + 3.26113 \times 10^{-8}\, LON^3 - 3.08525 \times 10^{-10}\, LON^4 \right) + \\
& 0.00586939\, fg^2 \left(4.44408 + 0.606423\, E^{-\frac{1}{200}(-10+LAT)^2 - \frac{1}{200}(-145+LON)^2} + 1.27363\, E^{-\frac{1}{800}(-18+LAT)^2 - \frac{(-10i+LON)^2}{1600}} + \right. \\
& 0.543769\, E^{-\frac{1}{200}(-5i+LAT)^2 - \frac{1}{800}(-28+LON)^2} - 2.46732\, E^{-\frac{LAT^2}{1600} - \frac{LON^2}{160000}} + 1.24301\, E^{-\frac{1}{72}(-i+LAT)^2 - \frac{1}{200}(7+LON)^2} + 0.985009\, E^{-\frac{1}{200}(-20+LAT)^2 - \frac{1}{128}(8i+LON)^2} - \\
& 0.00295028\, LAT - 0.00227557\, LAT^2 + 0.000032673\, LAT^3 - 1.24191 \times 10^{-7}\, LAT^4 - 0.000813164\, LON - 0.0000591988\, LAT\, LON + \\
& 1.23286 \times 10^{-6}\, LAT^2\, LON + 0.000014057\, LON^2 - 1.09912 \times 10^{-7}\, LAT\, LON^2 + 7.83228 \times 10^{-9}\, LON^3 - 2.29729 \times 10^{-10}\, LON^4 - \\
& 0.201139 \left(7.04001 + 1.31368\, E^{-\frac{1}{200}(-10+LAT)^2 - \frac{1}{200}(-145+LON)^2} + 2.37081\, E^{-\frac{1}{800}(-18+LAT)^2 - \frac{(-10i+LON)^2}{1600}} + 0.766558\, E^{-\frac{1}{200}(-5i+LAT)^2 - \frac{1}{800}(-28+LON)^2} - \right. \\
& 4.47446\, E^{-\frac{LAT^2}{1600} - \frac{LON^2}{160000}} + 2.50808\, E^{-\frac{1}{72}(-i+LAT)^2 - \frac{1}{200}(7+LON)^2} + 1.50452\, E^{-\frac{1}{50}(-20+LAT)^2 - \frac{1}{200}(50+LON)^2} + 1.63483\, E^{-\frac{1}{200}(-20+LAT)^2 - \frac{1}{128}(8i+LON)^2} - \\
& 0.00902149\, LAT - 0.00392949\, LAT^2 + 0.0000616678\, LAT^3 - 2.76814 \times 10^{-7}\, LAT^4 - 0.00201488\, LON - 0.000116053\, LAT\, LON + \\
& \left. \left. 2.19948 \times 10^{-6}\, LAT^2\, LON + 0.0000387636\, LON^2 - 4.49708 \times 10^{-7}\, LAT\, LON^2 + 3.26113 \times 10^{-8}\, LON^3 - 3.08525 \times 10^{-10}\, LON^4 \right) \right) \\
& \left(0.665418 - \right. \\
& 132.118\, (-0.740741 + 0.0333667\, fg)^2 \left(0.999375 - 11.4943 \sqrt{(-0.740741 + 0.0333667\, fg)^2}\, \text{ArcTan}\left[\frac{0.0869456}{\sqrt{(-0.740741 + 0.0333667\, fg)^2}} \right] \right) - \\
& \left. 132.118\, (0.740741 + 0.0333667\, fg)^2 \left(0.999375 - 11.4943 \sqrt{(0.740741 + 0.0333667\, fg)^2}\, \text{ArcTan}\left[\frac{0.0869456}{\sqrt{(0.740741 + 0.0333667\, fg)^2}} \right] \right) \right) - \\
& 0.00392386\, fg^2 \left(18.2482 \left(\text{Log}\left[\frac{(2 - 0.0333667\, fg)^2}{0.000749822 + (2 - 0.0333667\, fg)^2} \right] + \text{Log}\left[\frac{(3.959999999999999 - 0.0333667\, fg)^2}{0.000187456 + (3.959999999999999 - 0.0333667\, fg)^2} \right] + \right. \right. \\
& \text{Log}\left[\frac{\left(\frac{61}{10} - 0.0333667\, fg\right)^2}{0.000187456 + \left(\frac{61}{10} - 0.0333667\, fg\right)^2} \right] + \text{Log}\left[\frac{(2 + 0.0333667\, fg)^2}{0.000749822 + (2 + 0.0333667\, fg)^2} \right] + \\
& \text{Log}\left[\frac{(3.959999999999999 + 0.0333667\, fg)^2}{0.000187456 + (3.959999999999999 + 0.0333667\, fg)^2} \right] + \text{Log}\left[\frac{\left(\frac{61}{10} + 0.0333667\, fg\right)^2}{0.000187456 + \left(\frac{61}{10} + 0.0333667\, fg\right)^2} \right] + \\
& \left. \left. 27.7778\, \text{Log}\left[\frac{0.00111334\, fg^2}{0.000323595 + 0.00111334\, fg^2} \right] \right)
\end{aligned}
$$

Several Applications of Guaranteed Handover (*GH*) Service in Mobile Satellite System (MSS)

Gérard MARAL(1), Joaquín RESTREPO (2), Felipe CABARCAS (3), Santiago JARAMILLO (4), David RIVERA (5)

(1) Ecole Nationale Supérieure des Télécommunications (E.N.S.T.), Site de Toulouse, France
email: maral@supaero.fr

(2),(3),(4),(5) Universidad Pontificia Bolivariana (U.P.B.) and COLCIENCIAS, Colombia
email: (2) restrepo@logos.upb.edu.co (3) fcabarca@egresados.upb.edu.co
(4) sjara@egresados.upb.edu.co (5) drivera@pregrado.upb.edu.co

ABSTRACT

In previous works the *GH* concept has been developed [1], and modelled [2][3]. This concept will guarantee to some users of MSS, that their calls will succeed the handovers. This paper presents several innovative applications based on *GH* concept. Such applications are: *GH option*, offering a *GH* service as a preferred option, but not mandatory. *GH on demand*, asking *GH* users to specify in advance the time duration of *GH* service. *Leased Lines*, allowing the MSS operator to offer a permanent connection. Such options are explained, and analytical models are derived, allowing to dimension traffic performance of MSS. Furthermore, satellite diversity on *GH* service is described.

Keywords: MSS, *GH*, *GH* option, *GH* on demand, Leased Lines, Satellite Diversity

1. INTRODUCTION

Future mobile communication services will be based partly or totally on non-geostationary (*non-GEO*) constellations of satellites, namely Low Earth Orbits, *LEO*, and Medium Earth Orbits, *MEO* [4]. The capacity of the network is increased through frequency reuse not only among satellite footprints, but also within the footprints themselves [5], by dividing the footprint into cells, each one corresponding to a specific beam of the satellite. With *Satellite-Fixed Cell, SFC*, systems, beams maintain a constant geometry with respect to the spacecraft, and the cells on the ground move along with the satellite; they mainly apply for mobile communications, *e.g.*, Mobile Satellite Systems, *MSS* [5]. With *Earth-Fixed Cell, EFC*, systems, the beams steer so as to point towards a given cell on the earth during some time interval; they mainly apply for fixed communications [5][6]. This paper only deals with *SFC*, i.e. *MSS*

Due to the satellite motion with respect to the earth surface, an active user may change beam (*beam handover*), and eventually satellite *(satellite handover)*[5]. Whenever the incoming cell has not any idle channel, a handover failur occurs, entailing a forced termination of the on-going call [7]. This affects both fixed and mobile users [2]. Forced termination can also be caused by propagation impairments. Forced terminations are perceived by the user as frustrating events, and the system designer should aim at achieving a low forced termination probability.

Current terrestrial cellular networks and GEO systems are typically designed to provide a call forced termination probability of about 1% for mobile users [8], and even lower for fixed ones (about 0.01%). In these systems, users at fixed locations (i.e., *fixed users*) do not experience handover failures (because they never suffer handovers), whereas in MSS they could experience handover failures due to the satellite motion. Consequently, it makes sense to envisage that some fixed users will ask for an improved QoS when subscribing to the service. Additionally to this, also some *mobile users* could be interested in a better QoS.

In order to meet the expectation of such potential users, it is necessary to implement a *Guaranteed Handover* (*GH*) service. Users having subscribed to this service will be named *GH users*, and their calls, *GH calls*, while other users will be named *Regular users*, and their calls, *Regular calls*. This *GH* concept has been proposed in [1]. It should reserve *the strictly necessary capacity*, *during the minimum time*, and *as late as possible*, in the cells visited by *GH* user during a call. It requires a known user position at call set-up, and very low mobile velocity, with respect to that of subsatellite point (less than 100 m/s) [9].

One can envisage that MSS will simultaneously service two types of users (i.e., *Regular users* and *GH users*), each category stipulating a specific QoS. Hence, it is important to evaluate the impact on satellite capacity of servicing a mixed population of *GH* and Regular users. With this goal, analytical models for *GH* services have been developed and validated [2][3]. This paper presents several innovative applications based on *GH* concept. Such applications are:

- *GH option*: it offers *GH* service as a preferred option, but not mandatory. Hence, *GH* users would obtain a *GH* service only when capacity in future visited cells is available; otherwise, their calls could be accepted, but as regular ones.
- *GH on demand*: it allows *GH* users to specify in advance the *GH* service duration. In this manner, call is connected as *GH*, and this service is guaranteed until that specified

GH time duration; whenever a call exceeds the specified *GH* service duration, it becomes a regular one.

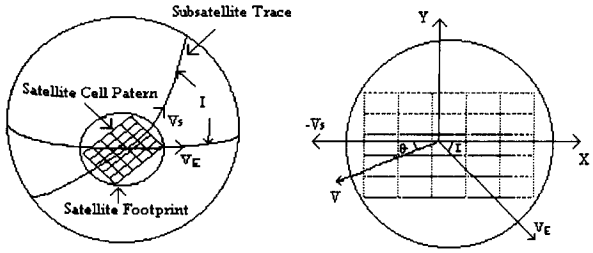
- *Leased Lines*, this service allows MSS operators to offer a permanent connection. Thus, providing adequate fading margins, they could offer communication services with fixed-likely QoS, becoming an attracting alternative for last mile solutions or thin routes, in rural telephony.

Such applications are explained and modelled, allowing to dimension traffic performance of MSS when improving them. Furthermore, satellite diversity on *GH* service is described. The organisation of the paper is as follows: Section 2 revisits the analytical model for basic *GH* service. Section 3 exposes and models *GH* option. Section 4 describes and analyses *GH* on Demand. Section 5 explains and models Leased Lines in MSS. Section 6 describes *GH* service when Satellite Diversity is available. Finally, Section 7 summarises the results and concludes.

2. BASIC *GH* PROCEDURE

This procedure has been initially studied and developed [1]. Latter it has been modelled, firstly by an one dimension model, thus, valid for LEO systems [2], and enlarged to a two dimensions model, taking into account the earth rotation effects, hence valid for any kind of orbit [3]. Here below, this last one is revisited, and taken as a reference, to be modified according to the behaviour of users along the different proposed *GH* applications.

2.1. Coverage geometry and mobility model:



a. Satellite cell pattern b. Satellite footprint excerption

Figure 1: Satellite coverage geometry

Figure 1a displays an earth view of one satellite in the constellation, following an orbit with inclination **I**. Figure 1b shows an enlarged view of satellite footprint, displaying the orientation of the cell pattern with respect to the two involved motions: that of the satellite, V_S, aligned with the X axis, and that of the earth rotation, V_E, at an angle I with respect to axis X, when the satellite is in the equatorial plane. At a given latitude φ, the velocity V_E is found by:

$$V_E = \Omega_E R_E \cos \varphi \qquad (Eq.\ 1)$$

where :

- Ω_E: earth angular velocity [10; p 263]
- R_E : earth radius (R_E = 6378 km) [10; p 257]

The relative user velocity V is the resultant of vectors $-V_S$ and V_E, and calculated by :

$$V = [-V_S + V_E \cos I]\ \mathbf{x} + [-V_E \sin I] \qquad (Eq.\ 2a)$$

And its apparent is inclined at an angle θ, defined by:

$$\text{atan}(\theta) = [-V_E \sin I] \ / \ [-V_S + V_E \cos I] \qquad (Eq.\ 2b)$$

2.2. GH procedure at call set-up time [3]:

Any cell is split into three regions, as shown in Figure 2. Regions II and III are separated by an oblique segment from a specific user path aligned with vector V.

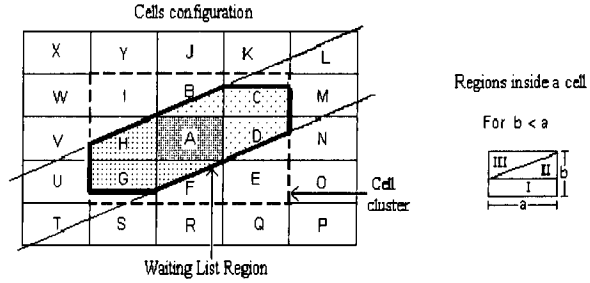


Figure 2: cell configuration

According to [1], Dmax is the longest path in a given cell, and is determined by [3]:

$$\begin{array}{ll} D_{max} = a/\cos\theta. & \text{if} \quad 0 < \theta < \text{atan}(b/a) \\ \quad\ b/\sin\theta & \text{if} \quad \text{atan}(b/a) < \theta < \pi/2 \end{array} \qquad (Eq.\ 3)$$

For any cell targeted for handover, and placed before Dmax at the *GH* call set-up instant, a *GH* user reserves a channel at call set-up. Should no channel be available in it at this time, the call attempt is refused. According to that, when analysing channel occupancy in cell A, eight regions are concerned [3], as shown in Figure 2: A_I, A_{II} , A_{III} , H_I, H_{II}, G_{II}, G_{III}, F_{III}. *GH* users placed into any of them, shall reserve a channel in the cell A, at call set-up, therefore, they must be considered on traffic load of cell A.

2.3. Waiting list for reserved call [3]:

For any cell targeted for handover, and placed beyond Dmax at the *GH* call set-up instant, a *GH* user reserves a channel when arriving at distance Dmax from the boundary of that cell. Should no channel be available in it at this time, the channel reservation request is placed in a waiting list. The requests in the queue are prioritised with respect to any other call. Channels liberated in visited cells are firstly allocated to users, whose requests are standing in the queue, following a FIFO discipline. In order to guarantee a channel to any *GH* handed-over user, it is mandatory that the number of active *GH* users do not exceeds channel cell capacity, C, during the whole residing period of a given *GH* user. This requires knowing the position of *GH* users at every *GH* call set-up (which is actually done in *MSS* [9]). Unlike the Waiting List model, *WL*, of [3], based on an unknown user position, this paper takes into account a known one, and the *WL* can be reduced to:

1- For each next visited cell, find out the probability that the number of future *GH* active user reaches the channel capacity C. This probability is named P_{bwl}

2- The blocking probability of *GH* procedure due to a unsuccessful Waiting List, P_{WL}, corresponds to the product of P_{bwl} by the mean number of visited cells, N_{wl}. As *GH* concept guarantees the success of handoff, $N_{wl} = T_{call}/T_c = \gamma_{GH}$ (see Eq. 7a)

2.4. User mobility model:

The user mobility can be characterised by a parameter γ, which is the inverse of parameter α introduced in [11][12]:

$$\gamma = T_{call} / T_c \qquad (Eq.\ 4)$$

where T_{call} is the mean call duration (an unique common value will be retained, both regular and *GH* users), and T_c is the mean residing time, i.e., the time it takes a user to proceed through a cell, and is given by:

$$T_c = L / V \qquad (Eq.\ 5)$$

where :

- L: mean length along the subsatellite point trace, of different users path in cell (see [3])
- V: the apparent user motion speed (Figure 1b)

2.5. Traffic assumptions [2][13]:

The following assumptions will be considered: Uniform traffic distribution per area unit; Equal number of channels per cell, C; Equal cell size; No priority discipline for handovers, then, equal blocking probability, P_b, in all cells.

2.6. Mobility for Regular users:

Mobility parameter for Regular users, γ_{REG}, which also represents the average number of handover requests of regular users, is defined by:

$$\gamma_{REG} = T_{call} / T_{cREG} \qquad (Eq.\ 6a)$$

T_{cREG} is the mean residing time for Regular users, given by:

$$T_{cREG} = L_{REG} / V \qquad (Eq.\ 6.b)$$

where L_{REG} is the average length of paths followed by Regular users into a cell.

2.7. Mobility for GH users:

The mobility parameter for *GH* users, γ_{GH}, which also represents the average number of handover requests of *GH* users (then, mean number of handovers), is defined by:

$$\gamma_{GH} = T_{call} / T_{cGH} \qquad (Eq.\ 7a)$$

T_{cGH} is the mean occupancy time for *GH* users, given by :

$$T_{cGH} = L_{GH} / V \qquad (Eq.\ 7b)$$

where L_{GH} is the average length of paths followed by *GH* during time they occupy channels in cell A (8 regions of Section 2.2) [3]

2.8. Analytic model for GH service:

Basic assumptions: Traffic model is based on the following assumptions: Sojourn times in cells are constant; Allocation of capacity is according to *FCA* (*Fixed Channel Allocation*); Channel capacity of each cell is a constant, *C*; Traffic model is an *exponential model* based on a Markov chain of the *M/M/C/S* type; All new call arrival and handover arrival processes are independent Poisson ones; Infinite population of users; channel holding times are exponentially distributed; uniform traffic generation (calls attempts/km2) is, corresponding to the peak value.

Handover assumptions: let it define P_{h1} as the probability for the first handover (into the *source cell*), and P_{h2} as the probability for subsequent handover (into the *transit cells*), [11][12]. Subsequently, resulting handovers probabilities for both Regular and *GH* users, are respectively given by :

$$P_{h1REG} = \gamma_{REG}\ (1-e^{-1/\gamma_{REG}}) \quad P_{h2REG} = e^{-1/\gamma_{REG}} \qquad (Eq.\ 8a)$$

$$P_{h1GH} = \gamma_{GH}\ (1-e^{-1/\gamma_{GH}}) \quad P_{h2GH} = e^{-1/\gamma_{GH}} \qquad (Eq.\ 8b)$$

Regular users traffic: The arrival of channel requests from Regular users in a cell is characterised by:

λ_{REGH}:mean call attempts rate of handed-over Regular calls

λ_{REGn}:mean call attempts rate of new Regular calls, given by:

$$\lambda_{REGn} = \lambda_{REG(S)}\ S \qquad (Eq.\ 9)$$

where $\lambda_{REG(S)}$ is the mean call attempts distribution per area unit, for Regular users, and S is the area of the cell

The following equation, which is derived from expressing flux equilibrium [7], determines the handover traffic:

$$\lambda_{REGH} = \lambda_{REGn}\ (1-P_b)\ P_{h1REG}/[(1-(1-P_b)\ P_{h2REG})] \qquad (Eq.\ 10)$$

GH users traffic: Similar to Regular users, *GH* users in a cell are characterised by the following average rates:

λ_{GHh} : mean call attempts rate of handed-over *GH* calls

λ_{GHn} : mean call attempts rate of new *GH* calls

The traffic component of new *GH* calls is given by [3]:

$$\lambda_{GHn} = \lambda_{GH(S)}(1-P_{WL})[S(A_I)(1-P_b)+S(A_{II})(1-P_b)^2+ S(A_{III})(1-P_b)^2+S(H_I)(1-P_b)+S(H_{II})(1-P_b)^2+ S(G_{II})(1-P_b)^2+S(G_{III})(1-P_b)^2+S(F_{III})(1-P_b)^2] \qquad (Eq.\ 11)$$

where:

$S(X_i)$: Area of region i (I, II or III) in cell X (A, H, G or F)

$\lambda_{GH(S)}$: mean call attempt rate per area unit, for *GH* users.

- P_{WL} : blocking probability of *GH* procedure due to a unsuccessful Waiting List, determined by:

$$P_{WL} = \gamma_{GH}\ \text{ErlangB}\ [\ (\lambda_{GH(A)}\ T_{call})\ ;\ C] \qquad (Eq.\ 12)$$

$\lambda_{GH(A)}$: mean call attempts from *GH* users placed into cell A only, and finding a channel into it. It can be deducted from Eq. 11, as follows:

$$\lambda_{GH(A)} = \lambda_{GH(S)}\ (1-P_b)\ [S(A_I)\ (1-P_b) + S(A_{II})\ (1-P_b) + S(A_{III})\ (1-P_b)^2] \qquad (Eq.\ 13)$$

As for Regular users, the *GH* handover traffic is derived from expressing flux equilibrium [7], and is given by :

$$\lambda_{GHh} = \lambda_{GHn}(1-P_b)\ \gamma_{GH} \qquad (Eq.\ 14)$$

Channel holding time in a cell: The mean channel holding time in a cell, $1/\mu_{Tcell}$ [12] [13], is given by:

$$1/\mu_{Tcell} = P_1E[T_{H1}] + P_2E[T_{H2}] + P_3E[T_{H3}] + P_4E[T_{H4}] \qquad (Eq.\ 15)$$

where P_1, P_2, P_3, and P_4 are the probabilities that a channel be occupied/locked by a new Regular call, by a handed-over Regular call, by a new *GH* call, and by a handed-over *GH* call respectively. Similarly, $E[T_{H1}]$, $E[T_{H2}]$, $E[T_{H3}]$, and $E[T_{H4}]$:are the mean channel holding time for new Regular calls, handed over Regular calls, new *GH* calls, and handed over *GH* calls rsepectively. P_1, P_2, P_3, and P_4 are given by:

$$P_1 = \lambda_{REGn}(1-P_b)/\Lambda \qquad P_2 = \lambda_{REGH}(1-P_b)/\Lambda$$
$$P_3 = \lambda_{GHn}(1-P_b)/\Lambda \qquad P_4 = \lambda_{GHh}/\Lambda \qquad (Eq.\ 16)$$

Λ is the mean rate of the total carried traffic, given by:

$$\Lambda = \lambda_{REGn}(1-P_b) + \lambda_{REGH}(1-P_b) + \lambda_{GHn}(1-P_b) + \lambda_{GHh} \qquad (Eq.\ 17)$$

The mean holding times are given by:

$$E[T_{H1}] = T_{callREG}(1-P_{h1REG}) \qquad E[T_{H2}] = T_{callREG}(1-P_{h2REG})$$

$$E[T_{H3}] = T_{callGH}(1-P_{h1GH}) \qquad E[T_{H4}] = T_{callGH}(1-P_{h2GH}) \qquad (Eq.18)$$

Total traffic load in a cell: The total mean call rate in a cell, λ_{Tcell}, is defined by:

$$\lambda_{Tcell} = \lambda_{REGn} + \lambda_{REGH} + \lambda_{GHn} + \lambda_{GHh} \qquad (Eq.\ 19)$$

The total traffic intensity per cell, ρ_{Tcell}, is given by:

$$\rho_{Tcell} = \lambda_{Tcell} / \mu_{Tcell} \qquad (Eq.\ 20)$$

Input parameters of GH traffic model: An input variable is the traffic load due to new calls *in a given cell* (both from *GH* and Regular users), ρ_{tncell}. As T_{call} is unique, the total new call attempt rate per cell, λ_{tncell}, is given by:

$$\lambda_{tncell} = \rho_{tncell} / T_{call} \qquad (Eq.\ 21)$$

In this manner, we have that:

$$\lambda_{tncell} = \lambda_{REGncell} + \lambda_{GHncell} \qquad (Eq.\ 22)$$

One introduce a new parameter, k_{GH}, defined as:

$$k_{GH} = \lambda_{GHncell} / \lambda_{tncell} \qquad (Eq.\ 23)$$

Thus, the call attempt rates are given by :

$$\lambda_{REGncell} = \lambda_{tncell}(1-k_{GH}) \qquad \lambda_{GHncell} = \lambda_{tncell}\ k_{GH} \qquad (Eq.\ 24)$$

Call attempts per area unit (Eqs. 10 and 12), is obtained by dividing the above variables by the cell surface:

$$\lambda_{REG(S)} = \lambda_{REGncell}/S \qquad \lambda_{GH(S)} = \lambda_{GHncell} / S \qquad (Eq.\ 25)$$

The total new calls traffic load per cell, ρ_{tncell}, and the fraction of *GH* calls, k_{GH}, will be considered as the input variables for the comparative analyses.

Markov's chain: Now, cells are modelled as *M/M/C/S* queuing systems with non-homogeneous arrival rates and S = 2C, since the number of *GH* users handed-over calls in the queue cannot exceed C. A Markov's chain based on this model is shown in Figure 3. In this Figure, μ and μ_{call} correspond to μ_{Tcell} (as in Eq. 15) and the inverse of T_{call} ($\mu_{call} = 1 / T_{call}$) respectively.

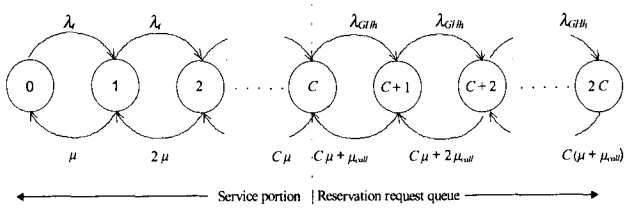


Fig. 3: Markov chain for a GH service with FCA.

This Markov's chain has been solved in [14][15] [16] [17]:

$$\frac{P_n}{P_0} = \begin{cases} \dfrac{\lambda_t^n}{n!\mu^n}, & 1 \le n \le C-1 \\ \dfrac{\lambda_t^C\ \lambda_{GHh}^{n-C}}{C!\ \mu^C \prod_{j=1}^{n-C}[C\mu + j\mu_{call}]}, & C \le n \le 2C \end{cases} \qquad (Eq.\ 26)$$

where the idle system probability, P_0, is given by:

$$P_0 = \left\{ \sum_{n=0}^{C-1} \frac{\lambda_t^n}{n!\ \mu^n} + \sum_{n=C}^{2C} \frac{\lambda_t^C\ \lambda_{GHh}^{n-C}}{C!\ \mu^C \prod_{j=1}^{n-C}[C\mu + j\mu_{call}]} \right\}^{-1} \qquad (Eq.\ .27)$$

Finally, P_b, is obtained as follows:

$$P_b = \sum_{n=C}^{2C} P_n \qquad (Eq.\ 28)$$

2.9. QoS parameters:

Through the analytical evaluation above described, one can determine the blocking probability for a given cell, Pb. From this value, it can be deduced the different parameters of QoS, both for Regular and *GH* users, defined as follows:

- *Call blocking probability, P_b*: the probability that a new call attempt is rejected at call set-up time.
- *Call dropping probability, P_{drop}*: the probability that a call attempt is lost due to an unsuccessful handover.
- *Non successful ending call probability P_{ns}*: the probability that a call attempt is not successfully ended, either due to *Call blocking* or due to *Call dropping.*

QoS for Regular users: defined by the following set:

$$P_{bREG} = P_b \qquad (Eq.29a)$$
$$P_{dropREG} = P_bP_{h1REG} / [\,1-(1-P_b)\,P_{h2REG}\,] \qquad (Eq.29b)$$
$$P_{nsREG} = P_b + (1-P_b)\,P_{dropREG} \qquad (Eq.29c)$$

QoS for GH users: defined by the following set:

$$P_{bGH} = 1-((1- P_{bWL}) / S) [S(A_I) (1-P_b)^2 + S(A_{II}) (1-P_b)^3 + S(A_{III}) (1-P_b)^3)]) \quad (Eq.\ 30a)$$

$$P_{dropGH} = 0 \quad (GH \text{ concept}) \quad (Eq.\ 30b)$$

$$P_{nsGH} = P_{bGH} \quad (Eq.\ 30c)$$

Extensive evaluation of this model, applied to the existing *MSS* [18][19][20], can be found in [3], concluding that there is not relevant difference among them. In this paper, in order to compare the different *GH* applications, only one of them is retained, whit the following parameters:

- Latitude: 0°
- Orbit Inclination (at RAAN): 86°
- Satellite Altitude: 780 Km
- Cell size: 500x500 Km
- Mean call duration: 3 min
- Number of Channels per Cell: 10

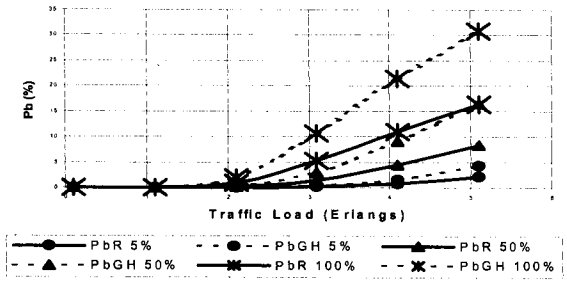


Figure 4: Standard GH service

Figure 4 displays results obtained for the above case. *X*-axis shows different traffic loads, and *Y*-axis shows the resulting P_b both for *GH* users (dotted lines) and Regular ones (continuos lines). Curves are displayed for different portion of *GH* users (5, 50 and 100%). It can be deducted that the P_b for *GH* users is roughly twice that of Regular ones, which is explained by the requirement of 2 channels at set-up for most of *GH* users. In order to study the effect of *GH* applications on the system performance, the next Sections will consider that thus *GH* service is the unique one, *i.e.*, traffic models do not mix different *GH* services. Furthermore, the scenario above described will be retained as a reference, in such a manner that its associated Figure 4 permits to establish relative comparisons

3. *GH* OPTION

In this *GH* application, *GH* users subscribe a *GH* service as a prime option. Hence, they firstly try to be connected as *GH* users; whether their attempt fails, they *always* accept to be treated as Regular ones. *GH* service should be refused by a different reason than lack of capacity into the source cell, because in this event, they can not be connected, regardless of being *GH* or Regular user.

3.1. GH Option procedure:

- For all of users (both *GH* and regular), if there is no channel into the source cell, the call will be rejected.
- For Regular users the set-up procedure remains the same: they only require an idle channel into the source cell.
- For *GH* option users the system will try firstly to connect them as *GH* standard, by guaranteeing the success of channel reservation in the next visited cells, *i.e.*, set-up procedure (cells before Dmax) and waiting list procedure (cells beyond Dmax). Whether such a procedure fails, a *GH* Option user accepts a regular *status* instead.

3.2. Analytical model:

GH users becoming Regular ones increase the traffic load of these last ones, therefore, they must be considered in the term expressing new call attempts from regular users λ_{REGn}. Hence this term shall be replaced by:

$$\lambda_{REGn(op)} = \lambda_{REGn} + \lambda_{GH \to REG} \quad (Eq.\ 31)$$

where:

$\lambda_{REGn(op)}$: modified new call attempts rate of Regular users (be actual regular users, or *GH* transformed ones)

λ_{REGn} : new call attempts rate of actual Regular (found through Equation 10, in Section 2)

$\lambda_{GH \to REG}$: new call attempts rate of *GH* users, becoming Regular ones. This term is determined by:

$$\lambda_{GH \to REG} = \lambda_{GH(N)} P_{(GH \to REG)} K_{GH \to REG} \quad (Eq.\ 32)$$

where

$\lambda_{GH(N)}$: new call attempts rate for *GH* users placed into source cell, found by:

$$\lambda_{GH(N):} = \lambda_{GHS} S \quad (Eq.\ 33)$$

with

λ_{GHS} : density of new call attempts rate for *GH* users (as in Section 2, Eq. 25)

S : whole area of one cell. ($S = A_I + A_{II} + A_{III}$)

$P_{(GH \to REG)}$: transformation probability of a *GH option* call into a regular one, found by:

$$P_{(GH \to REG)} = [1-[(1-P_{WL})[A_I(1-P_b)+2A_{II}(1-P_b)^2](1/S)]] \quad (Eq.\ 34)$$

P_{WL} : blocking probability of the Waiting List.

$K_{GH \to REG}$:Portion of *GH* users accepting *GH* option service

When modelling this *GH* application since the standard *GH* model, as seen in Section 2, all the Equations beyond Eq 11, and including λ_{REGn}, should apply $\lambda_{REGn(op)}$ instead. Additionally, the *GH* traffic load is not affected by this analysis, then, it remains as described in Section 2. On the other hand, the blocking probability for *GH* users (as in Eq. 30a), should be redefined for *GH Option*, as follows:

$$(1-P_{bGHopt}) = (1-P_{bGH}) + P_{(GH \to R)} K_{GH \to REG} (1-P_b) \quad (Eq.\ 35)$$

where:

$(1-P_{bGHopt})$: Total connection probability for *GH* option users (be as *GH*, be as Regular)

$(1-P_{bGH})$: Probability of *GH* call attempts to be granted as *GH* users

$P_{(GH \to R)}$: Probability of *GH* call attempts to be transformed into Regular ones.

Based on Eq. 34, this Equation can be rewritten as:

$$(1-P_{bGHopt}) = (1-P_b) [1 - P_{(GH \to R)}(1-K_{GH \to REG})] \quad (Eq.\ 36)$$

It can be shown that when K_{GH} = 100%, one matches the Blocking probabilities of *GH* and Regular users. Also, when K_{GH}= 100%, one obtains the basic *GH* value, P_{bGH}

3.3. System Performance:

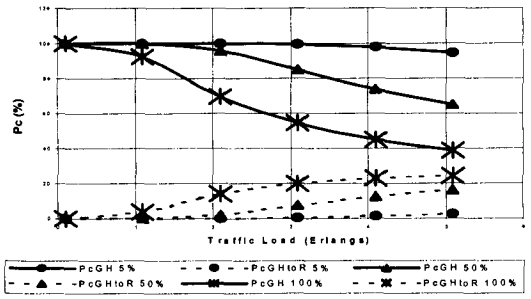


Figure 5: Pc vs. Traffic Load in GH Option

The Figure 5 has been obtained from the *GH* option model, for K_{GH} =100%. It shows the succeeding probability of *GH* users to a *GH* service (continuos lines), or being connected through a Regular *status* (dotted lines); the total probability of connection will be the sum of those. Curves are drawn for different portion of *GH* users. An in-deep comparison between Figure 4 (Basic *GH* service) and Figure 5 (*GH* Option), leads to the following statements:

- For high traffic loads and large k_{GH}, the blocking probability of *GH* option is greater than that of basic *GH*.
- For low traffic loads and small k_{GH}, the blocking probability of *GH* option, is lower than that of basic *GH*.

As matter of example, in the first case, the P_{bGH} is very high, and rejected calls being transformed into regular users, create a relevant traffic load, e.g., for k_{GH}=100% and Traffic load of 5 Erlg., P_{bGH} (as in *GH* basic case) is about 30%, thus, it generates a regular traffic load of about 1.5 Erlg, and it entails a modified condition, with about 6.5 Erlg and k_{GH} ~ 70%. It can be concluded that in the first case, the *GH* option is not recommended, because it leads to a higher Pb, affecting both *GH* and regular users. On the other hand, in the second case, the *GH* option allows a reduction of P_{bGH}, with a small augmentation of P_{bR}.

4. *GH* on DEMAND

In this application is considered that the *GH* users request the *GH* service during a given time, and after that time they become Regular Users.

4.1. GH on Demand Procedure:

- The set-up procedure remains the same, both for Regular and *GH* users.
- When a *GH* user try to connect, he must specify the time duration he ask for a *GH* service, T_{GH}.
- Once connected, it is guaranteed the success of all handovers occurring before T_{GH}. When the call duration exceeds T_{GH} , it becomes regular and might suffer a handover failure. Let be noted that T_{GH} should be equal to or greater than T_C. In fact, if T_{GH} is lower than the time to the first handover, this service has not sense

4.2. Analytical model:

Once *GH* calls are transformed in regular ones, they will be seen by the system as new regular connected calls, thus, they are able to be part of regular handover rate. This contribution can be determined by:

$$\lambda_{GHR} = \lambda_{GHh}\, Ph_{GHR} \qquad (Eq.\ 37)$$

where:

λ_{GHR} : Connected *GH* call rate becoming Regular ones.

Ph_{GHR} : Probability that a *GH* user becomes Regular. Since call duration follows a negative exponential distribution, this probability can be calculated from:

$$Ph_{GHR} = \int_{T_{GH}}^{\infty} \frac{1}{T_{call}} e^{\frac{-T}{T_{Call}}} dT = \exp^{(-T_{GH}/T_{call})} \qquad (Eq.\ 38)$$

Where T_{GH} is the average time that the *GH* users request the service before becoming regular users.

In this manner, Regular handover traffic is increased by λ_{GHR}. Hence, based on the flux equilibrium for Regular Users [7], Equation 10 becomes:

$$\lambda_{REGH}=[\lambda_{REGn}(1-P_b)+\lambda_{GHR}](1-Ph_{1REG})/[1-(1-P_b)P_{h2REG}] \qquad (Eq.\ 39)$$

Similarly, the *GH* traffic is reduced by λ_{GHR}, and, Equation 14 becomes:

$$\lambda_{GHh} = [\lambda_{GHn}(1-P_b)\,P_{h1GH}] / [\,1-(P_{h2GH} - Ph_{GHR})] \qquad (Eq.\ 40)$$

Comparing Eqs. 8b, 38, and 40, it can be deducted that $T_{GHR} > T_c$ (*i.e.*, *GH* duration should be larger than cell residing time, as assumed before).

4.3. System Performance:

The Figure 6 has been obtained from the *GH* on Demand model, for T_{GH} = T_{call} =3 min. It shows the blocking probability of both *GH* (dotted lines) and Regular ones (continuos lines). Curves are drawn for different portion of *GH* users.

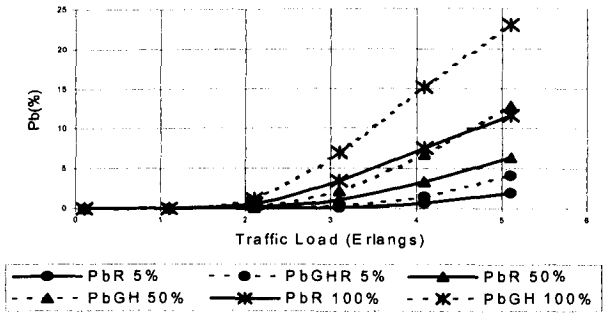


Figure 6: P_b vs. Traffic load in GH on Demand

When comparing this Figure with that of Basic *GH* service (Figure 4), it is shown that *GH* on Demand always leads to a lower P_b, both for *GH* and regular users. It is explained by a reduction of the mean number of reserved channels from handed over *GH* calls, which increase the mean channel availability. Furthermore, it is also interesting to look at the performance effects due to T_{GH} . As a matter of example, one has retained a traffic load of 3.1 Erlg., and

K_{GH} = 50 %. Then, a curve is obtained showing the Pb vs. T_{GH}, as in Figure 7.

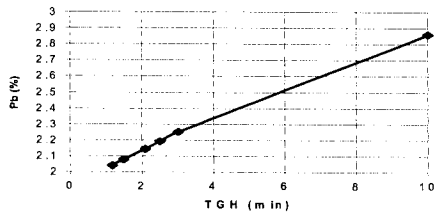


Figure 7: P_b vs TGH changes (3 Erlg, k_{GH}=50%)

Figure 7 shows that when T_{GH} grows, P_b improvement reduces, trending to that of basic *GH* service. This has been verified for other values of traffic load and k_{GH}. Hence, it can be considered a main characteristic of this application, resulting from less and less *GH* calls changing from *GH* to regular *status*.

5. LEASED LINES

Leased lines could be consider as already connected *GH* calls, never ending their calls, thus in constant handover

5.1. Leased Lines Procedure:

For such an application, P_{bGH} is not considered, because of the user can wait until the arrival of the appropriate conditions to be connected. After that connection, he is seen by the system as a permanent handed over call, *i.e.*, a connected call of infinite duration.

5.2. Analytical Model:

Whether mean call duration becomes infinite, it leads to a mean channel occupancy time equal to the residing time. As the set-up procedure is not considered, it can be assumed that λ_{GHn}=0, and P_{bGH}=0. Additionally, leased lines generate a *GH* handed over traffic, $\lambda_{GHh}=\lambda_{LL}$, given by:

$$\lambda_{LL} = \frac{n_{LL}(s) \cdot S}{T_C} \qquad (Eq.\ 41)$$

where $n_{LL}(s)$ is the Leased lines density (lines / Km^2)

Based on *GH* procedure from Section 2, the mobility parameter (Eq. 7a) becomes infinite, thus, in Eq. 8b hand over probabilities $P_{h1GH}=P_{h2GH}=1$. Furthermore, in Eq. 18, the mean occupancy time for *GH* handed over calls (i.e., Leased Lines) can be deducted from:

$$E(T_{LL}) = \lim_{T_{Call}\to\infty} T_{Call}\left(1-e^{-\frac{Tc}{T_{Call}}}\right) = T_C \qquad (Eq.\ 42)$$

where T_{LL} is the mean occupancy time of a leased lines in a cell. As they never end, they pass through all visited cells, and mean occupancy time is equal to the cell residing one, as demonstrated in Eq. 42. Furthermore, Eqs. 17, and 19, can be replaced by Eqs. 43 and 44 respectively, as follows:

$$\Lambda = \lambda_{Rt}(1-P_b) + \lambda_{LL} \qquad (Eq.\ 43)$$

$$\lambda_{TCell} = \lambda_{Rn} + \lambda_{Rh} + \lambda_{LL} \qquad (Eq.\ 44)$$

Figure 8 displays the Blocking Probability for regular users, resulting from the above model, for 3 different MSS: Iridium-Likely (as In Section 2), Globalstar-likely (I=52°, h=1440 km) and ICO–likely (I=45°, h=10340 km). Curves are displayed for 3 different percentage of Leased Lines: 25, 50 and 100%.

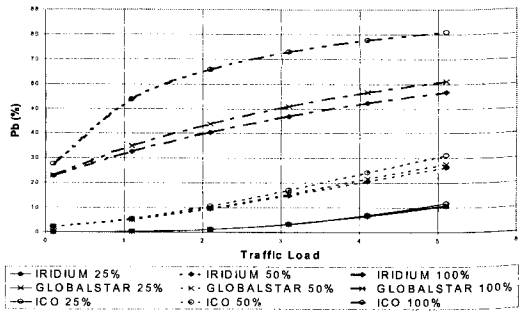


Figure 8: P_{bR} vs. Traffic Load for Leased Lines

When comparing this Figure with that of Basic *GH* service (Figure 5), it is shown that in this application, Blocking Probability for regular users is a lot larger than that of *GH* basic Service. It is also observed that for LL<25%, the blocking probability for the regular users does not depend on the constellation type, and it is relatively small (lower than 10%). For LL=50%, differences among constellation increase, in such a manner that the higher the constellation altitude, the greater the resulting Pb. This divergence between constellation is even more notorious for LL=100%. This is due to the channel liberation rate, which is smaller for higher constellations. On the other hand, forced termination will be larger in lower constellations, due to a larger handed over traffic.

It can be also noted that, in spite of having LL~100%, Pb is lower than 100%. It is due to an uniform distribution, which makes that not all Leased Lines enter in at the same instant. Subsidiary, it can be verified that, if the positions of Leased Lines are very close, Pb trends rapidly to 100%. Thus, MSS operator must be very careful about relative distance among Leased Lines, to avoid such a situation.

6. SATELLITE DIVERSITY IN *GH*

In this application, fading impairments might be overcome by the bias of satellite diversity, *i.e.*, handing over the *GH* call to another satellite in visibility. Two approach are being currently studied:

6.1. Stand-by Diversity:

In this case, *GH* users are connected uniquely to one satellite, and they appeal for a second one, only in the even of a deep fading. In this event, they should repeat the set-up procedure on the new satellite. In such a case, *GH* forced termination probability is defined by:

$$(1-P_{drop\text{-}SD})=(1-P_{f(S1)})+P_{f(S1)}P_{V(S2)}[1-P_{f(S1|S2)}](1-P_{bGH(S2)}) \qquad (Eq.\ 45)$$

where:
$P_{drop\text{-}SD}$: call drop probability with satellite diversity
$P_{f(S1)}$: fading probability in satellite 1
$P_{V(S2)}$: probability of having a second satellite in visibility
$P_{f(S1|S2)}$: probability of simultaneous fading in satellites 1 and 2
$P_{bGH(S2)}$: blocking probability for *GH* users in satellite 2

The *GH* call attempts rate in a given satellite will be:

$$\lambda_{GHT(S1)} = \lambda_{GHI(S1)}(1 - P_{f(S1)}) + \lambda_{GHC(S2)} P_{f(S2)})[1 - P_{f(S1|S2)}] \quad (Eq.\ 46)$$

where:
$\lambda_{GHT(S1)}$: *GH* call attempts rate, seen by satellite 1
$\lambda_{GHI(S1)}$: *GH* call attempts rate from users into satellite 1
$\lambda_{GH(S2)}$:*GH* connected call rate from users into satellite 2
$P_{f(S2)}$: fading probability in satellite 2

6.1. Permanent Diversity:
In this case, *GH* users are connected simultaneously to both satellites; in the event of deep fading, communication is kept through the remaining link. Now, *GH* forced termination probability is defined by:

$$(1 - P_{drop\text{-}SD}) = (1 - P_{f(S1)}) + P_{f(S1)} P_{V(S2)}[1 - P_{f(S1|S2)}] \quad (Eq.\ 47)$$

The *GH* call attempts rate in a given satellite will be:

$$\lambda_{GHT(S1)} = \lambda_{GHI(S1)} + \lambda_{GHI(S2)} \quad (Eq.\ 48)$$

Note that P_{drop} is reduced, but P_{bGH} at set-up will increase, due both to a doubled traffic load, and a duplicated *GH* procedure (2 simultaneous *GH* connection are required). Further studies will analyse these options, which must be compared with required fading margins, in the absence of diversity.

7. CONCLUSIONS

In this paper, the basic *GH* service has been revisited. Some new *GH* applications derived from it has been exposed and modelled. It has been shown that *GH* Option becomes an interesting application only for low traffic loads and low K_{GH} percentages. *GH* on Demand seems to be a very attractive application of *GH*, because it always increases the system performance. Leased Lines provide a permanent connection in MSS, but reduces strongly the performance of regular users.

Furtherworks : An in-deep study about Satellite diversity in *GH* service, compared with fading margins would be realised. Also, all these analytical models will be validated through a simulation campaign.

REFERENCES

[1] J. Restrepo, G. Maral, «Guaranteed Handover (*GH*) Service in a non-GEO Constellation with "Satellite-Fixed Cell" (SFC) Systems», *Proceedings of the NASA-JPL International Mobile Satellite Conference, Satcoms,* IMSC '97, Pasadena, 16-18 June 1997.

[2] G. Maral, J. Restrepo, E. Del Re, R. Fantacci, G. Giambene, «Performance analysis for a guaranteed handover service in a LEO constellation with a 'SFC' system», *IEEE Transactions on vehicular technology.,* Vo. 47, No. 4, November 1998.

[3] J. Restrepo, G. Maral et al, " Extended Analytical Model For Calculating The Quality Of Service (Qos) In A Mobile Satellite System (Mss) With a Guaranteed Handover (Gh) Service" *Procceding of the EMPS98* Venice, Italy; November 1998

[4] «Green paper on a common approach in the field of mobile and personal communications in the European Union». *European Commission*, 1994.

[5] J. Restrepo and G. Maral, «Coverage Concepts for Satellite Constellations Providing Communications Services to Fixed and Mobile Users», *Space Communications*, Vol. 13, No 2, pp. 145-157, 1995.

[6] J. Restrepo and G. Maral, «Constellation Sizing for Non-GEO 'Earth-Fixed Cell' Satellite Systems», *Proceedings of the AIAA 16th International Communications Satellite Systems Conference and Exhibit*, Washington D.C., U.S.A., Feb. 25-29, pp. 768-778, 1996.

[7] S. Rappaport, «The Multiple-Call Hand-off Problem in High-Capacity Cellular Communications Systems», *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 3, pp. 546-557, August 1991.

[8] ITU-T Recommendation E.771, «Network Grade of Service Parameters and Target Values for Circuit-switched Land Mobile Services». *Blue Book*, 1995.

[9] W. Zhao, R. Tafazolli, B. G. Evans, "A UT Positioning Approach for Dynamic Satellite Constellations," *Proc. of 4th IMSC'95, pp. 251-258*, Ottawa, Canada, June 6 - 8, 1995

[10] G. Maral, M. Bousquet, '*Satellite Communications Systems*' Wiley, 3d edition , 1998.

[11] E. Del Re, R. Fantacci, G. Giambene, «Performance Analysis of Dynamic Channel Allocation Technique for Satellite Mobile Cellular Networks», *International Journal of Satellite Communications*, Vol. 12, No. 1, pp. 25-32, 1994.

[12] E. Del Re, R. Fantacci, G. Giambene, «Efficient Dynamic Channel Allocation Techniques with Handover Queuing for Mobile Satellite Networks», *IEEE Journal on Selected Areas in Comm.*, Vol. 13, No. 2, pp. 397 - 405, February 1995.

[13] F. Dosiere, T. Zein, G. Maral., «A Model for the Handover Traffic in Low Earth-Orbiting (LEO) Satellite Networks for Personal Communications», *International Journal of Satellite Communications*, Vol. 11, No. 6, pp. 145-149, 1993.

[14] Duk Kyung Kim, Dan Keung Sung, «Handoff/Resource Managements Based on PCVs and SVCs in Broadband Personal Communication Networks», *Proceedings of IEEE GLOBECOM '96*, London, November 18 - 22, 1996.

[15] Yi-Bing Lin, Li-Fang Chang, A. Noerpel, «Modeling Hierarchical Microcell / Macrocell PCS Architecture», *IEEE Proceedings of ICC'95*, pp. 405-409, Seattle, June 1995.

[16] K. W. Ross. *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer-Verlag, 1995.

[17] D. Hong, S. S. Rappaport, «Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures», *IEEE Trans. on Veh. Tech.*, Vol. VT-35, No. 3, pp. 77 - 92, August 1986.

[18] Web page with address: http://www.iridium.com.

[19] Web page with address: http://www.globalstar.com.

[20] Web page with address: http://www.ico.com

Stochastic Optimization of Satellite Frequency Assignment

Axel Jahn
German Aerospace Center (DLR)
Institute for Communications Technology
P.O. Box 11 16, D-82230 Wessling, Germany
Email: Axel.Jahn@dlr.de

ABSTRACT

This paper presents stochastic optimization methods for frequency assignments in land mobile satellite (LMS) systems. Genetic algorithms (GA) and simulated annealing (SA) can be used for channel assignment in satellite multibeam antennas. The principles of stochastic optimization is described, the performance of the assignment is evaluated as fitness or cost criteria based on the co-channel interference. Several operators for selection, mutation and crossing are investigated. Finally, results show the applicability and power of the algorithms. The results are compared against regular reuse patterns for satellite antennas with equal beamwidth and equal cell area.

I. INTRODUCTION

Frequency assignment plays an important role in the design of satellite communication networks [1]. It determines the spectrum efficiency as well as the quality of service of the radio transmission link. Thus, efficient planning and performance evaluation tools for the channel assignment are prerequisite. Several approaches are known for channel assignment problems, for instance graph colouring [3], algorithmic approaches [4] and stochastic optimization methods [5], [6], [7]. If co-channel and adjacent channel interference between overlapping satellite cells and overlapping footprints is addressed, the frequency assignment problem is highly non-linear. Stochastic optimization promises an efficient and powerful approximation method whenever highly non-linear equations systems must be fitted to several constraints. Stochastic optimization methods are based on evolution processes in nature (genetic algorithms, GA) and on processes with minimal entropy (simulated annealing, SA). The applicability of GA and SA has been studied for many areas in communications. Several publications have already adopted GA [5], [7] and SA [6], [7] for channel assignment in cellular terrestrial networks.

II. STOCHASTIC OPTIMIZATION METHODS

A. Genetic Algorithms

The application of GA for engineering aspects was founded by the work of Holland [8], [9]. He coded complex structures by simple binary sequences that can represent all states of the system under investigation. Improvements of a performance measure can be achieved through modifications of the bit structures if the modification process is subject to a control process. Different bit sequences represent individuals of a population with different characteristics. Based on Darwin's evolution theory, the population is developing over several generations. The genes of the fittest individuals are selected for parents of the next generation.

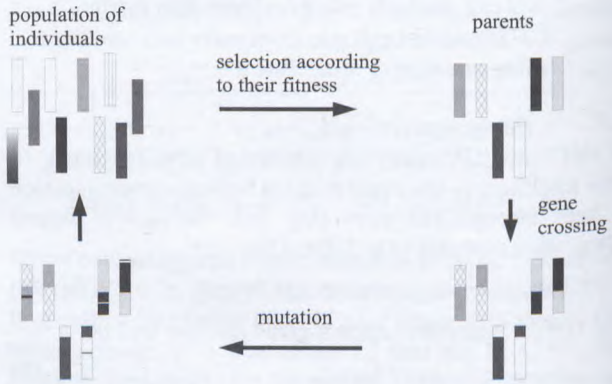


Fig. 1. Function of genetic algorithms

The function principles of genetic algorithms are illustrated in Fig. 1. GA start from a random initial population. The fitness of each individual of the population is assessed through an appropriate evaluation function. Parents are now selected according to their fitness. The genes of two parents are taken for reproduction. For this, the genes of the parents are crossed in pairs. Additionally, parts of the genes may be modified by random mutation. Iteration of this procedure will yield an improvement of the population fitness.

A.1 Chromosome Representation

Chromosomes **ch** represent the characteristics of the individuals. In general, different characteristics are separated in subsets, called genes g_i, such that

$$\mathbf{ch} = \{g_i | \, i = 1 \ldots N_g\} \, , \tag{1}$$

in which N_g denoted the genetic length of the code. The genes g_i are elements of a finite alphabet (often is $g_i \in \mathbb{B}$). The individuals of a population at a time k consist of N_{pop}

chromosomes

$$\mathbf{P}^k = \{\mathbf{ch}_i^k \mid i = 1 \ldots N_{\text{pop}}\} . \quad (2)$$

The chromosome representation of the channel assignment problem can be written as a matrix $\mathbf{S}$

$$\mathbf{S} = \begin{pmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{m1} & \cdots & s_{mn} \end{pmatrix} \text{ with } s_{mn} \in \mathbb{B} = \{0, 1\} , \quad (3)$$

with m network nodes and n available channels $c_j = 1, \ldots, n$. If the channel n is used at node m, it yields that $s_{mn} = 1$, otherwise $s_{mn} = 0$ [5], [7]. The assignment nodes usually correspond to the satellite spotbeams or earth-fixed cells. A chromosome of an individual of the channel assignment problem is formed by the arrangement of the matrix $\mathbf{S}$ as sequence of $i = 1 \ldots m$ genes $g_i = \{s_{ij}\}$ with $j = 1 \ldots n$ channel positions.

Multiple assignments of channels to nodes as required for dynamic channel assignment schemes (DCA) according to a request function $A(v) = (a_1, \ldots, a_m)$ are possible as well. A condition for the channel assignment can be formulated for the chromosome structure that reflects the multiple demand in each cell:

$$\sum_{j=1}^{n} s_{ij} \stackrel{!}{=} a_i . \quad (4)$$

The search space $\mathbb{B}^{m \cdot n}$ of the optimization can be substantially reduced when only such chromosomes $\mathbf{ch} \stackrel{\frown}{=} \mathbf{S}$ are considered for the optimization that fulfill Eq. (4).

A.2 Fitness Evaluation Function

A performance measure is required for the selection of the individuals. The function maps the performance of each individual in the search space $\mathbb{B}^{mn}$ to a positive real value called fitness $f(\mathbf{ch}) : \mathbb{B}^{mn} \rightarrow \mathbb{R}_0^+$. The fitness function may contain several performance aspects with different weights. Higher values of $f(\mathbf{ch})$ indicate a better channel assignment. The carrier to interference power ratio C/I is widely used to characterize the performance of a channel assignment. The C/I contains the noise power $N_0 \cdot B$ in the signal bandwidth B and all interferences coming from the own system (such as co-channel and adjacent channel interferences I_i , CCI and ACI, of users i) or from other systems I_F and inter-symbol interferences due to multipath fading effects [10], [11].

In satellite systems co-channel interference arises from insufficient co-channel suppression of the satellite antenna in neighbor spotbeams. In the following, we will consider the satellite up-link. A subscriber x_i appears in the satellite antenna spotbeam under an off-axis angle θ_i in a distance d_i. Furthermore, $j = 1 \ldots N_I$ interferers exist in the same or in neighbor cells with distances d_j and off-axis angles θ_j to its serving cell. An interferer j appears in the cell of the user i under an angle θ_{ij} in distance d_{ij}. The received power of the ith subscriber at the satellite amounts to

$$C_i = EIRP_i(\theta_i, d_i) \cdot L_{\text{FSL}}^{-1}(d_i) \cdot h_i^2 \cdot G_E(\theta_i) , \quad (5)$$

in which the EIRP of the user depends on a power control scheme P_{PC} to compensate for the free space loss $L_{\text{FSL}} = \left(\frac{4\pi d}{\lambda}\right)^2$ and the satellite antenna gain G_E: $EIRP_i(\theta_i, d_i) = P_S G_S \cdot P_{\text{PC}}(t, \theta_i, d_i)$, $P_{\text{PC}} = f(h_i(t), G_E(\theta_i), L_{\text{FSL}}(d_i))$. The transfer function of the land mobile satellite channel of the ith subscriber is denoted by $h_i(t)$.

The interference power of the co-channels and adjacent channels of a subscriber j is given by

$$I_{ij} = EIRP_j(\theta_j, d_j) \cdot L_{\text{FSL}}^{-1}(d_{ij}) \cdot h_{ij}^2 \cdot G_E(\theta_{ij}) \cdot \mu_j \cdot \gamma_{ij} , \quad (6)$$

with the additional factors μ_i for the mean voice activities and the orthogonality factor γ_{ij}. The mean voice activity μ_i takes the reduction of the co-channel interference during speech gaps into consideration. The orthogonality factor γ_{ij} characterizes the effective part of the interference from subscriber j for the reception of subscriber i. It yields [12]

$$\gamma_{ij} = \begin{cases} 1 & \text{for co-channels,} \\ 0 & \text{for orthogonal channels,} \\ \ll 1 & \text{for adjacent channels or multiple access interference in CDMA.} \end{cases} \quad (7)$$

In CDMA systems the interference contribution of a single interferer in a fading channel can be treated as additive white Gaussian noise (AWGN). The factor γ_{ij} amounts to [13]

$$\gamma_{ij} = \frac{\delta_c}{2G_p^2} \quad (8)$$

with the processing gain $G_p = \frac{R_S}{R_b}$ (R_S: chip rate, R_b: data rate) and the mean cross correlation factor δ_c of the spreading sequence

$$\delta_c = \begin{cases} 2G_p & \text{for A-CDMA with PN-Codes,} \\ 0 & \text{for O-CDMA with orth. WH-Codes,} \\ \frac{1}{N-1} & \text{for QO-CDMA with pp-Gold-Codes.} \end{cases} \quad (9)$$

Abbreviations: A-CDMA: asynchronous CDMA with *pseudo noise* (PN)-codes; O-CDMA: orthogonal synchronous CDMA with Walsh-Hadamard (WH)-codes; QO-CDMA: quasi-orthogonal synchronous CDMA with preferentially phased (pp) Gold-codes (N: number of active subscribers).

The total C/I of the ith subscriber is according to Eq. (6) and (7)

$$\left(\frac{C}{I}\right)_i = \frac{C_i}{kT_iB + I_F + \sum_{j=1}^{N_I} I_{ij}} \quad (10)$$

with I_F denoting the other-system interference.

The fitness of a channel assignment can thus be expressed as sum

$$f(\mathbf{ch}) = \sum_{i=1}^{m} \sum_{j=1}^{n} s_{ij} \left(\frac{C}{I}\right)_{i,c_j} \quad (11)$$

of all single C/I contributions in all cells. Regarding Eq. (12) the advantage of stochastic optimization is obvious: the contributions of all interferers are taken into account. Moreover, not only co-channel interference is considered but also adjacent channel interference and co-site constraints can be included through the orthogonality factor γ_{ij} in Eq. (8). This is an advantage over graph-theoretical approaches that can only consider the interference between one pair of cells.

If one wants to search also for inhomogeneous channel distributions on cell basis according to a channel request function, the fitness can reflect the matching between requested and assigned channel numbers by

$$f(\mathbf{ch}) = \sum_{i=1}^{m} \left(a_i - \sum_{j=1}^{n} s_{ij} \right)^2 + \sum_{i=1}^{m} \sum_{j=1}^{n} s_{ij} \left(\frac{C}{I} \right)_{i,c_j} . \tag{12}$$

A.3 Selection

The genetic manipulation of a population starts with the selection of suited chromosomes on the basis of their fitness $f(\mathbf{ch})$. Several selection operators $Sel(\mathbf{P})$ can be used [14]:

Sel_a **Roulette operator**: the operator chooses a chromosome with a probability proportional to the fitness:

$$\mathbb{E}\{\mathbf{ch}\} = \frac{f(\mathbf{ch})}{\sum_{i=1}^{N_{\text{pop}}} f(\mathbf{ch}_i)} . \tag{13}$$

Sel_b **Ranking operator:** the operator chooses n copies of the N_{pop}/n fittest individuals.

Sel_c **Tournier operator:** the operator chooses iteratively a pair of individuals with equal probability. The fittest of both is selected.

Sel_d **Monogamy operator:** a pair is chosen with a probability proportional to the fitness (like Sel_a) but each individual may be taken only once.

Sel_e **Mixture operator:** equal occurrence of the operators Sel_a, Sel_b and Sel_c

Sel_f **Mixture operator:** equal occurrence of the operators Sel_c and Sel_d

A.4 Crossing

The crossing of chromosomes $Cr(\mathbf{ch})$ is the core of genetic algorithms. Through crossing the search after optimal solutions of a problem is performed. A crossing operator takes two chromosomes $\mathbf{ch}_{1,2}$ and forms another pair $\widehat{\mathbf{ch}}_{1,2}$. The most important operators are:

Cr_a **One-Point Crossing:** the new chromosomes consist of the genes from $\mathbf{ch}_1$ until a crossing point l_k, then of the genes from $\mathbf{ch}_2$

$$\begin{aligned} \widehat{\mathbf{ch}}_1 &= \{g_{1,1}; \ldots; g_{1,l_k}; g_{2,l_k+1}; \ldots g_{2,N_g}\} \\ \widehat{\mathbf{ch}}_2 &= \{g_{2,1}; \ldots; g_{2,l_k}; g_{1,l_k+1}; \ldots g_{1,N_g}\} \end{aligned} \tag{14}$$

The crossing point l_k is equally distributed in $[1;\ N_g]$.

Cr_b **Multiple-Point Crossing:** like operator Cr_a, however with several crossing points

Cr_c **Gene Crossing:** each gene is taken from $\mathbf{ch}_1$ or $\mathbf{ch}_2$ with equal probability.

Cr_c **Mixture operator:** equal occurrence of the operators Cr_a and Cr_c

A.5 Mutation

The mutation of chromosomes prevents the optimization from converging to local maxima. For this purpose the chromosome structure is changed arbitrarily. For channel assignment the following mutation operators $Mut(\mathbf{ch})$ are suited (see Fig. 2) [7]:

Mut_a **new channel assignment per cell:** a new channel assignment is chosen for each cell with probability p_m.

Mut_b **channel exchange per cell:** a channel exchange is done in each cell with probability p_m.

Mut_c **new channel assignment, once per chromosome:** for one cell a new channel assignment is chosen with probability p_m.

Mut_d **channel exchange, once per chromosome:** for one cell a channel exchange is done with probability p_m.

Mut_e **channel exchange between cells:** a channel exchange is done between two cells with probability p_m.

Mut_f **channel assignment of the least used channel:** a channel of a randomly chosen cell is replaced by with a channel, that is used at least in the system.

Mut_g **withdrawal of the most frequently used channel:** the channel that is used most frequently in the system is withdrawn in one cell and replaced by a randomly chosen channel.

Mut_h **channel roulette with maximum interference:** a channel is replaced by a randomly chosen channel proportional to the interference contribution of the channels.

Mut_i **mixture operator:** equal occurrence of the operators Mut_d and Mut_e

Mut_j **mixture operator:** equal occurrence of the operators Mut_d and Mut_h

Mut_k **mixture operator:** equal occurrence of the operators Mut_d, Mut_e, Mut_f, Mut_g and Mut_h

Additionally an **Elite operator** $El(\mathbf{P})$ can be defined that selects a copy of the n_e best chromosomes in the new population if not already included. This operator accelerates the convergence of GA dramatically.

The new population at time $k+1$ is formed through the concatenated application of the operators for selection, crossing and mutation, and eventually the elite operator

$$\mathbf{P}^{k+1} = Mut(Cr(Sel(\mathbf{P}^k))) + El(\mathbf{P}^k) . \tag{15}$$

B. Simulated Annealing

Simulated Annealing (SA) is based on physical cooling processes which tend to minimize the energy in the system. The optimization procedure may be regarded as a special case of GA using one individual $N_{\text{pop}} = 1$ that is affected by mutation only. Kirkpatrick et al. [15] have first demonstrated the applicability of SA for combinatorical problems.

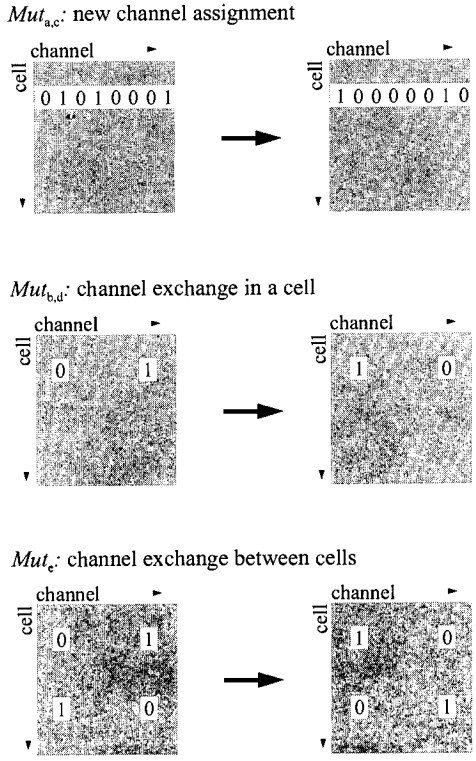


Fig. 2. Mutation operators for genetic algorithms

A cost function $c(\mathbf{ch}) \in \mathbb{R}_0^+$ is required for SA that maps a positive real number for each system state being represented by a chromosome structure **ch**. Lower cost coincides with a better state of **ch**. The cost function can be the reciprocal of the fitness $f(\mathbf{ch})$ (Eq. (12) or Eq. (13)), such that $c(\mathbf{ch}) = g(f(\mathbf{ch}))$. A suitable function for instance is $g(f) = \frac{1}{1+\sqrt{f}}$ [7].

Starting from a random initial state **ch**, adjacent states **ch'** are generated iteratively by mutation. The performance of the new states is evaluated using the cost function $c(\mathbf{ch})$. The new state is taken always for lower cost. For higher cost the state is taken with probability

$$P = \begin{cases} 1 & \text{for} \quad c(\mathbf{ch}') \le c(\mathbf{ch}) \\ e^{-[c(\mathbf{ch}')-c(\mathbf{ch})]/T_{SA}} & \text{for} \quad c(\mathbf{ch}') > c(\mathbf{ch}) \end{cases}. \tag{16}$$

Equation (17) reflects the Boltzmann distribution for thermo-dynamic processes at a system temperature T_{SA}. At higher temperatures T_{SA} a state change to higher energy (or cost) levels is more likely as changes at lower temperatures. The possibility to change to higher states allows the optimization procedure to escape from local minima. The system temperature is decreased in the course of the optimization. Thus, steady states will be reached with lower system energy. The start temperature T_{start} and the cooling behavior $T_{SA}(t)$ affect (beside the mutation operator) the convergence of the SA optimization.

III. Examples and Results

A. Regular Patterns

In this section we want to demonstrate the applicability and performance of stochastic optimization methods. Regular reuse patterns are used for comparison of the performance. These patterns are well known from terrestrial networks. The planning of regular patterns is based on hexagonal grids with equilateral triangles. Maximum dense co-channel assignments with n channels can be reached if the channel number follows the condition

$$n = i^2 + ij + j^2, \quad i, j \in \mathbb{N}_0 . \tag{17}$$

Regular patterns can also be adopted to multibeam satellite antennas. In this paper we will consider only one footprint, but the investigated methods can easily be extended to several footprints in a multi-satellite constellation with overlapping. The multibeam cells in a footprint are usually composed by circular arrangements of cells in several rings r_z where the number of cells n_z increases for each ring by six:

$$n_z = 1 + \sum_{2}^{r_z} 6(r_z - 1) = 1 + 3(r_z - 1) \cdot r_z . \tag{18}$$

The cell boarders can be interpreted either i) as contours of the beamwidth in true view angles of the radiation pattern, or ii) as the earth projection of the antenna contours. In the first case the antenna has cells all with equal beamwidth and gain, in the latter case with equal cell areas that can be approximately achieved by elliptically-shaped beams.

The normalized reuse distance d_s^0 between two co-channels can then be interpreted as spherical angle

$$d_s^0 = \frac{\theta}{\theta_{3\,\mathrm{dB}}/2} = \sqrt{3n} . \tag{19}$$

that is related to the 3-dB half beamwidth angle $\theta_{3\,\mathrm{dB}}$ of the cell for antennas with equal beamwidth, or as earth-centered angles of earth contours for antennas with equal areas, respectively.

To evaluate the performance of the channel assignment the contribution of the co-channel interference (CCI) to the $C/I_{\mathrm{tot.}}$ must be investigated. For this purpose Eq. (11) can be rewritten to

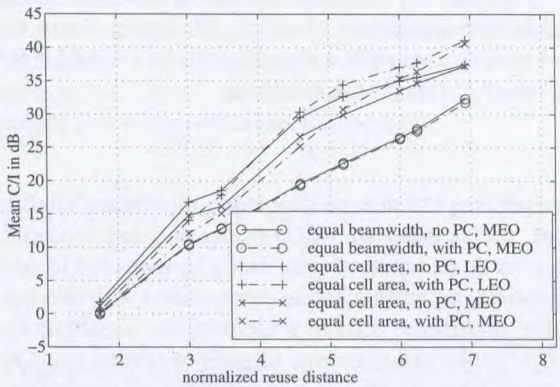
$$\frac{C}{I_{\mathrm{tot.}}} = \frac{1}{1/\frac{C}{N} + 1/\frac{C}{I_F} + 1/\frac{C}{I_{\mathrm{CCI}}}} , \tag{20}$$

denoting the sum of all CCI contributions by I_{CCI} and the noise power $N = N_0 B$. For the system dimensioning one can claim that the interference contribution I_{CCI} may deteriorate the C/N only by a factor x, so

$$\frac{C}{I_{\mathrm{tot.}}} \ge x \cdot \frac{C}{N}, \ (x \le 1) \quad \text{or} \quad \frac{C}{I_{\mathrm{CCI}}} \ge \frac{C}{N} \cdot \frac{x}{1-x} . \tag{21}$$

For instance if a deterioration by $x \doteq$ -0,5 dB is allowed, it follows $\frac{C}{I_{\mathrm{CCI}}} \ge \frac{C}{N} + 8,5$ dB.

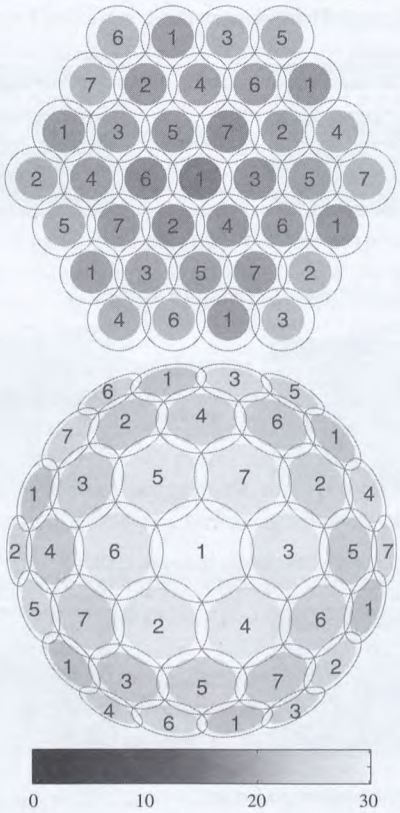
The level of the co-channel interference I_{CCI} that is received through the sidelobes of the satellite antenna depends on several factors: i) the radiation pattern of the antenna, ii) the reuse distance d_s^0 or the channel number n, iii) the position of the interferers in the cells, iv) the orbit height, and v) the power control. Figure 3 shows the numerical evalu-



C/I_{CCI} versus normalized spherical reuse distance d_s^0 in a satellite antenna with 37 cells

ation of the C/I_{CCI} in the up-link according to Eq. (11) for co-channels ($\gamma_{ij} = 1$) in a satellite antenna with four tiers (37 cells). The radiation pattern was assumed to be that of a the envelope of a generic taper antenna with a taper attenuation $T_{\mathrm{dB}} = 20$ dB and $p = 2$ according to [12]. The cell boarder corresponds to the 3-dB decrease of the antenna profile. The curves show the C/I_{CCI} as a function of the normalized reuse distance d_s^0 for different orbit heights. Furthermore, we distinguish between systems with and without power control (PC). If power control is adopted the antenna gain decrease $G_E(\theta)$ and the increasing free space loss $L_{\mathrm{FS}}(d)$ towards the cell boundaries is compensated by a PC correction factor $P_{\mathrm{PC}} \sim \frac{L_{\mathrm{FS}}(d)}{G_E(\theta)}$. In systems without PC the transmit power shall be adjusted in a way that users in the cell centers receive the same C/N, thus, $P_{\mathrm{PC},o} \sim L_{\mathrm{FS}}(\theta_c)$. In addition, a homogeneous user distribution is assumed in the cells.

The C/I_{CCI} ratio of antennas with equal cell area outperforms by several decibel that of antennas with equal beamwidth since cells appear here on the average with smaller diameters, and thus the co-channel suppression becomes higher. The influence of the orbit height is neglectible for antennas with equal beamwidth (the variation of the C/I_{CCI} from 700...36.000 km amounts to less than 0.2 dB). However, the interference contribution of antennas with equal beamsize increases with orbit height since the beams are closer collocated but the ratio of the beamwidth remains constant. Power control deteriorates the interference slightly since the compensation of the antenna gain decrease and the free space loss increase of the wanted user increases also the interference to other users due to the higher transmit power. This effect diminishes for higher channel numbers because of fewer co-channels in the pattern.

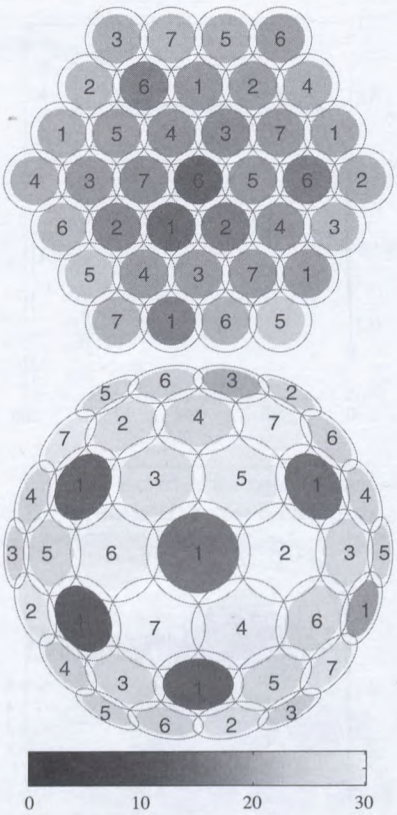


C/I_{CCI} in the antenna cells for a MEO satellite. Upper figure: antenna with equal beamwidth, lower figure: antenna with equal cell area

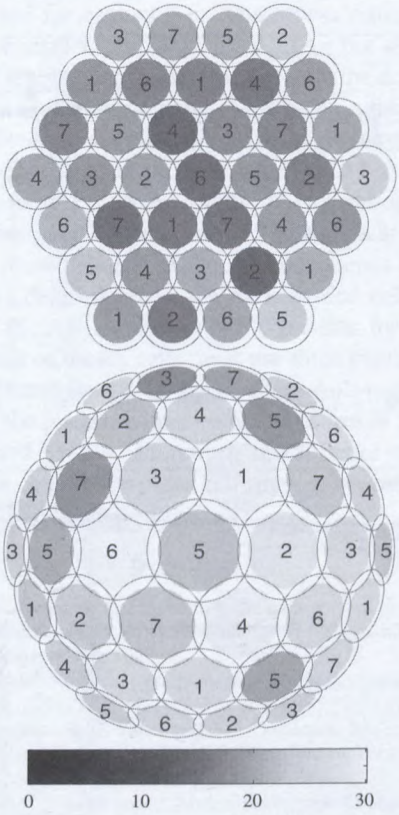
Channel numbers between $n = 3 \ldots 7$ will usually suffice for voice communications with low C/N requirements. The ICO system needs a $C/I_{\mathrm{tot.}}$ of 3,5 dB with QPSK modulation. A reuse pattern with $n = 4$ channels is used [16]. A lower bound for the C/I_{CCI} is 12 dB according to Eq. (22). This value can be almost reached with the pattern of size $n = 4$. Using a pattern of size $n = 7$ improves the interference quite a bit, at the expense of the higher bandwidth requirement.

Figure 4 compares the distribution of the mean values for the C/I_{CCI} in the antenna cells for a MEO satellite with $n = 7$ channels. Two opposite effects are apparent: while the worst values in antennas with equal beamwidth occur in the inner cells, the best values can be found here in antennas with equal cell area since the interferers in the outer cells transmit with lower power due to the higher antenna gain. A second effect lies in the inhomogeneous distribution of the CCI within the cells of a ring. This is caused by the limited expansion of the reuse pattern that yields a unequal distribution of the co-channels.

Above considerations showed that regular patterns may not be optimal in antennas with few cells. Furthermore, one could wish to find patterns with $4 < n < 7$ as a compromise between low co-channel interference and low bandwidth requirements. Stochastic optimization can be used for this



C/I_{CCI} in satellite antennas using genetic algorithms for an MEO satellite antenna diagram with equal beamwidth (upper fig.) and equal cell area (lower fig.)



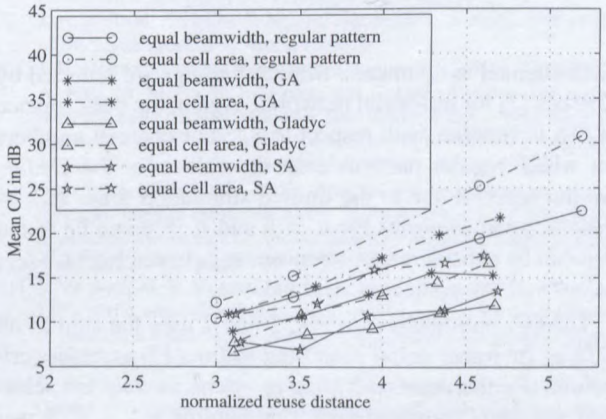
C/I_{CCI} in satellite antennas using simulated annealing for an MEO satellite antenna diagram with equal beamwidth (upper fig.) and equal cell area (lower fig.)

purpose.

B. *Channel Assignment by Stochastic Optimization*

We want now to demonstrate the application of stochastic optimization for the channel assignment in satellite multibeam antennas. The purpose is to get patterns for channel numbers between $n = 4$ and $n = 7$. To estimate the performance a comparison with the optimal pattern for $n = 7$ is given. Figures 5 and 6 show the channel assignments and the mean C/I_{CCI}-values in the cells for $n = 7$ that result from the application of genetic algorithms (GA) and simulated annealing (SA). Like in the previous section a MEO satellite with power control and a generic aperture antenna is used with a homogeneous user distribution on earth. None of the algorithms did find the optimal regular pattern in reasonable short time. The GA algorithm outperforms SA on the average. However, the variation of the mean C/I_{CCI}-values per cell is higher as for SA.

For smaller channel numbers $n < 7$, good results are also achieved by stochastic optimization methods. This can be seen from Fig. 7 that shows the average C/I_{CCI} of all cells for several channel numbers, or their reuse distance, respectively. Furthermore, the same problem was solved with a graph theoretical approach using the GLADYC-algorithm



C/I_{CCI} versus normalized spherical reuse distance d_e^0 with stochastic optimization of an 37-cell multibeam antenna.)

(see [17]). The graph algorithm yields worse values for the C/I_{CCI} than the stochastic optimization methods GA and SA. This is due to the fundamental problem of graph colouring problems that the coupling of nodes can consider only the interference between two nodes. Thus, high margins must protect against worst case interference, yielding ineffi-

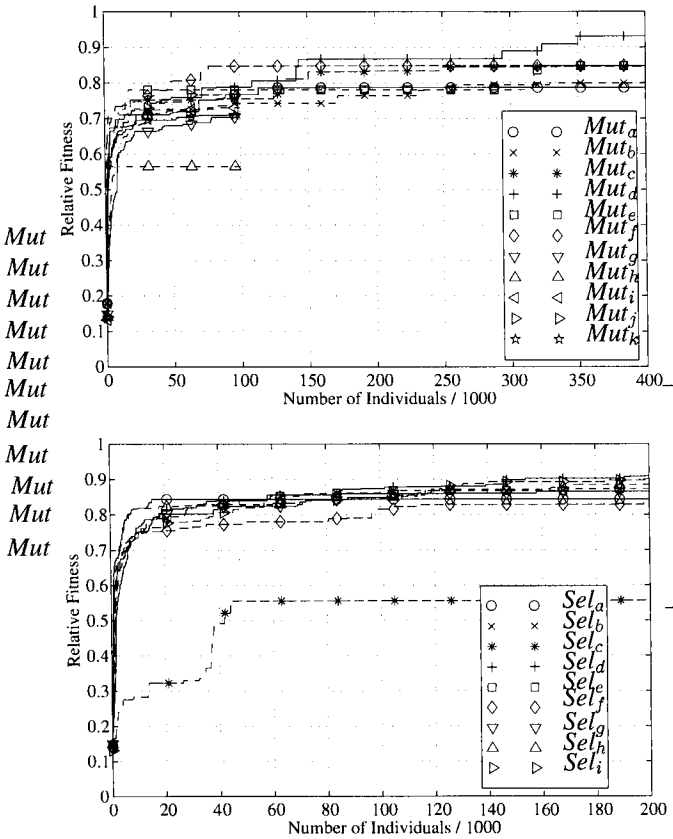


Fig. 8. Dependence of the GA convergence from the mutation operator (upper fig.) and from the selection operator (lower fig.) represented by the relative fitness $f(\mathbf{ch})/f(\mathbf{ch}_{\text{opt}})$ versus the number of simulated individuals $k \cdot N_{\text{pop}}$

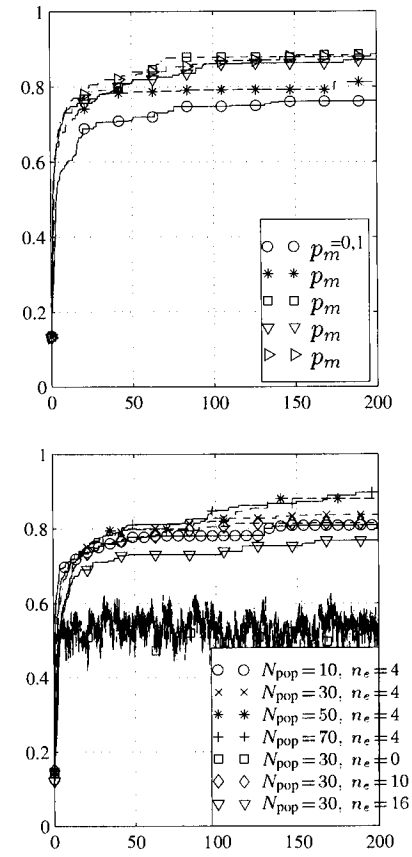


Fig. 9. Dependence of the GA convergence from the mutation probability p_m (upper fig.) and from the population size N_{pop} and n_e of the elite operators (lower fig.) represented by the relative fitness $f(\mathbf{ch})/f(\mathbf{ch}_{\text{opt}})$ versus the number of simulated individuals $k \cdot N_{\text{pop}}$

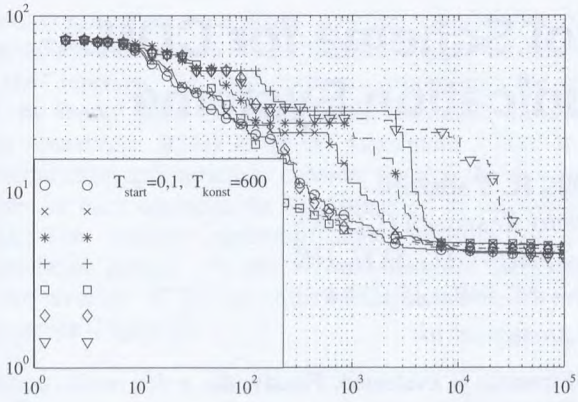
cient channel assignments. Similar findings are reported by Zuerbes [7] for terrestrial networks. Again, the performance of GA is superior with respect to SA. For channel numbers for which regular patterns exist the values for the C/I_{CCI} are not reached due to the limited simulation time. Nevertheless, good solutions for $n = 5$ and $n = 6$ can be found that can be considered as compromise between high C/I_{CCI} values and low frequency demand.

The GA optimization in Fig. 5 and 7 uses the sum of all C/I_{CCI} (in linear units) according to Eq. (12) as fitness criterium. Furthermore, the mixture operators Sel_f for selection and Mut_k for mutation wit probability $p_m = 90\%$ was taken for simulation, as well as one-point crossing Cr_a and the elite operator El with $n_e = 4$. The population size was chosen to $N_{\text{pop}} = 50$. The mutation operator performs the search for the global optimum whereas the selection operator optimizes locally. The dependence of the convergence from the chosen operators is depicted in Fig. 8 for $n = 7$ and an antenna with equal beamwidth. The time course of the fitness $f(\mathbf{ch})$ is normalized with respect to the fitness of the optimal regular pattern $f(\mathbf{ch}_{\text{opt}})$. In general, the mixture operators show faster convergence although the operators Sel_d (monogamy selection) and Mut_d (channel exchange once per chromosome) yield the best results. The mutation probability p_m (cf. Fig. 9) affects the convergence of the algorithms, too. High mutation probabilities $p_m \geq 0.5$ result in fast convergence. Bigger population sizes N_{pop} converge faster as well. The elite operator El has the biggest impact. Without the elite operator the convergence process appears very noisy and the final fitness reaches only 60% of the optimum.

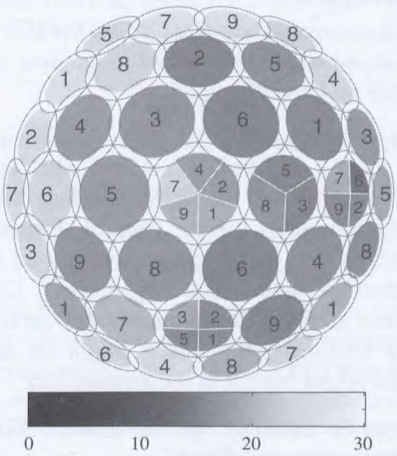
The SA simulation uses again the mutation operator Mut_k. An exponential temperature decay is assumed for the cooling process, thus

$$T_{\text{SA}}(t) = T_{\text{start}} \exp\left(-\frac{t}{T_{\text{const}}}\right) \qquad (22)$$

where T_{start} denotes the start temperature and T_{const} the decay constant. Figure 10 shows the effect of these parameters for $n = 7$ by means of the time course of the cost function $c(\mathbf{ch})$ normalized with respect to the cost $c(\mathbf{ch}_{\text{opt}})$ of an optimal reuse pattern. Cooling processes with high decay constant T_{const} yield good cost values after short simulation time, however, on a long-term perspective the better



T_{start} and from the cooling constant T_{const} represented by the relative cost $c(\mathbf{ch})/c(\mathbf{ch}_{\text{opt}})$ versus the number of simulated individuals $k \cdot N_{\text{pop}}$



C/I_{CCI} in satellite antennas with inhomogeneous traffic demand $A(v)$. The sectors in the cells represent an assigned channel for the whole cell, each.

values are reached by slow temperature decays. This behavior reflects nicely the underlying physical principles, e. g. of crystallization.

C. *Genetic Algorithms for DCA*

Stochastic optimization algorithms are also well suited for dynamic channel allocation (DCA) schemes. Figure 11 shows the result of a DCA channel allocation through GA. An arbitrary traffic demand vector $A(v)$

$$A(v) = \begin{cases} 5 & \text{for } v = 1 \\ 3 & \text{for } v = 2 \\ 4 & \text{for } v = 8 \\ 4 & \text{for } v = 17 \\ 1 & \text{otherwise} \end{cases}$$

was assumed for each cell v. The fitness criterium in Eq. (13) now considers not only the C/I_{CCI} but also that the number of assigned channels matches with the demand $A(v)$ in each cell v. The numbering of the cells in the multibeam footprint is done sequentially from inner to outer rings, counterclockwise in each ring. For Fig. 11 the same assumptions as in Fig. 5 are valid. The chromosome structure was designed to assign $n = 9$ channels at most. The minimum reuse distance of the inhomogeneous channel assignment is determined by the most loaded cells, i. e. the cells $v = \{1, 2, 8\}$. One can easily see that for $n = 9$ the reuse pattern of theses cells must use three channels. Thus, the lower bound for the C/I_{CCI} of these cells is given by the C/I_{CCI} of the pattern with $n = 3$. As shown in Fig. 11 this was achieved by GA. Moreover, much better values were reached for cells in the outer cell rings. The traffic demand is satisfied in all cells.

References

[1] A. Jahn. Resource management techniques applied to satellite communications networks. In *Proceedings of the International Zurich Seminar*, pages 1–8, 1998.

[2] W. K. Hale. Frequency assignment: Theory and applications. *Proceedings of the IEEE*, 68:1497–1514, 1980.

[3] S. A. Grandhi, R. D. Yates, and D. J. Goodman. Resource allocation for cellular radio systems. *IEEE Transactions on Vehicular Technology*, 46:581–587, 1997.

[4] F. J. Jaimes-Romero and D. Muños-Rodríguez. Channel assignment in cellular systems using genetic algorithms. In *IEEE VTS 46th Vehicular Technology Conference (VTC'96)*, pages 741–745, 1996.

[5] A. Pujante and L. de Haro. Optimisation of satellite frequency plans with balanced carriers by simulated annealing. *Electronic Letters*, 33:934–935, 1997.

[6] S. Zürbes. Frequency assignment in cellular radio by stochastic optimization. In *Proc. Second European Personal Mobile Communications Conference (EPMCC'97)*, pages 135–142, 1997.

[7] J. H. Holland. *Adaption in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975.

[8] J. H. Holland. Genetic algorithms. *Sci. America*, 267:66–72, 1992.

[9] A. Jahn, M. Sforza, S. Buonomo, and E. Lutz. Narrow- and wideband channel characterization for land mobile satellite systems: Experimental results at L-band. In *Proceedings Fourth International Mobile Satellite Conference IMSC'95*, pages 115–121, 1995.

[10] A. Jahn, H. Bischl, and G. Heiß. Channel characterisation for spread spectrum satellite communications. In *Proceedings IEEE Fourth International Symposium on Spread Spectrum Techniques and Applications (ISSSTA'96)*, pages 1221–1226, 1996.

[11] F. Vatalaro, G. E. Corazza, C. Caini, and C. Ferrareli. Analysis of LEO, MEO, and GEO global mobile satellite systems in the presence of interference and fading. *IEEE Journal on Selected Areas in Communications*, 13:291–300, 1995.

[12] R. De Gaudenzi, F. Giannetti, and M. Luise. Advances in satellite CDMA transmission for mobile and personal communications. *Proceedings of the IEEE*, 84:18–39, 1996.

[13] D. S. Weile and E. Michielssen. Genetic algorithm optimization applied to electromagnetics: A review. *IEEE Transactions on Antennas and Propagation*, 45:343–353, 1997.

[14] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[15] F. Makita and K. Smith. Design and implementation of ICO system. In *17th AIAA International Communications Satellite Systems Conference and Exhibit*, pages 57–65, 1998.

[16] M. Grevel and A. Sachs. A graph theoretical analysis of dynamic channel assignment algorithms for mobile radiocommunication systems. *Siemens Forschungs- und Entwicklungsberichte*, 12:298–305, 1983.

An Advanced Power Control Scheme for CDMA-based Satellite Communication Systems

P. Taaghol, S. Nourizadeh, R. Tafazolli
Centre for Communication Systems Research (CCSR)
University of Surrey, Guildford, Surrey GU2 5XH, UK
Tel: +44-1483-25 9810, Fax: +44-1483-25 9504
Email: P.Taaghol@ee.surrey.ac.uk

ABSTRACT

In this paper, a new speed adapted closed loop power control algorithm applicable to land-mobile satellite system is proposed. The proposed scheme takes advantage of an advanced speed estimation algorithm to adapt the user terminal step size for improved performance. Further enhancement of the algorithm is achieved by employment of similar speed estimation algorithm at the fixed earth station, enabling issue of more accurate commands. It is shown that through the use of the combined algorithm reductions in standard deviation of the PCE of up to 0.8dB is achieved.

I. INTRODUCTION

Since the capacity of the CDMA system is interference limited, minimising the effects of multiple access interference is of great importance. The use of asynchronous CDMA in the return-link (user terminal -SAT- fixed earth station) gives rise to tight power control requirements as the spreading sequences of different user terminals loose their orthogonality with respect to each other to a large extent. It is hence desired that the received signals from all the user terminals within a spotbeam, at the input of the Fixed Earth Station (FES) receiver to have the same power level, regardless of the position of the user terminals and the characteristics of their channels.

Due to inherent round-trip delays associated with satellite communications, the performance of conventional Closed Loop Power Control (CLPC), designed to track fast fading is greatly reduced . It is therefore desired to provide the User Terminal (UT) and the FES with some level of intelligence to respond to the fast variations appropriately by changing parameters such as the UT step size and the FES signal averaging period based on which commands are issued.

In this paper, it is demonstrated that through the use of a novel adaptive step-size power control algorithm, the user terminal would be able to respond to fast variations more appropriately, thereby reducing the standard deviation of the Power Control Error (PCE).

The paper is organised as follows- section-1 deals with the requirements and limitations of the CLPC in land-mobile satellite systems. This is followed by a description of the considered propagation channel in section-2. In section-3, a new power control algorithm is proposed and its performance is evaluated. Finally, the major results and the finding are discussed in section-4.

II. LIMITATIONS AND REQUIREMENTS

CLPC is designed to track the faster variations of the channel. The fact that a feedback loop exists in the system, imposes an inherent hysteresis phenomenon. That is, by the time the received power is measured (at FES) and the actual power control commands are received by the UT, the fading may have changed and the command to increase or decrease the power may no longer be valid.

As far as the accuracy of the CLPC is concerned, the long Round-Trip Delays (RTDs) associated with the satellite environment are by far the most limiting factor. This significantly limits the Power Control Command Rate (PCCR) which in turn reduces the capability of the CLPC, in tracking the fast fading. But nevertheless, CLPC could still prove very useful under circumstances such as a slow moving or a stationary UT which might be in fade for example in the return but not in the forward link.

Figure 1, shows the minimum and maximum RTDs for three typical non-geostationary orbits of,

- a 66 satellite polar LEO system with a minimum elevation angle of about 8°
- a 48 satellite inclined LEO system with minimum elevation angle of 10°
- a 10 satellite inclined MEO system with minimum elevation angle of 10°

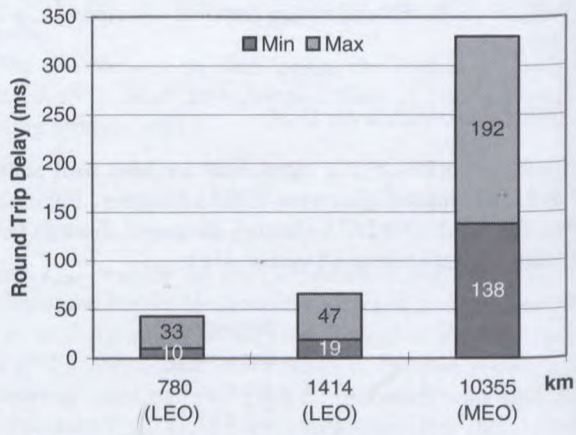


Figure 1: Typical round trip delays (UT-LES-UT)

When considering a particular satellite constellation the minimum and maximum RTDs could be slightly misleading as they represent the two extremes where both the user and the feeder links are either at the minimum elevation angle (maximum RTD) or at 90° (minimum RTD). It is therefore more appropriate to consider the average elevation angles for each constellation for estimation of the average RTD. The average elevation angle statistics of a constellation change with latitude and therefore knowledge of the average RTD for different latitudes is of vital importance (Figure 2).

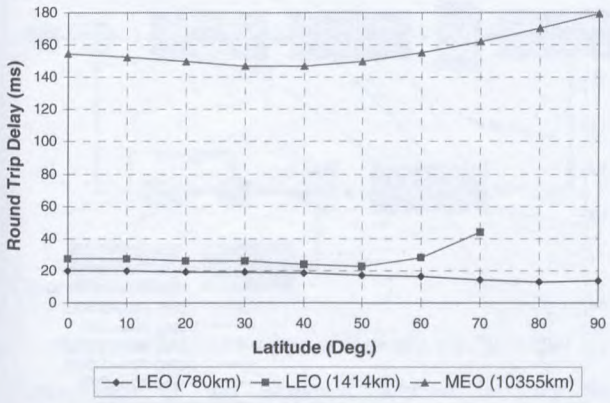


Figure 2: Typical round trip delays (UT-LES-UT)

Knowing the above, various CLPC algorithm parameters can be fine-tuned for a given constellation.

When considering CLPC, one must not also loose sight of the fact that even in the terrestrial environments, were the RTDs are much smaller, the use of CLPCs with high PCCRs is not common practice due to resource limitations. In fact at higher velocities interleaving becomes a more efficient technique in randomising the errors and hence maintaining the required quality of service. On the other and, increased interleaving depths would introduce longer delays which may not particularly suite a satellite system with an existing RTD. Therefore, striking the balance between the performance of CLPC and interleaving for a given constellation becomes an absolute necessity.

III. CHANNEL

In terrestrial cellular communication systems, the mobile user does not benefit from direct Line-of-Sight (LOS) to the base station most of the time. This implies that the mobile terminal would rely on the multipath reflection and echoes for communication. In relatively narrowband communication systems, the existence of multipath causes severe fading and degradation of the link quality. In wideband CDMA systems, however, the multipath environment can be exploited through RAKE receiver architecture, allowing signals arriving with different propagation delays to be independently received and combined to provide an additional gain. The use of RAKE receivers in terrestrial suburban and urban environments proves to be very effective due to the sufficient delay spread associated with such channels. However, it has been demonstrated in the wideband channel measurement campaigns of [1], [2] and [3], that in satellite channels the delay spread has an average of about 100ns. Hence any CDMA system designed to effectively make use of the spreading to combat multipath would have to be spreading by at least an amount greater than the coherence bandwidth, i.e. 10MHz or more. Even if the multipath in a satellite channel is resolved, it is very likely that the echoes arrive at much lower power compared to that of the LOS, limiting any potential multipath diversity gain. Nevertheless, RAKE receivers are still useful for combining different satellite diversity paths as discussed in [4].

Taking all the above into consideration, a widely used narrowband two-state Markov representation of the propagation channel could therefore be considered. Under such an assumption, the propagation channel can be categorised into a good and bad state, the parameters of which depend on the operational frequency and environment. The good state is characterised as a Ricean channel and the bad state as log-normal shadowing together with a Rayleigh channel. When in the bad state (shadowed), the signal level significantly drops. In practice, the Open Loop Power Control (OLPC) would track the large signal variations caused by shadowing. However, considering the limited link margins of the first generation mobile personal satellite communication systems (typically 8-16dB), it is very unlikely that the power control dynamic range can cope with large shadowing variations particularly in hostile environments such as urban and tree-shadowed. In this case the CLPC would not help due to the same limited dynamic range of the UT and the fast variations of the channel.

The CLPC is therefore used to combat the Ricean fading of the good sate which could introduce great improvements for a CDMA-based system with an asynchronous return-link. In this paper, the CLPC algorithm has therefore been optimised in correlated Ricean channels with Rice factors, representing two commonly experienced elevation angles.

IV. SPEED ADAPTED CLOSED LOOP POWER CONTROL

a. Conventional CLPC

Before describing and evaluating the proposed power control scheme, it is important to establish the reference conventional CLPC by which comparisons will be made.

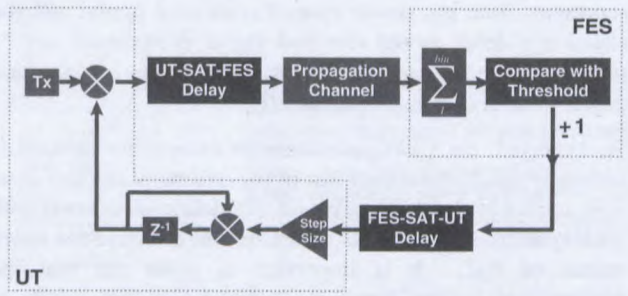


Figure 3: Conventional CLPC model

Figure 3, shows the considered conventional CLPC model [5]. In this model, it is assumed that the power control commands are all received by the UT correctly in the forward link. It is further assumed that the power control bits are not encoded and hence do not experience any additional decoding delay. Due to the correlated nature of the channel, it is extremely difficult to analytically derive the power control error [6]. Simulation is therefore used to evaluate the performance of the loop.

In conventional CLPC, different UT power control step sizes may be negotiated during the call set-up. But nevertheless, the power control step size is assumed to be fixed for the entire duration of the call. This is to some extent limited as the fixed step CLPC performance varies under different,

- vehicular speeds
- operating centre frequencies
- propagation channels
- round trip delays

In fact different values of step size can enhance the performance for particular combination of the above operational parameters.

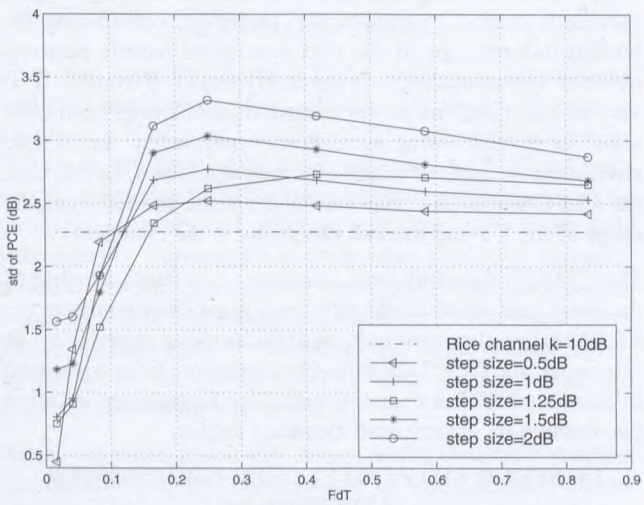


Figure 4: Performance of fixed step power control in Ricean channel, Rice factor of 10 dB

Figure 4, shows the experienced power control error (PCE) versus F_dT, where, F_d represents the Doppler frequency (the combined speed and centre frequency effect), and T represents both the power control command period and the round trip delay as the received signal is averaged over a period of a single round trip delay. This implies a maximum power control command rate of $1/T$.

As expected, the CLPC performance in satellite systems is generally much worst than that of the terrestrial cellular case. This is simply due to long round trip delays associated with such systems. Nevertheless, CLPC is still effective for lower values of F_dT. It is important to point out that the performance figures above show that for higher values of F_dT, an increased PCE compared to that of the Ricean channel with no CLPC (the dashed line) is experienced. For higher values of F_dT, clearly, the lower the step size is, the lower PCE would be.

From the above it can also be observed that the performance of the CLPC varies with different power control step sizes. This can be better observed in the magnified version of same plot in Figure 5.

As it can be seen below, a single step size does not necessarily have the best performance at all the F_dTs'.

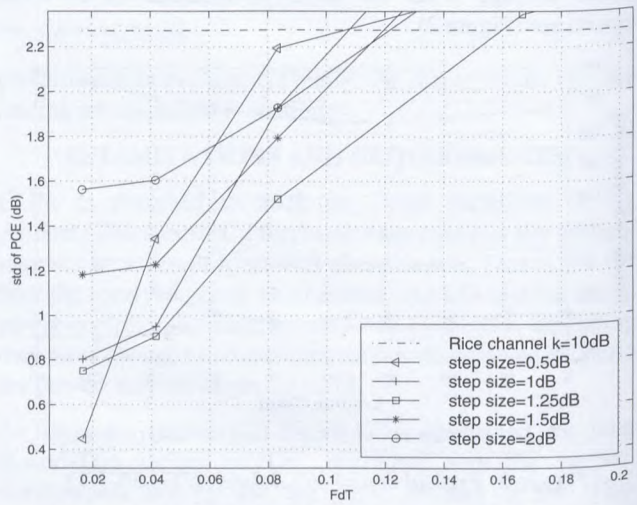


Figure 5: Magnified performance of fixed step power control of Figure 4

Figure 6, shows the performance of the conventional loop in a Ricean channel with a lower factor of 8dB.

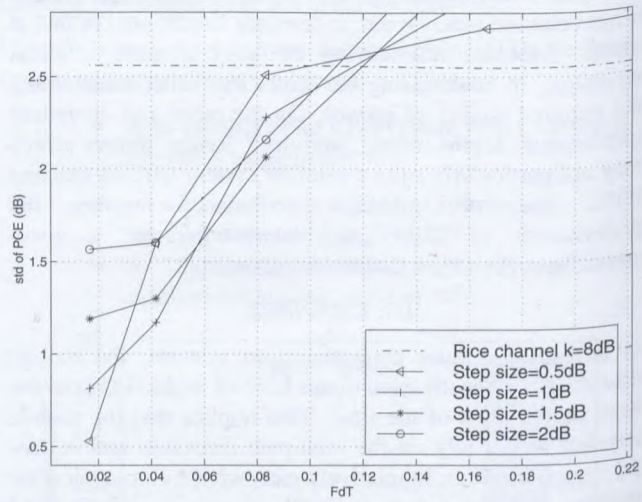


Figure 6: Performance of fixed step power control in Ricean channel, Rice factor of 8dB

A slight degradation in the CLPC performance compared to the case of Figure 4, can be observed as the considered channel in Figure 6, fluctuates relatively faster. Furthermore, note that in the above case the step size of 1.25dB is not depicted as it did not produce a different result compared to that of the 1dB case.

b. Speed-adapted step size CLPC

Assuming knowledge of the vehicular speed through a speed estimation algorithm, the UT would then be able to select the best step size by using the appropriate lookup tables (Figure 7) . This will lead to an overall improvement of the CLPC performance for a wide range of vehicular speeds. Moreover, under such a scheme the CLPC algorithm can be switched-off after certain vehicular speeds where significant degradation in the performance is expected.

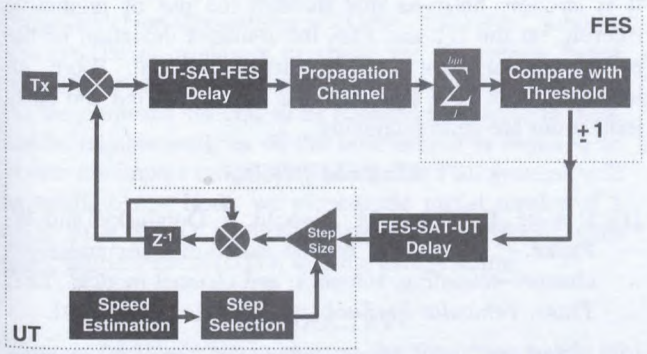


Figure 7: Speed-adapted step size CLPC model

Performance of the proposed scheme is very much dependant on the accuracy of the estimated user terminal speed. Evaluation of a novel environment and speed detection algorithm (pending patent) has shown that high accuracy estimations of the speed at UT and FES are possible. The proposed algorithm takes advantage of several information such as average level crossing, channel impulse response and local mean variations of the received signal to estimate the user terminal speed accurately under a wide range of operational environments.

Figure 8, shows the performance comparison between the conventional fixed step and the speed-adapted scheme.

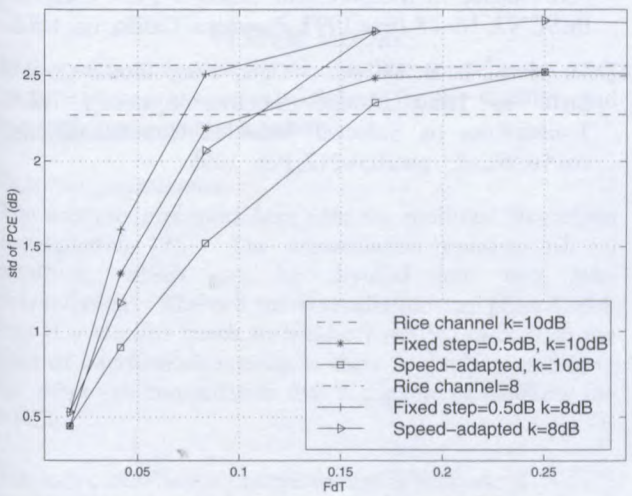


Figure 8: Comparison between the conventional 0.5 dB fixed-step and speed-adapted CLPC

From the above performance gains of more than 0.5dB for some values of F_dT can be observed.

b. Speed-adapted bin size CLPC

Considering the long RTDs of the channel, it is beneficial to reduce the averaging periods (bin sizes) based on which the commands are issued at the FES. That is to issue commands only based on a much smaller portion of the received signal between two consecutive issuing. Under the proposed scheme, two different approaches were investigated.

In the first approach, control commands are issued based on the last *0.25T* seconds of the received signal, where *T* is the power control command period or the round trip delay. This requires no additional estimation algorithm at the FES as the power control command rate is set for a given system.

In an alternative approach the decision is made over an a varying averaging period which always corresponds to the time taken for the UT to travel 1/500 of a wavelength. This approach can only be implemented with the aid of a speed estimation algorithm in the FES, as shown in Figure 9.

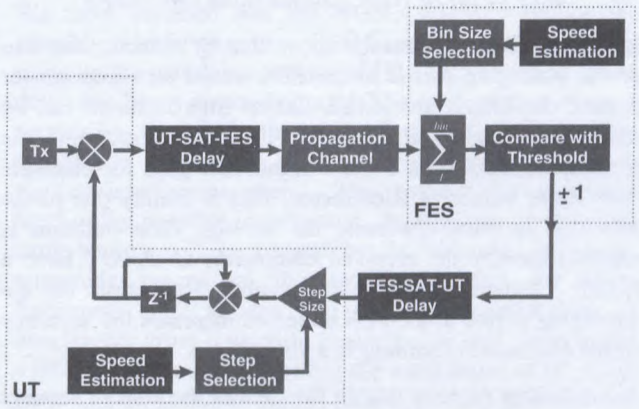


Figure 9: Speed-adapted step and bin size CLPC model

Simulations results for both options in Ricean channels with factors of 8 & 10dB are shown in Figure 10 &Figure 11, respectively.

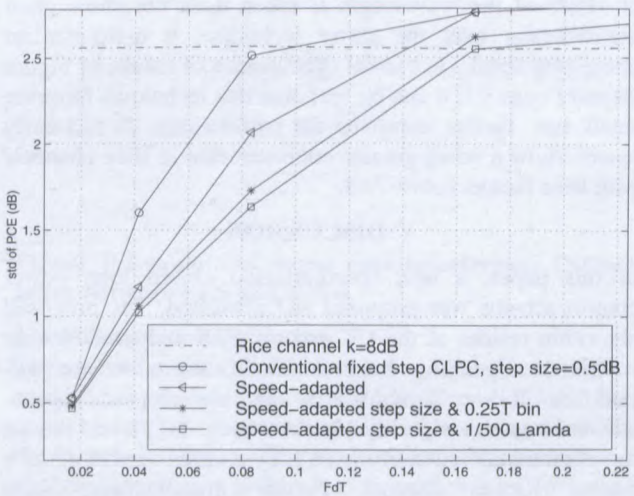


Figure 10: Speed-adapted step and bin size CLPC model, Rice factor of 8dB, Lamda is the wavelength

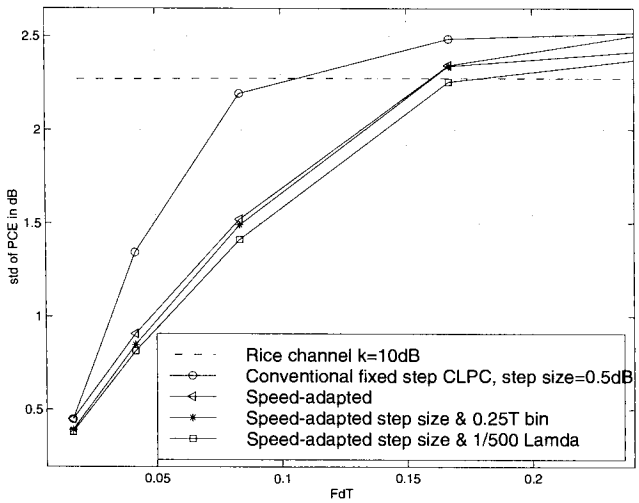


Figure 11: Speed-adapted step and bin size CLPC model, Rice factor of 10dB, Lamda is the wavelength

It can be observed from the above that by reducing the size or the averaging period in the FES, based on which power control commands are issued, further improvements can be made. From Figure 10 and Figure 11, it can be seen that this techniques introduces a more significant gain for channels with lower values of Rice factor. This is mainly due to the fact that in such channels, the average fade duration is shorter, thereby the received commands at the UT have a higher probability of being invalid. Reduction of the averaging period at the FES therefore increases the accuracy of the commands resulting in a visible gain.

It can further be seen that in the case of the Ricean channel with a factor of *8dB* (Figure 10), the first technique in which averaging periods of *0.25T*, are taken improves the PCE by a further *0.3dB*, compared to that of the speed-adapted step size algorithm.

Although the second technique whereby an averaging period of 1/500 of the wavelength is taken does not show great improvement over the above technique, it highlights an interesting trend. On careful examination of results of Figure 10 and Figure 11, it can be seen that this technique, however small, has further improved the performance. In fact early results show a much greater improvements in Rice channels with Rice factors below *7dB*.

V. DISCUSSION

In this paper, a new speed-adapted closed loop power control scheme was proposed and evaluated. The proposed algorithm resides at the UT and the FES and introduce no additional signalling load or modifications to the air-interface. It was demonstrated that the proposed speed-adapted step size algorithm (placed at the UT) could reduce the standard deviation of the PCE by a little over 0.6 dB in a typical Ricean channel. Further improvements were introduced through the use of shortening the signal averaging periods based on which power control commands are issued. Two techniques for this were proposed and evaluated in conjunction with the rest of the alternatives. The first techniques uses a fixed averaging period equal to a quarter of the power control command period (or the round trip delay). The second technique takes advantage of speed estimation in order to keep the averaging period fixed to 1/500 of the wavelength. Both the techniques introduce the most significant gains under more sever fading channels with lower Rice factors. Improvements of up to *0.3dB* are experienced.

It is strongly believed that through the use of predictive filtering at the UT and FES the standard deviation of the power control error can be further reduced. Work on introduction of such algorithm has already started and early indications are quite promising.

VI. REFERNCES

[1] E. Lutz, D. Cygan, M. Dippold, F. Dolainsky, and W. Papke, " The land mobile satellite communication channel–recording, statistics, and channel model,"*IEEE Trans. Vehicular Technol.*, vol. 40, no. 2, May 1991.

[2] N.Kleiner, W.J.Vogel, "Impact of Propagation Impairments on Optimal Personal Mobile Satellite Communication System Design", November 1992, Australia.

[3] M A N Parks, B G Evans, G.Butt and S Buonomo; Simultaneous Wideband propagation measurement applicable to mobile satellite communication systems as L–band and S–band; AIAA96

[4] P. Taaghol, A. Sammut, R. Tafazolli, B. G. Evans, "Satellite Diversity and its Implications on the RAKE Receiver Architecture for CDMA–Based S–PCNs", IMSC'95, Ottawa, Canada, June 1995.

[5] P. Taaghol, R. Tafazolli, B. G. Evans, "Power Control Performance in Measurement–based S–PCN Channel", IMSC'97, 16–18 June 1997, Pasadena, California, USA.

[6] M. Monk, L. B. Milstein, "Open–Loop Power Control Error in Land Mobile Satellite System", IEEE Transactions on Selected Areas in Communications, vol.13, No. 2, pp.205–212, Feb. 1995.

Analysis and Simulation of Interference from NGSO Satellites to GSO Earth Stations

R.W. Kerr[1], M. Moher[1], M. Caron[1], and V. Mimis[2]
[1]Communication Research Centre, 3701 Carling Ave., P.O. Box 11490 Station H, Ottawa, Ontario K2H 8S2, Canada
[2]Space Engineering, Industry Canada, Jean Edmonds Tower North, 300 Slater Street, Ottawa, Ontario, K1A 0C8

ABSTRACT

Proposed Non-Geostationary Orbit (NGSO) Satellite services have proposed to share the frequency bands currently used by Fixed Satellite Services (FSS). In order for the proposed systems to be licensed to operate in these bands, reliable analysis of the interference is required to ensure the impact on existing and future FSS systems will be negligible. Here, we present the initial results of a study of a NGSO satellite constellation interfering with Geostationary Orbit (GSO) satellite earth stations.

INTRODUCTION

There is great interest in sharing the frequency bands of geostationary orbit (GS0) satellite services with non-geostationary orbit (NGSO) satellite services. However, the amount of interference that would be caused by the NGSO services to the GSO satellite earth stations is a concern to the GSO operators and licensing bodies. In this paper, the interference from an NGSO satellite into a reference GSO earth station is calculated through simulation of a NGSO satellite constellation and analysis. The effect on interference for earth stations at different latitudes is examined by placing the reference earth station at three different latitudes: 0°, 20°, and 53°.

In the following sections, we will present the system model, the simulation results and the interference calculations for a reference earth station.

SYSTEM MODEL

In this section, we describe the satellite constellation used, an interference mitigation technique for the satellites and the earth station antenna pattern.

Satellite constellation

The analysis presented here uses the modified Skybridge constellation [2]. The constellation consists of 80 satellites, which can be divided into two sub-constellations. The two sub-constellations are phased such that if a satellite needs to hand-off traffic, such as in the case of interference avoidance, there is another satellite in the other sub-constellation that is capable of handling the traffic.

The sub-constellation parameters are as follows:

- circular orbits with 53° inclination
- 10 orbital planes spaced 36° at the equator
- 4 satellites per plane spaced 90° apart in true anomaly
- altitude of 1469.3 km and orbital period of 115 min.

The relative phasing between the two sub-constellations is -18 between the ascending nodes and 14 spacing between the mean anomalies. The relative spacing between satellites in adjacent sub-constellations is 45

The constellation repeats the same ground track every 35.6 days.

Interference reduction technique

We have assumed that the NGSO satellites reduce the possible interference to a GSO satellite earth station by using an arc avoidance technique [1]. This entails shutting down the beam directed towards the earth station when it approaches near the main beam of the earth station. This prevents a main-beam (satellite) to main-beam (earth station) interference event. This reduces the maximum level of the possible interference. Interference to the GSO earth station will be caused by the sidelobes from the remaining beams of the satellite sharing the same frequency. We assume that the main contribution of interference from a satellite constellation will occur when a NGSO passes directly within the main beam of the earth station. This scenario of the NGSO passing between the earth station and the GSO is illustrated in Figure 1.

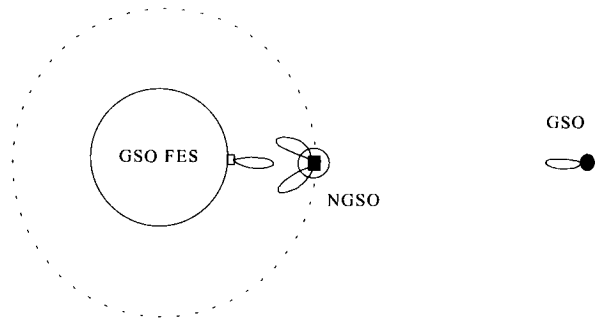


Figure 1 Example of worst case interference (NGSO inline between FES and GSO)

Earth Station Antenna Pattern

The antenna pattern used in this paper for the earth station antenna is modified from ITU-R Rec. 1323. The modification was proposed in [3] and attempts to define the close-in sidelobes. The modified antenna pattern has a slightly higher gain in the close-in sidelobes as compared with the ITU-R Rec. 1323 antenna pattern. The antenna gain, $G(\theta)$, as a function of off-axis angle, θ, for the modified characteristic is defined as:

For $0 < \theta \le \phi_m$,

$$G(\theta) = G_{\max} - 2.5\times10^{-3}(\theta \times D/\lambda)^2 \text{ dB}$$

For $\phi_m < \theta \le \phi_R$, $G(\theta) = G_1 = -1 + 15\log(D/\lambda)$ dB

For $\phi_R < \theta \le 48$, $G(\theta) = 32 - 25\log(\theta)$ dB

For $48 < \theta \le 180$ $G(\theta) = -10$ dB

Where $\phi_m = (20\lambda/D)\sqrt{(G_{\max} - G_1)}$,

$\phi_R = 20.85\times(\lambda/D)^{0.6}$,

λ is the wavelength in metres,

D is the antenna diameter,

$G_{\max} = 20\log(Df) + 18.53$ dBi (antenna efficiency is η= 0.65, f is frequency in GHz)

The antenna gain as a function of off-axis angle is shown in Figure 2.

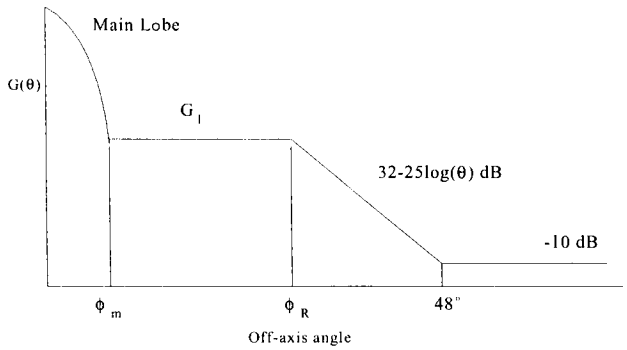


Figure 2 Sidelobe pattern used for the GSO earth station

SIMULATION RESULTS

In order to gather statistics on the frequency and duration of interference events, the satellite constellation was simulated. The simulator was written by Caron et al. [4]. The software records the probability density function (pdf) of the off-axis angles of the satellites from the GSO satellite earth station. The range of off-axis angles defined as "interference events" is specified at the beginning of the program. The software then records the interfering satellite number and the duration of the interference event. The percentage of time that multiple satellites are interfering is also recorded.

The Skybridge constellation was simulated for a period of 35.6 days to gather statistics on the interference events for each earth station position. This period is the amount of time required for the constellation to repeat its exact ground path. The positions were calculated for time increments of one second, therefore the resolution of the event durations are in one second steps.

The simulations were carried out for earth stations on the equator, at a moderate (20°) and high (53°) latitudes. In all cases, the antenna is pointed in the direction that maximizes the interference event duration. In [5], a method is presented that gives the required latitude and difference in longitude between the earth station and the GEO satellite, which maximizes the interference event durations for the NGSO constellation. The differences in longitude were calculated using the method in [5] for the earth stations at 0°, 20°, and 53° latitude and are presented in Table 1.

Table 1 GSO satellite earth station relative location

Station	Latitude of earth station	Difference in Longitude
Equator	0°	0°
Moderate	20°	13.7°
High	53°	58.4°

The statistics on the interference events near the boresight of the earth stations' antenna are presented in Tables 2 to 4 for the earth stations at the 0°, 20°, and 53° latitude, respectively. From the tables, it can be seen that the length of the interference events tends to increase as the latitude increases. This is due to the fact that the satellite orbit transits through the main beam of the earth station at an oblique angle for the high latitude station but passes through perpendicularly for the equatorial station.

INTERFERENCE CALCULATION

As the NGSO satellites orbit the earth, they will inject interference into the GSO satellite earth station's receiver. The amount of this interference is dependent on the flux density produced on the ground by the NGSO satellites, and the earth station's antenna pattern in the direction of the satellites. As the positions of the NGSO satellites relative to the earth station change with time, it is important to find the interference events that will cause harmful interference to the GSO satellite earth station. Here, "harmful interference" is defined as sufficient interference to disrupt communications of the GSO satellite earth station. The earth station locations are the same as presented in the previous section.

The NGSO satellites are iso-flux and thus produce the same flux density on the ground regardless of direction and range of the satellite from the earth station (when the beam is operational). We make the assumption that the flux density produced by the sidelobes of interfering beams is also constant. The probability density functions (pdf) of the off-axis angles, which were found by simulating the satellite constellation, are used in this section. With these pdfs, we are able to find the pdfs of the equivalent power flux density (*epfd*), the interference power received by the GSO satellite earth station, and the interference-to-long-term noise power ratio (I/N).

The link margin of the GSO satellite earth station channel defines the maximum tolerable value of I/N. An outage in communications for the GSO satellite earth station occurs when the value of I/N is above this threshold.

From the pdfs of the I/N for the three earth stations and two antennas, we can find the percentage of time that the threshold value of I/N is exceeded and the earth station is deemed to have a communication outage.

Flux density

The Skybridge system uses iso-flux satellites; thus, the flux density at the surface of the earth is constant. In [1] the on-axis flux density on the ground for different services (i.e., different communication channels), ranges from −160.7 to −155.4 dBW/m^2/4kHz. The sidelobe contribution is calculated assuming 6 beams surrounding a central beam. We assume that the adjacent beams have antenna gains of −10 dBi in the area covered by the central beam. With this assumption a single adjacent beam would contribute −181.8 dBW/m^2/4kHz and the six beams would create an *efpd* of −174 dBW/m^2/4kHz.

The equivalent flux power density (*efpd*) is calculated as follows [5]:

$$efpd(\theta) = 10\log 10(\sum_{i=1}^{Ns} 10^{pfd_i/10}) - G_{\max} + G(\theta_i) \quad \text{dBW/m}^2 \qquad (1)$$

where

pfd_i is the power flux density (dBW/m^2) for each satellite

G_{max} is the maximum on-axis gain (dBi)

$G(\theta_i)$ is the gain of the GSO antenna at θ_i degrees off-axis, in dBi

The flux density for a given off-axis angle is calculated using Eq. 1 and the following conditions:

1) For off-axis angles less than 10°, the main beam is shut down and only the sidelobes contribute to the flux density at the ground. In this case we assume that there is another satellite which has an off-axis angle greater than 10° serving the cell.
2) For off-axis angles greater than 10°, both the main beam and the sidelobes contribute to the flux density.
3) There is only one main beam serving the cell.

Conditions 1 and 2 represent the procedure suggested by Skybridge [1] for reducing interference to the GSO earth stations. The satellites will shutdown the beam covering the GSO earth station area when the satellite is within 10° of the boresight of the GSO earth station. This is to prevent main-beam to main-beam interference between the NGSO and GSO earth station. Prior to shutdown, the traffic is handed over to another satellite with an off-axis angle greater than 10°. Thus, the earth station will have a satellite with only the sidelobes of surrounding beams illuminating the main beam of the earth station and another satellite with a main beam illuminating a sidelobe of the GSO earth station.

The *efpd* is computed versus off-axis angle by Eq. 1. The percentage of time that a satellite is at a given angle was obtained by simulation. With the information obtained from simulation, we compute the complementary distribution function (cdf) for the *efpd*. The cdf for the *efpd* is plotted in Figure3. For comparison with the simulation results, the proposed limits from the WARC 97 for a 10 m antenna [3] operating at 11.7 to 12.2 GHz in Region 2 and 12.2 to 12.5 GHz in Region 3 are included in the plot. The high latitude station exceeds the proposed flux density levels while the equatorial and mid-latitude stations are below the proposed levels.

To compute interference peaks in terms of the long-term noise power (i.e. I/N) for the single satellite case, we use the following equation:

$$\frac{I}{N} = epfd + 10\log_{10}(\pi D^2\eta/4) - k - T - 10*\log 10(W) \text{ dB} \qquad (2)$$

where *epfd* is the equivalent power flux density at the antenna, D is the antenna diameter, η is the antenna efficiency (0.65), k is Boltzman's constant (-228.6 dB/K), T is the system noise temperature (dBK), and W is the reference noise bandwidth (4 kHz, in this case)

The cdf of the I/N can be found using the cdf of the *epfd* and Eq. (2). The cdf of I/N is shown in Figure 4 For a given link margin, we can calculate the corresponding value of I/N and from Figure 4, we can find the percentage of time that the I/N value is exceeded. When the value of I/N is exceeded an outage occurs at the earth station. For example, consider a link margin of 1 dB. It can be shown that the corresponding value of I/N is −5.86 dB. From Figure 4, the probabilities that this value is exceeded are 0.0005, 0.0006 and 0.006 for the earth stations at 0°, 20° and 53° latitude away from the equator, respectively. Thus, the associated availability, due to the NGSO satellite constellation interference alone, for each of the earth stations would be 99.9995, 99.9994 and 99.994.

CONCLUSIONS

We have provided digital link availability results for three GSO satellite earth stations that receive interference from an NGSO satellite constellation operating in the same frequency band. From the simulation of the NGSO satellite constellation and analysis, it is apparent that with our assumptions and system model, the high latitude GSO satellite earth stations can receive higher levels, more frequent and longer duration short term interference. Our interference mitigation technique was to turn off the beam that would transmit into the earth station as suggested in [1]. It may be necessary to find a better mitigation technique, such as turning off more than one satellite beam when crossing through a GSO earth station's main beam. Since the completion of the analysis reported in this paper, it is understood that Skybridge have refined their orbital avoidance technique in order to comply with the provisional *epfd* limits.

REFERENCES

[1] *Application of Skybridge L.L.C. for Authority to Launch and Operate The Skybridge System,* FCC filing, Doc# DC1:52000.3 1394A, February 28, 1997.

[2] Input to ITU-R meeting in July 1998, Document 4-9-11/192(Corr.1)-E

[3] ITU-R Document 4-9-11/Can.8. NGSO Interference Into GSO Earth Stations in the Ku-band Fixed Satellite Service, June 18, 1998

[4] M. Caron, D. Andean, D.J. Hindson, "On a semi-analytical tool to characterize interference between GEO and non-GEO systems", Third Ka-Band Utilization Conference, Sorrento, September 15-18, 1997.

[5] ITU-R Document 4-9-11/13-E, Characterizing the time profile of peaks of interference from NGSO satellite to large-dish GSO earth station antenna, Feb. 23, 1998.

Table 2 Interference durations for off-axis angles near the main beam of the earth station antenna. Earth station located on the Equator.

Range (degrees)	$0 < \theta \leq 0.2218$	$0.2218 < \theta \leq 0.5521$	$.5521 < \theta \leq 1$
Mean Duration (s)	1.4118	2.0769	2.7246
St. Dev	0.5073	1.0504	1.5031
Max. Duration (s)	2	4	6
Min. Duration (s)	1	1	1
No. of events	17	65	135

Table 3 Interference durations for off-axis angles near the main beam of the mid-latitude (20°) earth station antenna.

Range (degrees)	$0 < \theta \leq 0.2218$	$0.2218 < \theta \leq 0.5521$	$.5521 < \theta \leq 1$
Mean	1.8750	2.5479	3.2378
St. Dev	0.3416	1.3848	1.7718
Maximum (s)	2	5	8
Minimum (s)	1	1	1
No. of events	16	71	160

Table 4 Interference durations for off-axis angles near the main beam of the high latitude (53°) earth station antenna.

Range (degrees)	$0 < \theta \leq 0.2218$	$0.2218 < \theta \leq 0.5521$	$.5521 < \theta \leq 1$
Mean	4.1489	6.0186	8.2374
St. Dev	1.5378	3.1710	5.0141
Maximum (s)	7	15	26
Minimum (s)	1	1	3
No. of events	94	323	653

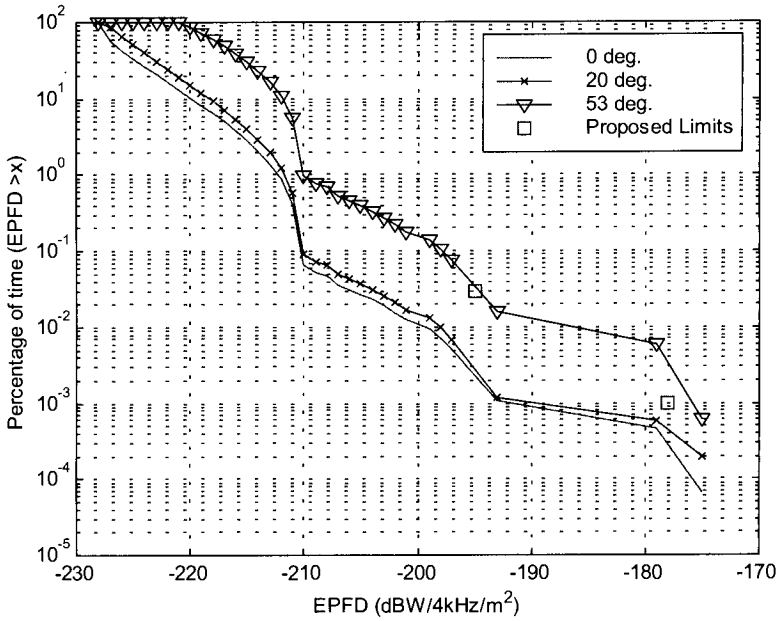


Figure 3 Calculated flux density distribution. Earth stations located 0°, 20°, and 53° latitude away from the equator. Proposed *efpd* limits from WARC-97 for 10m [3].

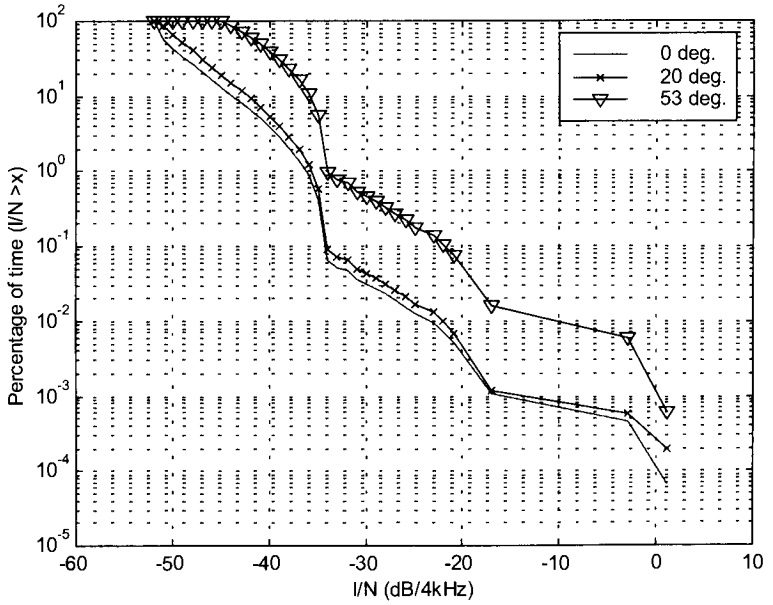


Figure 4 Percentage of time that I/N exceeds a given value. (Antenna diameter = 10m, Noise figure 33.6 dBK.)

Alternatives for the Next Generation of Mobile Satellite Services

Roger J. Rusch, President
TelAstra, Inc.
P.O. Box 4620, Palos Verdes, CA 90274, USA
Email: RogerRusch@aol.com

ABSTRACT

We live at the beginning of the age of personal communications. Few could deny the enormous benefits of these mobile communications services. Satellites provide a unique means to provide universal personal communications. Today, we expect that the ideal personal communicator will be small, convenient, reliable, and ubiquitous. Technology has enabled voice communications, but email and Internet services are the logical extension of the digital revolution. Within the next decade we can expect that new technologies will enable improvements in terminal size, service reliability, user cost, and data rates. It appears that satellites will provide better services that are more affordable than prior satellite systems.

This paper will examine business prospects for the new systems that have been proposed for the next generation of Mobile Satellite Services (MSS). We will examine the proposals for the use of new frequency bands and the new services which will be offered. The source material is drawn from documentation submitted to the US FCC, public forums, and interviews with representative from these and other companies to discuss future plans.

The FCC has received nine applications for use of the 2 GHz spectrum that has been allocated for MSS services after 2000. The desired spectrum is in the allocated 2.0/2.2 GHz bands for mobile communications services. The following systems filed with the FCC (or notified intention to use) in late September 1997. 2 GHz MSS (Boeing), CANSAT M3 (TMI), CELLSAT (Celsat America), Constellation II (CCI), Ellipso 2G (MCHI), GS-2 (Globalstar LP), Horizon (Inmarsat), ICO (ICO Services), and MACROCELL (Iridium LLC). Other companies that are planning next generation service include ACeS of Indonesia, ACS of India, APMT of SE Asia, and Thuraya for the Middle East.

STATUS OF MOBILE SATELLITE SYSTEMS

Inmarsat has been providing MSS for over 20 years. Inmarsat services have expanded from maritime to aeronautical to land mobile services. Some services are for voice and others are data. After eight years, Inmarsat had attracted 10,000 users and today there are over 130,000 subscribers of which about 70,000 are voice customers. Other geostationary (GEO) MSS systems have been built for North America (AMSC and TMI), Mexico (Solidaridad), Australia (Optis), Europe, and Japan. Each of these systems has a very small number of voice users. Inmarsat carries most of the voice service.

In 1995, several new constellations of Low Earth Orbit (LEO) MSS satellites began construction. These included Iridium, Globalstar, and ICO Global. Iridium started operational service in late 1998 and the other two will begin service in late 1999 or early 2000. All are 3.5 years later than originally planned. During 1998, both the Ellipso and Constellation started construction of new MSS systems for operation in 2001. ACeS of Indonesia will launch a GEO satellite in 1999 and Thuraya will launch a GEO satellite in 2000. Qualcomm has developed a major space based data service. The system is called Omnitracs in North America and Euteltracs in Europe. It has nearly 200,000 subscribers worldwide, but the traffic volume and revenues are much smaller than for Inmarsat. Orbcomm has initiated a global data service using a constellation of LEO satellites for messaging, remote control, and data collection. The Orbcomm system has the capacity of less than 100 voice circuits. Several other LEO systems are in the planning stage.

Since the data services have been at very low rates, a much larger share of the resources and revenues are related to voice services. The MSS voice services are 100 to 1000 times larger than the data services and have been growing rapidly, at rates of 30% to 60% per year. We expect that growth to continue. There is a wide variety of opinions on future growth patterns for MSS.

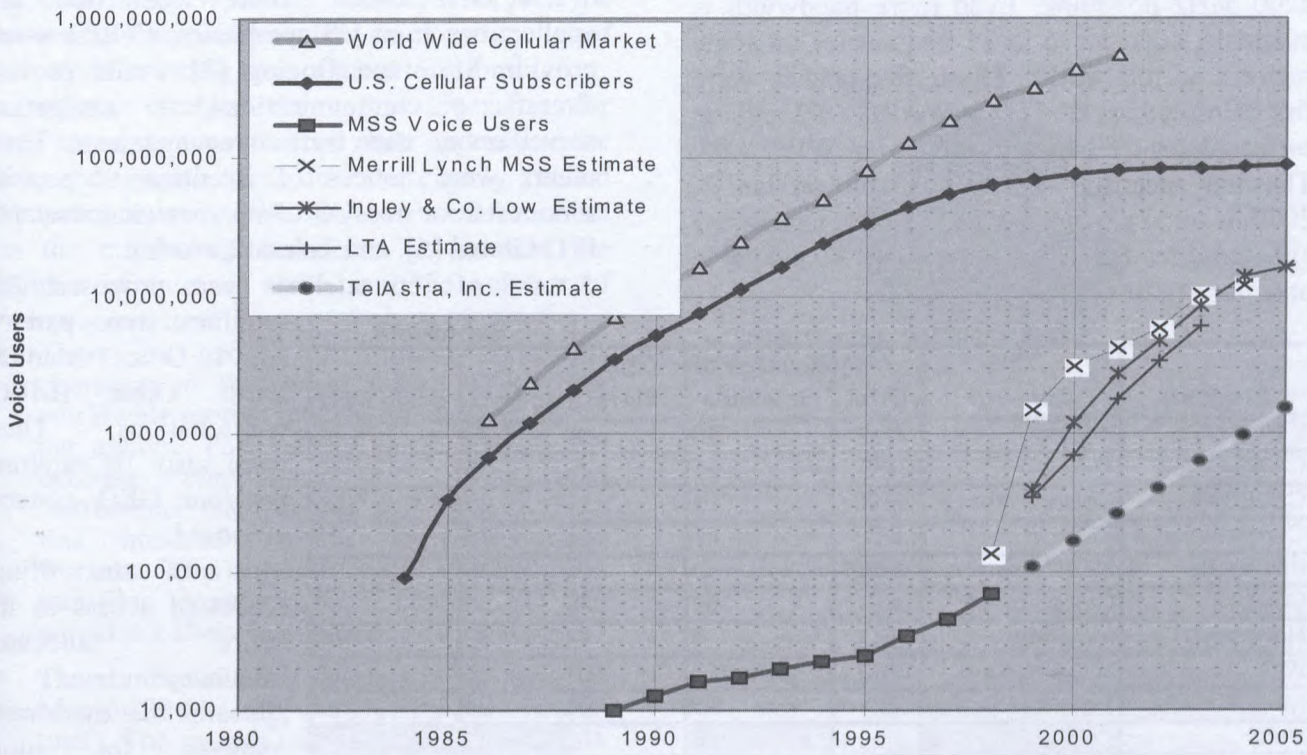


Figure 1. Growth of Mobile Voice Services

Figure 1 shows the explosive growth in mobile telephone services. The trend lines show past growth and future projections for global cellular, US cellular, and the space based voice services. The most recent optimistic projections have been prepared for the latest generation MSS voice systems like Iridium, Globalstar, and ICO Global. Several market analysts expect that there will be high growth (an inflection point) when the new LEO MSS services are introduced. We do not agree with these optimistic projections.

Promoters of new technologies sometimes predict demand that is inconsistent with past experience or past trends. The problem with large estimates for MSS demand is that they seem to disregard elasticity of demand and the experience of other global, space based services. Iridium, Globalstar, and ICO services are extremely expensive compared to terrestrial cellular.

The excessive projections are based on comparisons to the success of terrestrial cellular. These projections are based on securing a small percent-age of the very large cellular market. However, satellite based mobile voice service has been available for 20 years from Inmarsat. Satellite based mobile voice services have grown slower than cellular and have only captured 0.03% of the cellular market.

The Iridium terminals are smaller than the Inmarsat Mini-M terminals, but the Inmarsat terminals are less expensive, and the Inmarsat airtime prices are lower than for Iridium in most cases. The Iridium terminal costs ten times as much as a terrestrial terminal, and the airtime prices are about five times as high. How could we expect that the MSS market share will grow dramatically faster than Inmarsat or cellular service has grown? It is impossible to believe that Iridium will attract 335 times as many subscribers as Inmarsat within eight years.

THE NEXT GENERATION

Several systems are planning for the next generation of space based mobile services. An obvious trend is to provide a larger number of voice circuits. This will require additional bandwidth. Fortunately, the ITU has made global provision for additional MSS spectrum in the range of 1980 to 2010 MHz uplink and 2170 to 2200 MHz downlink. Even more bandwidth is available, adjacent to these frequencies, in some regions of the world. These frequencies were first allocated by the ITU at WARC – 92 and the service dates were moved forward at WRC – 95. The new spectrum will be available on Jan. 1, 2000 in some parts of the world. In other regions and countries the spectrum will not be available until Jan. 1, 2005.

					Bandwidth	
Program	Sponsor	Orbit	Number of Satellites	GEO Slots	up MHz	down MHz
2 GHz MSS	Boeing	MEO	16	0	8.25	8.85
CANSAT M3	TMI	GEO	1	1	35	40
CELLSAT	Celsat America	GEO	1	1	25	25
Constellation II	CCI	LEO	46	0	45	35
Ellipso 2G	MCHI	LEO	26	0	35	35
GS-2	Globalstar LP	LGEO	64	4	35	35
Horizons	Inmarsat	GEO	4	4	45	40
ICO	ICO Services	MEO	12	0	35	35
MACROCELL	Iridium LLC	LEO	96	0	35	35
Total	9		266	10	298.25	288.85

Figure 2. FCC Filings at 2 GHz

Inmarsat and ICO Global have been the primary advocates of the new spectrum at 2 GHz. ICO Global is building a fleet of 12 MSS satellites which will operate at 2 GHz from Medium Earth Orbit (MEO). The first ICO satellite launch is scheduled for the middle of 1999. Inmarsat and ICO took a big risk in selecting the new bands that had not been approved for satellite use anywhere in the world. These bands are currently used for tropo-scatter in the UK, for radar in the Middle East, for microwave links in Italy and South America, for military communications in Germany, and for news gathering in the United States. One major issue in the US is the cost of displacing existing services. Current US policy dictates that the new service will compensate the old service for displacement. This could be a major cost of entry for new MSS systems.

The FCC requested filings for the new band to be submitted by Sept. 26, 1997. By that date there were nine applications which are listed in **Figure 2**. Several applications [1] [2] were submitted by four firms that already had received licenses to provide MSS in the 1.6 GHz / 2.4 GHz bands. There were two new applications from US companies; Cellsat would provide MSS and Boeing [3] would provide aircraft communications, navigation, surveillance, and traffic management. Three others were notices of intention to provide services from non-US companies (Inmarsat [4], ICO Global [5], and Telesat Canada).

A total of 266 satellites were proposed, 256 satellites were part of six LEO or Medium Earth Orbit (MEO) constellations. There were also 10 satellites for four GEO systems proposed.

Most of the filings requested access to the entire allocated frequency band. It appeared that there were requests for more spectrum than had been designated by the ITU for MSS. It looked like the FCC would have some difficulty resolving the overlapping applications and assigning spectrum. Fortunately, there were hints in the filings that each company would be willing to share the total spectrum with other systems. Several companies proposed both CDMA and TDMA access methods to increase the prospects for a compromise in coordination.

In the future, more spectrum will be needed. The mobile satellite community should long term planning at the national and international levels. One possibility would be the conversion of the existing C-band FSS frequencies to MSS use.

HIGHER RATE DATA SERVICES

For many systems, the next generation of MSS will concentrate on higher capacity systems for

voice services. Examination of the filings shows that the new satellites and systems have substantially higher capacity for voice circuits.

Several operators are interested is developing and producing "medium rate" mobile terminals which would provide communication service through geostationary L-band or S-band satellites. The service would be rolled out on a regional basis. It would provide internet type access for business travelers or circuit switched video conferencing from remote locations.

In order to establish the market dimensions for high speed data services we must understand the size and growth rates for the closely related computer industry. We have some information on the number of computers in service, the number of Internet users, and the number of laptop computers.

- There are about 320 million computers operating in the world today. The recent worldwide growth rate has been about 16.6% per annum. This rate is likely to taper off because computers are sophisticated instruments that provide information for a few hundred million knowledge-based workers. This could change if computers transition to entertainment centers, but that would not alter the conclusions of this paper.
- The number of US computers is between 100 M and 120 M (30% to 38% of the worldwide total). The growth rate has been 11.6% for a few years. This rate will also decline or the number of computers would greatly exceed the population of the US within a decade. Although this situation is conceivable, it is not significant for this paper. The US has dominated in the use of computers by absorbing about 38% of the computers produced and seems likely to be the largest computer consumer for a long time.
- About 39 M US households have personal computers, and many households have more than one computer. (The author's household has three frequently used PCs, one laptop used for travel, and two relatively obsolete, but functional computers, for a total of six computers in the household.) The remainder is used by schools and businesses. We could safely assume that there are about 40 M "personal computers" used for US business.
- The worldwide number of laptop computers is currently about 20 M and could grow to 60M by 2005 if the growth rate is 16.6% per annum. This growth rate seems to be a little optimistic considering the maturity of the US market and the economic difficulties in Asia and other parts of the world. Most of the lap-top computers are business related, primarily for mobile businessmen. It is important to recognize that there are significantly more computers than computer users.

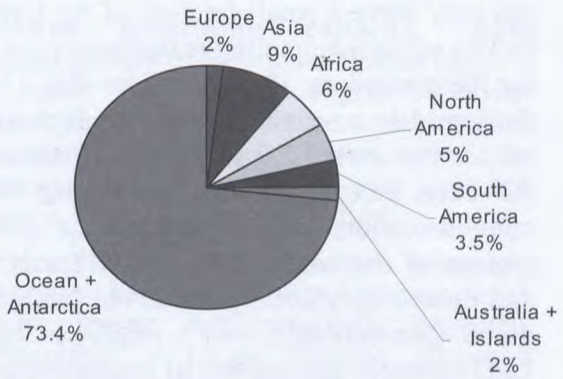


Figure 3. Surface Areas of Earth

- The computer industry is producing about 85 million computers per year. The average age of a computer is 3.15 years. This means that part of the annual production is used for replacements and part is for new users.
- Less than 15% of household computers had modems in 1995. No doubt the number is much higher today. There are about 100M Internet users (31% of computers). We think that at least 50% of laptop computers have modems.

This background information suggests that there is likely to be a niche market for business travelers, if the price for the service and terminal are affordable.

DISTINCTIVE CONSTELLATIONS

Most analysts have been surprised by the dramatic changes in communications satellite design and production. For 30 years, nearly every communications satellite was launched into GEO because this orbit greatly reduced the capital cost (compared to large numbers of LEO satellites) and simplified in-orbit operations.

LEO constellations offer the potential advantage of lower propagation time delay. LEO satellites are much closer to the Earth than GEO satellites. The LEO satellites range from 700 km to 1400 km in altitude whereas the GEO satellite altitude is nearly 36,000 km. GEO satellites have an inherent propagation round trip propagation time of ¼ second.

However, there are major cost problems directly related to the LEO architecture. Each satellite can only serve a small fraction of the Earth (1% to 3%) while maintaining reasonable view angles for the customers. If we consider that 73.4% of the world is covered by water or is Antarctica, we see that most LEO satellites will not be used. An even more dramatic fact is that 90% of communications occurs between the developed regions of the world. The United States, Japan, and Europe only occupy 4% of the surface of the globe. Consequently, only a small percentage of LEO satellites are capable of producing revenue. The net effect is that the service prices for LEO services are not generally affordable.

At the same time the quality of service is generally lower because of the likely obstruction of signals by terrain, buildings and trees at low elevation angles. Reports by users of Iridium, which is capable of a 16 dB link margin, are generally unfavorable because of call dropouts.

Ellipso has developed a clever constellation using elliptical orbits that dwell over the major land areas in the Northern Hemisphere. This system requires fewer satellites than the LEO systems. There is a risk of radiation damage due to the severe ionization environment in the Van Allen Belts.

Similarly, ICO Global and Boeing have selected MEO orbits that provide much less propagation time delay than GEO satellites. Because of the higher altitudes, the satellites can serve large regions of the world and fewer satellites are required.

At least one of the proposals suggests a hybrid network of GEO and LEO satellites, communicating through Inter-Satellite Links (ISL).

We are convinced that GEO satellite solutions will remain the most cost-effective approach. The question remains: Will customers pay a premium price for lower delay satellite services?

NEW TECHNOLOGIES

The new generation of spacecraft generally embodies several new aspects:

- New satellite platforms optimized for lower cost or mass production,
- Onboard processors for call routing and beam switching,
- Phase array antennas rather than horns or parabolic reflectors, and
- Intersatellite links for communications between satellites.

All of these technologies incorporate significant risks. We will only comment on the first item.

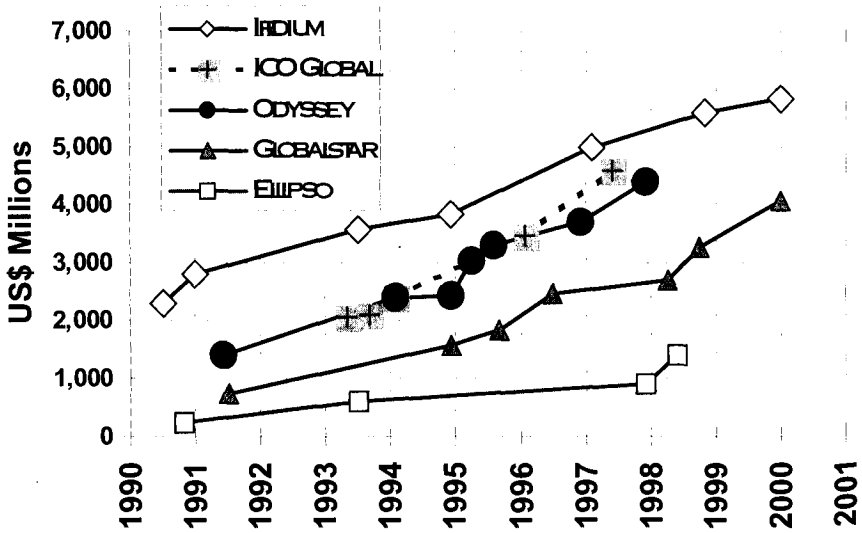


Figure 4. Cost Growth for Big LEO systems

Meticulous quality control and rigorous testing processes have been cut back or eliminated to reduce capital cost. For 30 years, satellites have used "high reliability" electronic parts that have been individually tested for extended periods to ensure reliable performance. Parts have been tracked for identification of defects. On some

new programs, the construction methods are based on mass production line processes and use of untested, commercial parts. Business and technical debates raged initially within the space industry when the proven methods were abandoned. Many challenged the wisdom of adopting concepts, which were discarded as not technically viable, expensive, and too risky. We are seeing very high failure rates in the Iridium constellation today as a consequence.

SERVICE COST COMPARISONS

The demand for terminals and services by satellite is highly elastic. Demand increases steeply as the prices are reduced. This assertion is based on both indirect evidence and market research. We have 20 years of data which compares the relative market shares and cost of service for space based and terrestrial mobile services. There is a sharp drop in demand when airtime prices are raised above $1 per minute in the US or $2 per minute in Europe.

Prices for all communications services are decreasing steadily. The continuous improvements in efficiency have led to price cuts that have stimulated demand and maintained high growth rates for more than a decade. We observe that the average monthly cellular telephone bill in America has dropped from $100 in 1988 to $39 in 1999. Inmarsat airtime rates for voice services dropped from $15 per minute in 1980 to about $2.50 per minute (with the Mini-M) in 1999.

Unfortunately, these FCC filings did not provide estimates of system cost. Therefore, we must use some approximate methods to estimate costs relative to the Big LEO satellites that are plotted in **Figure 4**. The important factors for determining service costs are:

- **Capital cost.** These costs include the cost for the satellites, launch vehicles, insurance, Earth stations, and financing costs. It is easy to underestimate these costs as illustrated in **Figure 4.** Here we see that each of the Big LEO systems grew in cost by a factor of two to four from the time that it was proposed.
- **Service provider costs.** It is important to remember that all of the revenues cannot be returned to the satellite operator. In practice, about 5/8 of the revenue is needed by the service provider for advertising, billing, distribution, servicing, and collections. Sometimes, the operator must provide financing for his customers to purchase the terminals. Satellite architects often underestimate this cost element.
- **Operational costs**. These costs include preparation of software to control the flow of communications traffic and to maintain orbit control. The cost of replacing failed satellites may also be included as an operating cost.
- **Revenue generating capacity**. Some systems greatly overestimate system capacity to make the system appear efficient. Capacity must be established under realistic operational conditions. This means the capacity that can be achieved with large link margins for multi-path and fading.
- **Service lifetime**. This element is extremely important because service take-up may require longer than originally expected. Furthermore, a longer lifetime allows a longer amortization of the capital investment.

In very simple terms, the airtime costs are directly related to the sum of the cost elements divided by the revenue generating capacity and service lifetime. We strongly believe that there is a continuing need to reduce the cost of service while increasing the quality of service. The obvious way to reduce the cost of service is to reduce the cost elements while increasing capacity and lifetime.

The next generation of satellites has greater capacity by about a factor of three, but the capital costs are also likely to be higher. It is not yet certain that these systems will yield affordable prices to the subscribers. One thing is becoming very clear: LEO systems are far too expensive.

FREQUENCY LICENSING AND ACCESS

The FCC has initiated a proceeding on the use of 2 GHz by satellites. On March 14, the FCC proposed that the 2GHz band should be split among the nine companies that have proposed use of this band (averaging less than 5 MHz each) The US has not allocated the bands in accordance with the ITU regulations. The US has given 10 MHz of the "Worldwide" MSS

uplink (1980 to 1990 MHz) to terrestrial PCS use.

The following is a summary of FCC proceeding IB 99-81 which was released on March 18, 1999. This Notice of Proposed Rulemaking would implement 2 GHz mobile satellite systems and establish service rules by amending existing Big LEO rules. The FCC proposes not to adopt financial criteria qualification in view of spectrum availability for all nine proposed systems. It will choose among four spectrum assignment options: "flexible band," "negotiated entry," and "traditional band," including auction, arrangements. Comments are requested for appropriate license terms, transponding capability requirements and the role of 2 GHz systems in rural service and incentives in system implementation milestones in exchange for rural service commitments. Comments also sought on orbital debris mitigation practices, out-of-band emissions requirements and international coordination with the European Union 2 GHz MSS rules.

CONCLUSIONS

If the MSS industry is to grow substantially, it must drive down costs to an affordable level. Fundamental design choices are crucial. More bandwidth and larger system capacity will help, but LEO systems are unaffordable.

REFERENCES

[1] In re Application of Globalstar, L.P. For Authority to Launch and Operate a Mobile-Satellite Service System in the 2 GHz Frequency Bands, September 26, 1997.

[2] Application of Iridium LLC for authority to launch and operate the MACROCELL Mobile Satellite System, 187-SAT-P/LA-97(96) September 26, 1997.

[3] The Boeing Company Application for Authority to Construct, Launch, and Operate a Non-Geosynchronous Satellite System in the 2 GHz Mobile-Satellite Serive and the Aeronautical Radionavigation-Satellite Service, 179-SAT-P/LA-97(16), September 26, 1997.

[4] Letter of Intent to provide Mobile Satellite Services to, from and within the United States by the Inmarsat fourth generation Mobile Satellite Services (MSS) systems, called hereafter 'Horizons', in the 2 GHz MSS Bands. 190-SAT-LOI-97(4), September 26, 1997.

[5] ICO Services Limited Letter of Intent to Access 2GHz MSS Frequency Bands at 1990-2025/2165-2200 MHz, 188-SAT-LOI-97, September 26, 1997.

The Accommodation of Spectrum Capacity for Mobile-Satellite Systems in the 1-3 GHz Frequency Range

by

Ali Shoamanesh, Robert Bowen, Gerard Kingsbury

(Telesat Canada)

ABSTRACT

This paper considers the options to mobile-system operators to individually and collectively obtain the necessary spectrum-orbit resources to carry out their business plans. This is becoming a serious real problem, given the number of systems currently in operation, under construction, and being planned, and the limited amount of spectrum allocated to the mobile-satellite service.

The problem is addressed here by first considering the evolution of mobile-satellite systems and their characteristics to date in bands below 3 GHz, briefly estimating the possible expansion of the mobile-satellite industry in the next decade, if sufficient spectrum-orbit resources could be found to accommodate that growth considering the possible availability of new mobile-satellite bands to meet that increasing requirement, and finally by examining possible alternatives available in the design of mobile-satellite systems to improve the utilization of the available spectrum-orbit resource.

2.0 EVOLUTION OF MOBILE-SATELLITE SYSTEMS TO DATE

Mobile satellite systems have been in operation about two decades. The initial geostationary L-band mobile satellite systems of Inmarsat were planned over twenty years ago, and have been in operation since the early 1980's. Since that time there has been implementation of regional and domestic geostationary mobile satellites, such as the MSAT systems of TMI and AMSC, the Asia Cellular Satellite System (ACeS), Asia-Pacific Mobile Telecommunications (APMT), and Thuraya in the 1525 to 1660.5 MHz portion of the spectrum global "Big-LEO" non-geostationary (NGSO) mobile satellites of Globalstar, Iridium, ICO, and others in portions of the 1610 – 1626.5 MHz, 1980 - 2025 MHz, 2170 – 2200 MHz, and 2483.5-2500 MHz portions of the spectrum, and global "Little-LEO" NGSO mobile satellites of Orbcom and Starsys in the 130-150 MHz portion of the spectrum.

As part of this development, user terminals have evolved from the relatively huge ship-borne terminals of Inmarsat in the early 1980's to vehicle-mounted terminals and more recently hand-held terminals of Globalstar, ICO, Iridium ACeS, APMT, Thuraya in the 1990's.

This development has taken place at a rapid pace, in parallel to development of terrestrial mobile systems, first as vehicle-mounted 800 MHz cellular systems in the 1980's to hand-held terminals at both 800 MHz and in the 1.8 GHz portion of the spectrum. Both first-generation cellular and second-generation PCS terrestrial mobile systems are now in operation, and third-generation IT-2000 systems are being planned.

These two areas of the provision of mobile communications are converging, in that several companies are or soon will be offering terminals that will provide service through either a mobile-satellite or a terrestrial mobile system, whichever provides the better response to a specific request for service, without any overt action required by the user except to specify the number to which he wishes to communicate.

These satellite systems are in part designed for and will function as satellite components of dual-mode mobile networks that provide mobile services to users through either their satellite component or their terrestrial component, whichever is more appropriate for each user. The uses of both mobile-satellite systems and terrestrial mobile systems are expanding rapidly. This expected expansion will require a similar expansion of the effective capacity of their available radio spectrum. It is this expansion of the effective capacity of the available mobile-satellite spectrum resource that is the topic of this paper.

3.0 MOBILE-SATELLITE SYSTEM CHARACTERISTICS

Any mobile-satellite system consists of one or more satellites, a large number of user terminals, and one or more gateway stations. Communication to and from the mobile user terminals is through the satellite or satellites to a gateway station and from that station either to the final destination through the public switched network or to another mobile-satellite user terminal through a second path through the satellite. That portion of the path between the user terminal and the satellite is usually in the 1 GHz to

3 GHz portion of the frequency band for systems that carry voice or data traffic, although the transmission of short alphanumeric messages is possible through Little-LEO satellites in the 130 to 460 MHz portion of the radio spectrum. The links between the satellite and the gateway stations is usually in the 4 GHz to 30 GHz SHF portion of the radio spectrum. There is also parallel development of both GSO and NGSO fixed-satellite systems to serve portable or transportable terminals carrying high-capacity multimedia messages in the 14/12 GHz and 30/20 GHz portions of the spectrum, but such systems are designed to provide a different service than the mobile-satellite systems being implemented in the 1 to 3 GHz band. This paper addresses only the availability and utilization of spectrum in the 1 to 3 GHz portion of the radio spectrum, the "service band" of current commercial mobile-satellite systems.

3.1 User Terminals of Current Commercial Mobile-Satellite Systems

Recently, user terminals have decreased in volume and weight until they are currently being implemented for use in a similar manner to current cellular or PCS mobile telephones. They are hand-held, operated in the same position as a mobile telephone, and stored in the user's pocket or purse. This is a giant step from fifteen years ago when they were permanently installed in a ship's radio room and mast.

This evolution has required, however, that mobile-satellite systems be designed and operated differently. System limitations imposed by the terminal design include: omni-directional user terminal antennas: one cannot ask the user to know where the satellite is and to point the antenna towards the satellite, nor can one afford the volume, weight, and cost of high-gain self-directing user terminal antennas; and limited power and energy of the user terminal transmitter, limited by both the battery size of the user terminal and the need to limit the EIRP of the terminal transmitter for health reasons.

In terms of mobile-satellite system design, an important characteristic of the system is the pre-detection carrier-to-noise ratio C/N of the link, either from the satellite to the user terminal or from the user terminal to the satellite. This term C/N can be described by the expression

$$C/N = \{EIRP + G/T\} - 20 \log (d) + \text{other terms} \qquad (1)$$

where EIRP is the effective isotropic radiated power of the transmitter, in either the user terminal or the satellite, G/T is the gain to noise temperature of the receiver, in either location, and "d" is the distance between the transmitter and the receiver.

The point here is that the size, weight, and health requirements of the user terminal puts severe limitations on the both the EIRP and the G/T of the user terminals. This means that the required sum {EIRP + G/T} has to be made up by more powerful and more sensitive satellites than used in the early days of huge terminals bolted down in ships. The other term, 20 log (d), simply tells us that these satellite parameters must be even better when the satellite is approximately 36,000 km away, depending on the latitude and relative longitude of the satellite and the user terminal, rather than at low-Earth orbits 800 km to 1400 km in altitude of low-Earth-orbiting satellites.

3.2 Mobile-Satellite Space Station Characteristics

The three important characteristics of mobile-satellites from the perspective of this paper are: the orbits in which they are placed; the gain of their antennas, the resulting frequency-reuse factor achieved, and the number of separate antenna beams that are simultaneously used; and the radio-frequency power transmission capability of the satellite.

These characteristics each have a major role in determining the telecommunications capacity of the satellite system.

3.2.1 Mobile-Satellite Orbits: Mobile-satellites may be placed in any one of four types of orbit: a geostationary orbit (GSO), approximately 36, 000 km above the earth, a circular low-Earth orbit (LEO), with altitudes ranging from 800 km to 1400 km above the Earth's surface; a similar so-called mid-Earth orbit (MEO), a circular orbit in the order of 10,000 km above the Earth's surface; and a highly elliptical orbit with a perigee only a few hundred kilometres and an apogee beyond the 36,000 GSO altitude. There are advantages and disadvantages to each type:

GSO Orbits: A single satellite in a GSO orbit can theoretically serve over a third of the Earth's surface. Inmarsat, in fact, can serve ships of the Earth's oceans from three orbital positions, and regional GSO mobile-satellite systems (ACeS, APMT, etc.) plan to provide service to a large portion of the earth's surface from a single orbit location. It should be noted, however, that a major spectrum-conservation advantage of the GSO, the ability of high-gain antennas to point to a GSO satellite continuously without any dynamic tracking capability, as applied in the implementation of fixed-satellite systems (FSS) and broadcasting-satellite systems (BSS), is lost when applied to mobile-satellites in the 1-3 GHz band because the user terminals of such systems have omni antennas (see above) and so no frequency reuse is possible between two visible GSO satellites.

LEO Orbits: It is common to implement large constellations of LEO satellites to provide mobile-satellite service over the whole globe. As examples, the Globalstar system uses 48 satellites at a 1400-km altitude, and the Iridium system uses 66 satellites in a lower-altitude orbit.

At lower altitudes more satellites are required, because the Earth's horizon is closer at a lower altitude, requiring more satellites to provide complete Earth coverage. The orbit geometry of these systems is quite complex; it involves a number of satellites in a given orbital "ring" around the Earth, a number of "rings" at different relative longitudes, and a certain "inclination" of each of these orbital rings. These orbits are designed to provide continuous service by at least one satellite anywhere on the Earth's surface.

MEO Orbits: An example of a satellite constellation in a MEO orbit is the ICO system. This orbital constellation is chosen so that the Earth's surface can be served from much fewer satellites. In comparison with a satellite system in a LEO orbit, however, the distance "d" is much larger, so a considerably larger {EIRP + G/T} product is required to provide the same level of service. This is important, as we will see later, when an increased {EIRP + G/T} can be used alternatively to provide increased telecommunications capability from a given orbit.

Highly-Elliptical Orbits: These orbits are also being developed for MSS applications, including the planned Ellipsat system. A highly elliptical orbit is one in which a satellite "ring" is highly inclined and has a low perigee only a few hundred kilometres high but an apogee beyond the GSO altitude. A satellite spends most of its time close to apogee when in such an orbit. Three satellites in such an orbit can provide continuous service to a large portion of the Earth surface, including the portion near the poles. Three such "rings" may be necessary to provide continuous service in the Northern Hemisphere, and another three rings necessary to provide continuous service in the Southern Hemisphere.

3.2.2 Satellite Antenna Gains: First-generation geostationary mobile satellites used simple wide-angle beams, with beam-widths about 17 °. Current-generation geostationary mobile-satellites in operation use several beams, primarily to increase the EIRP and the G/T of the satellite. In contrast, systems under construction use a hundred or more beams to increase their frequency-reuse factor as well. The ACeS system, for example, uses this technique to full advantage.

First-generation non-geostationary mobile satellites use a multiple-beam antenna system both to increase the satellite's EIRP and G/T, and to obtain significant frequency re-use. The Globalstar system, for instance, has a 16-beam antenna system; other non-geostationary systems have similar antenna structures. It may be noted that it is easier to implement a multiple-beam antenna system in a LEO orbit than in a GSO orbit, because from GSO the visible service arc is about 17° wide, whereas from a LEO location it is about 160° wide.

3.2.3 Satellite Power Limitations and The Implications of that Limitation. Because of the need to use an omni receiving antenna in the user terminals, the EIRP per voice channel has to be high if the system {EIRP + G/T} is to be large enough to receive the signal at the user terminal with sufficient quality. This can be done in three ways:

1) by employing high-gain narrow-beam transmitting antennas;

2) by generating enough radio-frequency power in the satellite, limited by the capability of the satellite platform; and

3) by using robust wide-band digital FM, BPSK or QPSK modulation to transmit the intended information. In some cases the detection of the signal is enhanced through the use of code-division multiplex systems to increase the bandwidth and so lower the required power of the transmitted signal.

4.0 MEANS OF INCREASING MSS SPECTRUM AVAILABILITY

Section 4.2.6 of the Conference Preparatory Meeting (CPM) Report to the 1997 World Radiocommunication Conference (WRC-97) indicates that by the year 2010 there will be a requirement for about 250 MHz of mobile-satellite spectrum in each direction in the 1 to 3 GHz frequency range, if conservatively-designed current-generation systems are used. This requirement is consistent with the trends outlined in Section 1 above that mobile-satellite systems will provide the satellite component of rapidly expanding personal mobile services in the 1-3 GHz band.

More recent estimates developed by Working Party 8D of the ITU-R suggests that based on conservative mobile-satellite traffic-growth estimates and the application of more recent realistic MSS spectrum-efficiency techniques, spectrum requirements will vary between 125 MHz and 145 MHz in each of the uplink and the downlink by the year 2010. This requirement is described in the current ITU-R Temporary Document 8D/TEMP/94 (Rev. 1).

There are basically two ways of meeting this requirement, or a combination of these two ways:

1) the allocation of new mobile-satellite bands by a WRC of the ITU. This is likely to require the ability of mobile-satellite systems to share spectrum with other radiocommunication services; and

2) use of available frequency bands more efficiently, possibly through the use of multiple higher-gain satellite antennas to achieve greater frequency reuse by a single satellite system, higher satellite powers and so greater frequency reuse by a given satellite system, through the use of narrower-bandwidth modulation techniques, more efficient source encoding techniques, signal processing in

the spacecraft to eliminate the addition of uplink and downlink powers in a power-limited system, sharing of the spectrum by multiple mobile-satellite systems, and new satellite orbits (This option is included here for completeness, although it offers very limited frequency-reuse potential, as seen below.)

4.1 The Allocation of New Mobile-Satellite Frequency Bands

Allocation of new mobile-satellite frequency bands is the most direct way to meet the anticipated increase in mobile-satellite spectrum requirements in the next decade. However, there has been limited success within the ITU in finding new mobile-satellite spectrum in the 1-3 GHz range. The primary bands used for GSO MSS systems are the 34 MHz wide bands 1525 to 1559 MHz downlink and 1626.5 to 1660.5 MHz uplink. Similarly, the bands used for current generation non-geostationary (NGSO) systems are:

1) the 16.5 MHz wide band pair 1610-1626.5 MHz uplink and 2483.5-2500 MHz downlink, being implemented by the Globalstar and Iridium systems;

2) portions of the 40 MHz wide band pair 1980-2025 MHz in the uplink and 2160-2200 MHz in the downlink, being implemented first by the ICO MSS system; (The term "portions of" must be used in this case because the allocations are not the same in all three ITU Regions, a distinct disadvantage if the bands are to be used by global NGSO systems.) and

3) the 20 MHz wide band 2500-2520 MHz downlink and 2670-2690 MHz uplink, a band which although it has been allocated on a global basis has not been implemented because of serious difficulties in sharing with other co-primary radio services.

There were efforts at two recent WRC's to harmonise the 2 GHz band in all three ITU Regions, and allocation of a portion of the 15 MHz wide band 1675-1690 MHz uplink to be paired with a similar downlink band, but the ITU has to date made very little progress in these efforts. The problems encountered are inter-service sharing between future mobile-satellite systems and currently implemented fixed, mobile, and meteorological-aids systems.

The allocation of spectrum to the mobile-satellite service in the 1-3 GHz portion of the radio spectrum is not an agenda item of the next WRC in the year 2000. The only WRC-2000 agenda items on the subject are

1. **Agenda Item 1.6.1:** review of spectrum and regulatory issues for advanced mobile applications in the context of IMT-2000, noting that there is an urgent need to provide more spectrum for the terrestrial component of such applications and that priority should been given to terrestrial mobile spectrum needs, and adjustments to the Table of Frequency Allocations as necessary;

2. **Agenda Item 1.9**: consideration of the allocation of the mobile-satellite (space-to-Earth) service in the band 1559-1567 MHz, taking into account relevant ITU-R sharing studies; and

3. **Agenda Item 1.10**: to re-consider the changes made at WRC-97 in the frequency bands 1525-1559 MHz and 1626.5-1660.5 MHz which replaced allocations to the maritime mobile-satellite service and the aeronautical mobile-satellite service to the generic mobile-satellite service.

An additional agenda item 1.11 continues the search for additional spectrum for Little-LEO mobile-satellite spectrum below 1 GHz, but does not address the need for additional mobile-satellite spectrum above 1 GHz. This is not because there is not a perceived need for additional mobile-satellite spectrum in the 1-3 GHz band, but because the subject was discussed at length at WARC-92, at WRC-95, and at WRC-97. In drawing up the agenda for WRC-2000, WRC-97 evidently concluded that other matters required ITU attention with a higher priority for consideration by a WRC in the 1998-1999 interval. The subject is also not on the preliminary draft agenda of the next WRC, in the year 2001 or 2002. However, that could be revised by WRC-2000 if there is new information to substantiate the need for additional spectrum and if there are indications that consideration of the subject might be successful in finding additional spectrum.

Even if all of the above bands were fully allocated and could be used effectively, their total bandwidth is only 110.5 MHz in each direction. To meet the CPM-97 estimate of 250 MHz of mobile-satellite spectrum being required by 2010, every available spectrum-enhancement technique will have to be applied, in some cases resulting in considerable additional cost to the implementation of mobile-satellite systems. These techniques are briefly considered below in Section 4.2.

ITU-R Working Party 8D is considering the use of possible new MSS bands, even though the allocation of such bands will not be considered by a WRC for at least four years. Such new bands include possible use of the 1675-1690 MHz band in the Earth-to-space direction. This band was considered by WRC-95 and by WRC-97, but the problems of sharing with the Meteorological Aids and Meteorological-satellite services were not resolved at those Conferences. Possible use of the band 1660.5 -1668.4 MHz in the Earth-to-space direction. In this band the sharing of MSS systems and radio astronomy sites has to be resolved. Possible use of the band 2475-2483.5 MHz band in the space-to-Earth direction, as an extension of the existing 2483.5-2500 MHz band. This use would require the sharing of MSS systems with un-licensed fixed

systems. ITU-R Working Party 8D is considering the possible use of this band by MSS systems outside of the urban areas where most of the unlicensed fixed systems are located.

These are options for future new bands. In any case, mobile-satellite operators will have to demonstrate that the most effective utilization possible is being made of existing bands. That is the topic of Section 4.2 below.

4.2 Techniques to Increase the Utilization of Available Spectrum

4.2.1 Use of Multiple High-Gain Satellite Antennas. All of the antenna discrimination used for isolation between different mobile-satellite systems and for frequency re-use within a given mobile-satellite system has to be located in the satellite, rather than in the user terminal, because of the size and convenience constraints on the design of the user terminal (see Section 3.1 above). There are definite trends in that direction in current-generation mobile-satellite systems, but the use of even narrower spot beams, the result of using larger antennas on the satellite, would both increase the number of times a given satellite could re-use the same block of spectrum, and also increase the satellite's EIRP and G/T. This is potentially the greatest improvement in spectrum utilization by mobile-satellite systems.

As it was stated above, large deployable reflector antennas have been instrumental in meeting the demanding mobile satellite communication systems requirements of high gain and multiple beams. Starting with the Canadian MSAT system, antennas have increased in size from 6m to 15m with the EAST program. With very little spectrum available (34 MHz in the 1.5-1.6 GHz range and 35 MHz in the 2.0-2.2 GHz range) it is essential to achieve a high level of frequency reuse. Of the three main geosynchronous mobile communications systems to be deployed within the next 3 years, all propose a large number of narrow beams, from 140 to 250 beams with beam radii of 0.5° and 0.7° respectively at L-Band frequency. It has been claimed [1] that a design with 40-m antennas can be integrated with communications systems that are launched in medium launch vehicle. The beamwidth of a 40-m antenna at L-Band and S-Band would be about of 0.3° and 0.2° respectively.

Such antenna reflector requires being lightweight, rigid, easily deployable and free of passive inter-modulation (PIM). To meet those stringent requirements, existing designs use a highly reflective gold-plated molybdenum mesh stretched on bonded graphite composite over an aluminium truss structure. To mitigate the risk of PIMs, a dual-antenna design such as ACeS may be used, or in the case of single reflector, the reflector itself along with boom structure are designed to be PIM-free and non-charging. In the dual-antenna design, beam congruency could present a challenge, especially when the beam diameter becomes very small.

To complete the multiple-beam antenna, a phased array feed, in the offset position, provides the multiple beam

function. At present two main designs are being deployed: the passive beamformer with hybrid matrix power amplifiers generating fixed beams (ACeS design [2]); and the digital beam former along with dedicated SSPAs per feed element, also known as direct radiating array feed, forming very flexible beams with power re-allocation capability. Each design requires key technologies to be developed in order to meet the requirements, the antenna reflector being certainly one of them.

4.2.2 Higher Satellite EIRP and G/T Values. As suggested above, higher satellite EIRP's, and a corresponding increase in satellite G/T's, has the advantage of allowing higher order modulation schemes as well as the frequency reuse associated with more antenna beams. The higher link {EIRP + G/T} has to come from improvements in the spacecraft performance, because the need for omni antennas and low transmitter powers in the hand-held user terminals puts severe limitations on these improvements coming from re-design of the user terminals. If such improvements in the spacecraft can be realised, use of higher-order modulations would improve the spectrum utilization. The use of modulation formats such as 8 ϕ PSK or 16 QAM, for instance, would allow higher data rate than current QPSK systems, and so greater effective use of the available spectrum.

It should be noted here that use of higher power levels and more antenna beams may increase the intra-system and inter-system interference levels. It may be necessary, as a consequence, to operate the systems with a higher I/N ratio than practised in current-generation systems, but at higher power levels with the same or increased pre-detection C/(N + I) ratios as a result of significantly higher pre-detection carrier levels. Note that the use of a larger number of higher-gain antenna beams will by itself increase the EIRP in a given antenna beam, but the correspondingly larger number of beams will require greater rf power to be generated in the satellite even if the rf power per beam is not increased. Thus the ability to generate higher levels of rf power, and so have greater dc power-generating capability, is an essential component of the evolution to satellite systems with higher utilization of the available spectrum.

4.2.3 More Efficient Source-Encoding Techniques. There has already been considerable progress in this area. The digital rates required to transmit a voice signal have been reduced from the conventional telephony 64 kbps to a 2.4 kbps rate. Significant further improvement in this area cannot be expected.

Source-encoding can be expected to provide improvements in the transmission of video or "multi-media" signals. If duplex video communication were to be offered through mobile satellites, significant further improvements would be required. This improvement may be forthcoming, however, as a result of a similar need if duplex video communication is to be provided through similarly bandwidth-limited terrestrial mobile systems.

4.2.4 Signal Processing in the Satellite. Simpler mobile-satellite spacecraft are simply "bent pipes", i.e. are simply amplifiers and frequency translators of the desired signals. In such systems the noise and interference power in the uplink and the downlink simply add. If, however, the uplink signal is processed in the satellite before re-transmission downwards, it is the error probabilities that add, not the noise powers. This results in up to 3 dB improvement in the performance of the system, without increasing the EIRP or the G/T of any of the components of the system.

This technique is already being applied in the Iridium system, for instance, primarily to enable routing of user traffic between satellites, but will have the added improvement of a lower error rate than if the same system acted as a "bent pipe". Other systems may to follow the same route with later generations of their equipment. All other systems being equal, this will allow "softer" modulation such as 8 ϕ PSK rather than QPSK, with the corresponding increase in spectrum utilization.

Deployment of more complex on-board signal-processing satellite systems will provide increased capability and flexibility in the operation of such systems, and in so doing will improve the utilization of the available spectrum. Different types of non-regenerative digital signal processors will be deployed shortly as part of the ACeS [2] and ICO [7] programs. Each of these will offer a multiple beam forming and switching function, which allows a large degree of frequency reuse to maximise system capacity and minimise the spectrum required. Frequency re-use consists of using the same frequency band several times through orthogonal polarization and in the case of multi-beam satellite through isolation resulting from antenna directivity, to increase the total capacity of the network without increasing the allocated bandwidth.

ACeS, a geostationary satellite, only needs a fixed cellular beam pattern. An analog low-level beam forming network, using simple passive components, and an on-board digital switch provide dynamic routing of individual frequency sub-bands to any beam and frequency slot. The L-Band coverage has 140 beams, which are formed by separate transmit and receive 12-m antenna subsystems. This multi-beam configuration can support a frequency reuse factor of 20 with a 7-cell frequency reuse pattern.

The L-Band feed-arrays and reflectors are identical for both transmit and receive. Each beam is formed by the low-level beam forming network which provides the amplitude and phase weighting. Signals are then presented to the multiport power amplifiers and the transmit antenna feed assembly. The role of the 88 L-Band cup-dipole radiator feed [4] array is to receive a pre-determined distribution of RF signals from the matrix power amplifiers and to illuminate the mesh reflectors to form 140 spot beams. The feed elements are shared between beams and the power is shared between multiport power amplifiers. This enables a high degree of power distribution flexibility between beams to accommodate traffic variation among beams while minimising the number of feed elements and power amplifiers [5]. The antennas' deployable reflectors, made of gold-plated molybdenum, provide the gain for communication links to handheld phones 40,000 km away [6].

In contrast, in the ICO non-geostationary satellite system, utilising 10 satellites in two medium earth orbits, a more flexible digital signal-processing technique is needed. The ICO on-board processor provides digital channelization and beam forming in the frequency domain. This allows creating 490 channels 170 kHz wide to be routed to any of the 163 beams at any frequency slots within the S-Band 30 MHz bandwidth used by the ICO system. This channel to beam routing can be adapted continuously to changes in traffic and interference on demand, simplifying the frequency coordination of the network.

More advanced regenerative on-board processors are planned for future systems, offering further improvements in utilization of the allocated spectrum. In such systems received radio-frequency signals are de-multiplexed and demodulated to their original baseband form, routed to the correct downlink beam, and then re-modulated and multiplexed, upconverted to the downlink carrier frequency, amplified, and transmitted. This demodulation and remodulation process offers the advantage of separating the uplink from the downlink, and in so doing reduces the overall link error rate performance. Assuming that both links have the same bit-error-rate (BER), a 3-dB improvement is realised over bent-pipe transponders. As stated earlier, this 3-dB improvement can be traded for a more bandwidth efficient modulations to further increase spectrum efficiency.

4.2.5 Sharing of Spectrum by Multiple Mobile-Satellite Systems. Sharing of spectrum by many different satellite systems is the normal process with fixed-satellite and broadcasting-satellite systems at higher frequencies. The primary mechanism used to isolate co-frequency FSS or BSS satellites at higher frequency bands, however, is a combination of polarization isolation and isolation due to a combination of earth-station antenna and satellite antenna isolations. Neither polarisation isolation nor earth-terminal antenna isolation is possible with mobile-satellite systems in the 1 to 3 GHz frequency range, however, because of the necessary omni-direction characteristics of the antenna of the user terminal (see Section 3.1 above).

One way in which different mobile-satellite networks might share use of the spectrum is if their respective co-frequency antenna footprints on the Earth's surface do not overlap. This would allow two different GSO satellites serving different service areas to share use of the spectrum from similar orbit positions. This sharing technique would be made easier if the satellite antenna complex was a large number of small beams, covering the exact service area of each system. This is an additional advantage of a complex spacecraft antenna structure, over that discussed in Sections 4.2.1 and 4.2.2 above.

Other than through spacecraft antenna isolation, there are limited opportunities for two satellite networks to share the spectrum when the two are visible from the same location. In Section 4.2 of the CPM-97 Report to WRC-97, it is explained that two mobile-satellite systems using time-division multiple-access (TDMA) or frequency-division multiple-access (FDMA), or one using either FDMA or TDMA and the other using code-division multiple-access (CDMA), cannot effectively share use of the radio spectrum. This is why the 1610-1626.5 MHz band is segmented between the Globalstar CDMA system and the Iridium TDMA system. The CPM-97 Report did indicate, however, that two CDMA systems can share the same spectrum, as long as the two systems employ relatively orthogonal CDMA codes. In practical terms, this is why the Globalstar system and the now-cancelled Odyssey system, both CDMA systems, could have shared the same spectrum. It should be pointed out, however, that the signals from one CDMA system imposes high levels of "noise" on the other system, reducing the signal-carrying capacity of both systems. For this reason, there are limited advantages to frequency sharing between two CDMA systems over segmenting the available band between the two systems.

4.2.6 Use of Different Types of Orbits By Different Mobile-Satellite Systems. As indicated in Section 3.2.1 above, mobile satellites in the 1-3 GHz band occupy or are being placed in several different types of orbit, including the classical geostationary orbit, low-Earth orbits, medium altitude or mid-Earth orbits, and highly elliptical orbits. However, again because of the omni-directional characteristics of mobile-satellite user terminals, a user terminal would transmit to or receive a signal from a satellite in any of these orbits. For that reason, putting two satellites in different orbits or even types of orbit does not result in their being able to share the radio spectrum beyond that discussed above in Section 3.2.5.

5.0 SUMMARY

We have seen that options for increasing the utilization of available spectrum by mobile-satellite systems in the 1-3 GHz frequency range is limited by the need for the user terminals to be small, convenient to use, and to not pose a health risk when used as a mobile telephone. Those constraints require that the user terminals have low EIRP's and omni antennas, characteristics that limit their ability to be modified to significantly increase their system's spectrum utilization.

At the same time, the rapid growth in demand for mobile services in general indicates that the current increases in implementation of mobile-satellite systems will continue in the coming decade. This increase will require a significant increase in the effective availability of radio spectrum to permit those systems to be implemented. There are, however, major difficulties within the ITU in the effort to allocate new frequency bands to the mobile-satellite service in the 1 to 3 GHz frequency range. These difficulties are centred around the problems of sharing spectrum between mobile-satellite systems and current fixed, mobile, radionavigation, and meteorological aids services.

With that as a background, the most promising alternatives for acquiring the necessary increase in effective bandwidth for an expanding mobile-satellite industry is in the design of larger and more complex spacecraft. If the forecast need for increases in the effective mobile-satellite spectrum are to be met, future generations of mobile-satellite spacecraft will have to evolve in even more sophisticated spacecraft.

6.0 REFERENCES:

[1] Mark W. Thomson, "The AstoMesh Deployable Reflector", International Mobile Satellite Conference 1997, June 16-18 1997, pp. 393-398.

[2] Stuart C. Taylor & Adi R. Adiwoso, "The Asia Cellular Satellite System", 16th AIAA International Communications Satellite Systems Conference, Washington, DC, 1996, pp. 1239-1249.

[3] J. Alexovich, L. Watson, A. Noerpel, and D. Roos, "The Hughes Geo-Mobile Satellite System", International Mobile Satellite Conference 1997, June 16-18 1997, pp. 159-165.

[4] M. Forest, S. Richard and C. A. McDonald, "ACeS Antenna Feed Arrays", International Mobile Satellite Conference 1997, June 16-18 1997, pp. 387-391.

[5] Nam P. Nguyen, Pedro A. Buhion Jr., and Adi Adiwoso, "The Asia Cellular Satellite System", International Mobile Satellite Conference 1997, June 16-18 1997, pp. 145-152.

[6] Barry Miller, "Satellites Free the Mobile Phone", IEEE Spectrum, March 1997, pp. 26-35.

[7] Fumio Makita and Keith Smith, "Design and Implementation of ICO System", 17th AIAA International Communications Satellite Systems Conference and Exhibit, 23-27 February 1998, pp. 57-65.

Mobile Satellite Data Communications and the Internet

David Dawe
Stratos Mobile Networks
P.O. Box 5933, St. John's, NF A1C 5X4, Canada
E-mail: dave_dawe@stratos.ca

ABSTRACT

Associated with any satellite communication is some form of terrestrial backhaul. In the case of satellite data communications, the landline component has traditionally been carried over telex, PSDN (Public Switched Data Network, normally X.25), or PSTN (Public Switched Telephone Network, using modem or fax).

Meanwhile, the popularity of the Internet continues to grow rapidly. Every day more and more people have access to, and are accessible via, the Internet. Today's mobile satellite service provider must tap into this medium, meaning that terrestrial access to its satellite data communication systems must include the Internet.

This paper discusses the importance of integrating on-demand mobile satellite data communication services with the Internet, and explores some technical issues and solutions associated with such integration – with a focus on store-and-forward messaging.

INTRODUCTION

Data and the Internet

In the terrestrial setting, voice communication has always dominated. However, the Internet is changing this. With the Internet centered around data communications, and the growing popularity of the Internet, data is taking on an ever increasing role in keeping people in contact. E-mail even appears to be more widespread than voice mail.

Not only is the Internet pervasive, but there is generally no incremental cost in using an existing access for additional applications – such as mobile satellite communications. Most users have the necessary infrastructure already in place. Overall, the Internet is much less expensive than the traditional networks, and location on the globe is no longer of much concern.

Data and Satellite Communications

Information technology plays an increasingly important part in today's business operations. Many companies using mobile satellite communications recognize the benefit of "extending their data networks" over the satellite – which is most easily accomplished by using the Internet. Accordingly, data is becoming more important in the satellite world, and the Internet is becoming the terrestrial backhaul of choice.

Most mobile satellite communications are characterized by on-demand, rather than dedicated, access. When satellite communication is required, a "call" is established. Once the information has been transferred, the call is cleared. This on-demand access best suits the needs of the typical user, where the relatively low volume of traffic does not justify the cost of dedicated access.

However, care must still be taken to keep down the cost of on-demand access. For example, browsing the web over a satellite call is impractical. Every minute spent "on-line" costs dollars, not pennies. In general, current satellite data communication facilities are not well suited to interactive access. The key is to employ on-demand access effectively, by establishing a call only when needed, and keeping it up only as long as necessary.

Store-and-forward Data

Store-and-forward data communication – as opposed to real-time data communication – involves the transfer of messages only. That is, information is always encapsulated into messages before transmission. In the case where the originating and destination systems do not directly communicate, a message travels through intermediate nodes along a path between the two systems. Each such message can be assigned characteristics (such as priority, expiry time, etc.) based upon its content. This permits intelligent routing, which includes identifying when and how each message should be forwarded.

When to Forward. The ability to identify when to forward a message is an important characteristic of a store-and-forward messaging system. This is particularly so for satellite communications, where the cost is much higher than in the terrestrial case, as shown in Table 1.

Based on the properties of a message, its forwarding can be deferred seconds, minutes, or in some cases hours. The longer a message can be held in queue, the more likely it can be sent cost effectively. By the time the call is made, other messages will likely have been queued, and they can be sent in the same call. Furthermore, once the call is established, the called system may have messages to send

in the return direction. The more messages that are transferred in the one call, the less overhead there is per message.

Table 1: Typical Data Communication Costs

Service	Speed	$US per minute	$US per kilobyte	Minimum $US
Inmarsat-B	9600 bps	$4.00	$0.06	$2.00
Inmarsat-C	n/a	n/a	$8.00	$0.25
Int'l PSTN	14400 bps	$0.80	$0.01	$0.80
Int'l PSDN	n/a	n/a	$0.20	$0.01

How to Forward. The problem of identifying how to forward a message can be considered when multiple networks are available for message transfer. This is often based upon the total length of the messages queued. Typically, the more data in the queue, the more cost effective a circuit-switched medium becomes. Alternatively, a particularly urgent message may be routed over the medium that can provide the quickest delivery time.

Benefits. In summary, store-and-forward messaging is a natural fit with on-demand communication access. It facilitates the efficient use of expensive satellite communication links, while permitting minimal delays for high priority messages.

Circuit switched and packet switched media can be used together in a hybrid system that chooses the medium best suited to each message transfer. In fact, a circuit switched medium is transformed from a means for potentially expensive interactive communications to a means for cost effective store-and-forward messaging.

An important aspect of store-and-forward communications is that each link along the message path behaves independently. That is, problems encountered over one link do not impact upon performance over other links.

Satellite Communications and the Internet

Mobile satellite service providers have traditionally marketed their data communication services as "pipes". The end-to-end communication path would typically be real-time in nature, with the terrestrial component carried over Telex, X.25 or PSTN.

However, the Internet can be prone to real-time communication delays. Internet delays are generally acceptable for terrestrial data communications where the costs imposed by such delays can be negligible or nil. In contrast, such delays cannot be tolerated for satellite data communications – where time is truly money.

So that the presence of Internet delays does not increase satellite communication costs, the actual satellite calls must be isolated from those delays. As illustrated in Figure 1, the solution is the insertion of a store-and-forward gateway between the Internet and the satellite network.

For example, there are cases (especially when using higher speed satellite services) where it can help to have a terrestrial-based "relay" FTP server, acting as a store-and-forward node. The server would be accessible to a satellite user without having to go through the Internet, and accessible to a terrestrial user without going over the satellite. Consequently, any Internet delays would not affect the length of the satellite call.

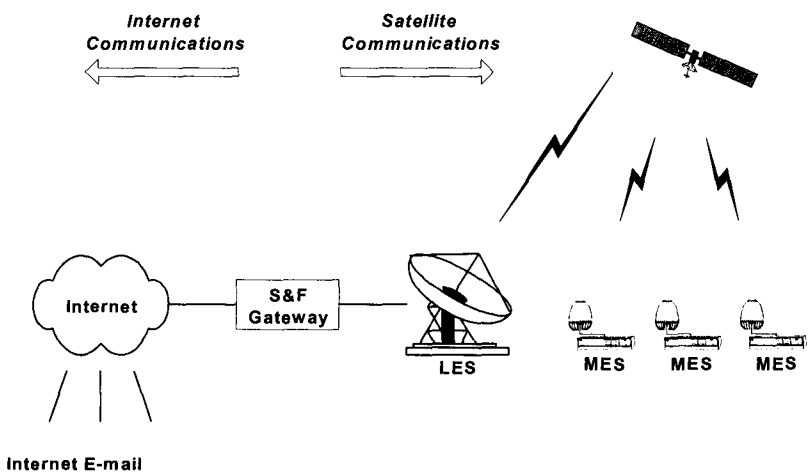


Figure 1: Insertion of a Store-and-forward Gateway

INTERNET E-MAIL GATEWAY

Stratos undertook development of the Internet E-mail Gateway and put it into service in early 1995, to provide Internet e-mail access to and from Inmarsat-C MESs (Mobile Earth Stations). It was the first gateway of its kind.

In the following, many of the features of the gateway are discussed below. Most of these features were originated by Stratos, and many of them remain exclusive to Stratos.

Inmarsat-C

Introduced in the early 90s, Inmarsat-C was the first digital data service offered by Inmarsat, and is, in itself, a store-and-forward messaging system. Originally conceived in the context of ensuring maritime safety, Inmarsat-C incorporates both conventional and broadcast messaging services. To address commercial interests, position reporting and group call services are also included.

An Inmarsat-C MES (Mobile Earth Station) includes a transceiver, omni-directional antenna, optional GPS receiver, and a PC running software to interface with the transceiver. Terrestrial access to Inmarsat-C is through any of about twenty LESs (Land Earth Stations) worldwide, via telex, PSDN, or PSTN – using either a two-stage or one-stage access method.

Two-stage terrestrial access. Because Inmarsat-C is a store-and-forward system, sending a message to an MES can be done by transferring the message to an LES accompanied by instructions for message delivery. The LES then attempts to transfer the message to the destination MES, and eventually reports message status back to the originator. This is known as two-stage access whereby an intermediate node (the LES) was explicitly involved in delivery of the message.

One-stage terrestrial access. Directly establishing a "call" to the destination MES is known as one-stage access. Although Inmarsat-C is a store-and-forward system, the terrestrial originator gets the impression of communicating directly with the MES. In fact, it is the LES that receives the terrestrial call, and identifies and emulates the dialed MES. After the message transmission is completed and the call is cleared, the LES then attempts the message transfer to the MES.

Although available for PSDN, one-stage access is most popular for telex. This arises from the fact that prior to Inmarsat-C, maritime satellite messaging most often involved sending telex messages over Inmarsat-A. One-stage access to Inmarsat-C behaves quite similarly to Inmarsat-A telex (a circuit switched medium). The emulation is effective enough that, to this day, many telex subscribers remain unaware that Inmarsat-C is a store-and-forward system.

Billing

An important aspect to using Inmarsat-C is the associated cost. It varies among service providers (and by location of the terrestrial endpoint) but a typical rate is $US 0.25 for every 32 bytes. A 10-line message, say about 500 characters, would cost $US 4.00. This means that e-mails need to be billed for – a notable break in Internet tradition.

Therefore, another important aspect is billing for the traffic. Accordingly, access to Inmarsat-C from the Internet requires that Internet e-mail addresses wishing to send messages to Inmarsat-C be registered. The registration process includes identification of billing information. Though not required for MES users, most MESs originating messages to the Internet through Stratos are registered. Then an organization has the option of consolidated billing for its outbound (to-MES) and inbound (from-MES) traffic.

Once registered, sending a message from the Internet to an Inmarsat-C MES is the same as sending any other e-mail. In this case the e-mail would be sent to an address along the lines of *mes-number*@stratosmobile.net, where *mes-number* is a 9-digit Inmarsat Mobile Number (IMN), beginning with "4". Optionally, to simplify use of the system, the MES can be assigned an "alias" so that *mes-number* is a more easily remembered "MES name".

Using the above model, each message is typically billed to its originator. However, there is the occasional MES that opts for reverse billing of outbound traffic. That is, the MES user pays for messages sent *to* it, not just *from* it. In rare circumstances, such an MES may also wish to allow any Internet e-mail user to send a message to it. Of course, this can be dangerous. Unexpected e-mails could be sent to such an MES, perhaps from a newsgroup, or worse still, from an "e-mail spammer". With the complexity of messaging permissions already in place, the ability to prevent messaging from such originators is challenging to resolve.

Attention to Cost

Part of the difficulty in integrating Internet e-mail with Inmarsat-C comes in the "inattention" to cost associated with many e-mail systems. Such systems do not consider the cost of communication links when forwarding e-mails.

For example, a typical RFC822 header [1] (Internet e-mail header) can easily be 800 bytes or longer. Transmission of such a header over Inmarsat-C would add more than $US 6.00 to the cost of a message. It becomes important to intelligently process the header, rather than just forwarding it blindly. The most important pieces of header information are the originating e-mail address (including the real name) and the message subject.

However, in some cases, an MES may exchange messages with only a limited number of Internet e-mail addresses. Therefore, even the smaller header may be unnecessary, and it may be desirable to suppress it. As well, headers can be suppressed on a per message basis.

In summary, the general idea is to reduce unnecessary costs due to transmitting header information over Inmarsat-C. Without such measures, messaging becomes significantly more expensive than it should be, and far less attractive to potential users.

Multi-part Messages

One of the most important conveniences of any e-mail system is the handling of attachments. In the context of Internet e-mail, this means the processing of MIME (Multi-part Internet Mail Extensions) formatted e-mails.

For integration with Inmarsat-C, this means processing outbound MIME messages so that they can be conveniently received by MESs without custom software, and encoding inbound messages into MIME where needed. Each attachment in an outbound MIME message is sent in its own Inmarsat-C message, so that each binary attachment appears in its own file at the MES. One exception is that adjacent textual attachments are

coalesced. While the cost benefit of coalescing such attachments is negligible, it does reduce the total number of messages and the total amount of time to transfer them.

MIME processing also does its share to help reduce Inmarsat-C costs. Of particular interest are "multi-part alternative" e-mails. Such messages, including those generated by popular Internet e-mail user agents from Netscape and Microsoft, transfer message content as both plain text *and* HTML (Hypertext Markup Language).

It is impossible to educate a large user community regarding the presence of multi-part alternative encoding, its effects on Inmarsat-C costs, and how to disable such encoding. Therefore, it is imperative to detect such multi-part alternative e-mails and suppress the HTML before transmission over Inmarsat-C. For example, a 1-kilobyte textual message, forwarded without message header and multi-part alternative processing, could cost three to four times more than without such processing.

There are also savings to be had in the transfer of binary attachments. Simply converting base64-encoded information to its binary equivalent results in a saving of 25%.

Delivery Notices

Background. A typical feature of store-and-forward messaging systems is the return of DNs (delivery notices) to message originators. An NDN (non-delivery notice) indicates failure to deliver a message, whereas a PDN (positive delivery notice) indicates success.

An NDN is generally considered more important than a PDN, because it is failure of a message that normally triggers an originator to take further action. In an environment where only PDNs are generated, the originator can never be certain whether a message (for which a PDN has not been returned) needs to be requeued. The store-and-forward system may still be attempting delivery, or the message may have expired and been discarded.

This is reflected in the nature of Internet e-mail DNs normally found. The portion of Internet e-mail systems that provide bounces (e-mail NDNs) appears close to 100%, while the portion supporting return receipts (e-mail PDNs) is typically just a little over 10%.

For Inbound Inmarsat-C messages. Again, the cost of Inmarsat-C becomes an important consideration. Even with the RFC822 header processed and reduced, a receipt message can still be very large. In fact, an e-mail DN often quotes some or all of the content of the referred-to message.

Therefore, while it would be simpler to just forward the content of the e-mail DN as-is to the MES – leaving it for human interpretation – economics prohibit this solution. Receipt messages, like RFC822 headers, should be processed and distilled down before being transmitted over Inmarsat-C.

Unfortunately, just distinguishing a normal message from an e-mail DN can be quite difficult. An e-mail DN is often more suited for human interpretation than automatic processing, and its format differs greatly among various mail systems. Furthermore, its content must be processed to identify the referred-to message, its delivery status, why it failed, etc.

With e-mail DN processing in place and well established, Stratos is able to properly return a DN to the MES. The DN is concise and cost effective, and the rules for shipment are consistent with general NDN/PDN forwarding. An NDN is returned unconditionally, whereas a PDN is returned only if requested by the message originator.

For Outbound Inmarsat-C Messages. There is no cost associated with returning an e-mail DN to the originator of the original message. Accordingly, Stratos always attempts to do so, whether the DN is a bounce or return receipt.

On the surface, this may appear to present very little complication. However, an e-mail DN generated toward the Internet may occasionally trigger a further DN in reply. With the help of the e-mail DN processing for inbound Inmarsat-C messages, such a DN must *not* be forwarded to an MES. This would have the potential of creating a "message loop". Hence, it becomes even more important to properly recognize and process e-mail DNs from the Internet.

Traffic Levels

Perhaps the best measure of success is customer acceptance. Although traffic growth was limited through to 1996, it has been strong from 1997 onward. Figure 2 illustrates the overall trend of Inmarsat-C / Internet e-mail traffic through Stratos since the gateway was put into operation in 1995.

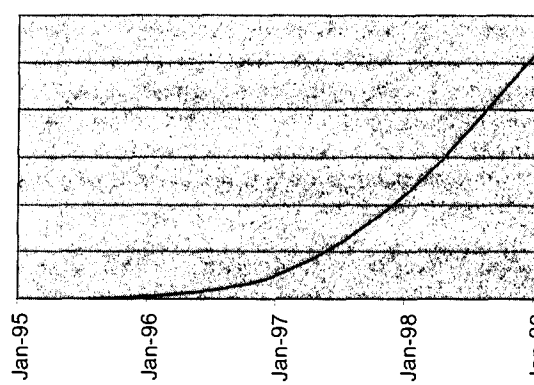


Figure 2: Inmarsat-C / Internet E-mail Traffic Trend

Inmarsat-A / Inmarsat-B Telex

More recently, support for Inmarsat-A and Inmarsat-B telex has been added to the Internet E-mail Gateway. This means that an Internet e-mail user that is registered for Inmarsat-C access can now also exchange messages with Inmarsat-A/-B MESs over telex. Although this may not appear to be a step forward, telex remains the most commonly accepted and used maritime data communication system.

Outbound messaging. For the Internet e-mail user, sending a message to an Inmarsat-A/-B MES appears the same as for Inmarsat-C. The only difference is that the *mes-number* (of the *mes-number*@stratosmobile.net e-mail address), if numeric, is in a different range. Even the underlying implementation is relatively straightforward.

Inbound messaging. For the Inmarsat-A/-B telex user, sending a message to an Internet e-mail address uses secondary addressing – whereby the address information is placed in the first few lines of message text – modeled after the technique used for Inmarsat-C.

However, there are complications, the most fundamental of which is the 5-bit ITA2 (baudot) character set used for telex. Many characters that can appear in an e-mail address are not available in ITA2. This was circumvented by defining various encoding rules. Perhaps the most notable such rule is that the sequence "(A)" in a secondary address is converted to "@" in the e-mail address.

Another carryover from Inmarsat-C secondary addressing is that multiple address can be specified. This permits one message (with one satellite space-segment charge) to be delivered to multiple Internet e-mail addresses.

Integration of Internet e-mail with Inmarsat-A/-B telex represents the marriage of the most popular terrestrial and maritime messaging technologies utilizing existing equipment.

On the Horizon

The Internet E-mail Gateway permits the exchange of Internet e-mail messages over Inmarsat-C and over Inmarsat-A/-B telex, with the focus on MESs that do *not* have additional equipment. However, there remains a significant need for communication systems that *do* utilize additional equipment to exploit the benefits of store-and-forward communications.

The Stratos Open Systems Messaging (OSM) initiative is a step in that direction. OSM is a development effort that provides a platform for cost-effective store-and-forward satellite communications. This involves a terrestrial-based message hub communicating with maritime and terrestrial systems incorporating Stratos technology. Some key features of OSM are:

- Support for circuit-switched media, such as Inmarsat-A/-B/-M, in conjunction with store-and-forward media, such as Inmarsat-C.
- Most appropriate routing (which is often least cost routing), and intelligent call scheduling over multiple satellite communication networks.
- A circuit-switched communication protocol, designed specifically for the needs of satellite communications, that permits full-duplex and restarted data transfers, and adapts to poor link conditions. The protocol uses a selective reject acknowledgement mechanism that keeps the data flowing while avoiding unnecessary retransmissions.
- Terrestrial access through the Internet, with full support of multi-part and multi-recipient messaging in both the inbound and outbound directions.
- APIs (application programming interfaces) to permit integration with third-party packages. The initial APIs will be COM (Microsoft Common Object Model) and POP3/SMTP (Internet Post Office Protocol / Simple Message Transfer Protocol).

While there are other satellite data communication systems with similar features, they are typically "closed" in nature. That is, they rely mostly upon integrated user interfaces and the like for their messaging traffic.

Meanwhile, OSM focuses primarily on the use of third-party off-the-shelf mail user agents, and third-party developed applications, to generate the traffic.

REFERENCES

[1] RFC822, Internet E-mail header standard.
[2] RFC2045-2049, Multi-part Internet Mail Extensions (MIME) standard.
Note: RFC refers to Request for Comment, the means by which Internet standards are drafted and finalized.

Research Elements Leading to the Development of Inmarsat's New Mobile Multimedia Services

Eyal Trachtman, Terry Hart
Inmarsat
99 City Road, London EC1Y 1AX
United Kingdom,
Eyal_Trachtman@Inmarsat.org

ABSTRACT

Inmarsat has devoted research efforts in recent years in developing Enabling Technologies for a new family of multimedia services. This paper describes a range of research elements leading to the key technologies that have been developed by Inmarsat as part of this effort.

These key technologies span a wide range of telecommunications system disciplines, and include the development of 16QAM and Turbo-coding based air-interface at the physical layer; the Inmarsat generic Medium Access Control (MAC) and radio resource management for Inmarsat shared bearer systems; a transparent ISDN interface for providing seamless circuit-switched inter-networking capability; and the research into compatibility between emerging terrestrial multimedia applications and Inmarsat multimedia systems.

1. INTRODUCTION

A growing demand has been evident in recent years, for increasingly higher data rate mobile communications, supporting a wide range of tele-services, remote access, and mobile-office applications, also known as mobile multimedia applications and services. This new market trend was seen as an opportunity as well as a technical challenge for Inmarsat. Key technologies had to be developed for enabling circuit and packet based, high data rate multimedia services, that are power and spectrum efficient, are operating under optimized radio resource management, are capable of seamless inter-networking with the terrestrial digital networks, and are compatible with the emerging multimedia applications and services.

The provision of power and spectrum efficient multimedia services over the current and future Inmarsat satellite constellations was made possible using a unique air-interface design based on 16 Quadrature Amplitude Modulation and Turbo-coding. This new air interface design has been adapted for both circuit and packet based services. Section 2 provides more details on this research activity.

The support of seamless operation of multimedia services over the Inmarsat space segment was facilitated by the development of a transparent ISDN interface. The interface extends the terrestrial ISDN network to the mobile node, presenting the mobile user with a compatible ISDN interface. Section 3 provides an overview of this research and development activity.

An Inmarsat generic Medium Access Control (MAC) layer was developed and introduced as part of the Inmarsat new packet data service, aiming to provide a suitable definition for the split between the satellite radio interface and higher layer protocols. Further research into Radio Resource Management (RRM) algorithms, will aim to optimize the utilization of shared satellite bearer resource for Inmarsat emerging and future packet systems. Sections 4 and 5 provide an overview of the MAC and RRM research elements respectively.

Finally, the compatibility between the emerging Inmarsat multimedia systems, and the terrestrial multimedia telecommunications applications has been researched aiming to ensure seamless inter-working for circuit and packet based applications. Section 6 provides a summary of this research element.

2. NARROWBAND CHANNEL TECHNOLOGY BASED ON 16QAM AND TURBO-CODING

A narrowband air-interface technology was developed following extensive studies of advanced modulation and coding techniques concentrating on a range of high order modulation and state-of-the-art coding and decoding techniques. The result of this R&D effort is a Narrowband Technology which is based on 16QAM modulation and Turbo-coding. The Narrowband Technology provides significant reduction (more than 50%) in the required bandwidth for mobile satellite channels while improving on satellite power efficiency, when compared to the more traditional channel design of QPSK modulation and convolutional coding. The combination of 16QAM

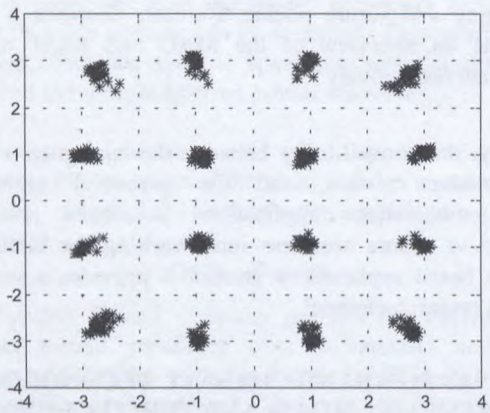
and Tubo-coding is particularly suitable for the Inmarsat emerging Multimedia services. The use of 16QAM modulation optimizes the mobile satellite channel spectral efficiency and ensures maximum flexibility in utilizing the increasingly congested mobile satellite spectrum resource. The use of Turbo-coding provides the additional power efficiency which is required to offset the higher power requirements of 16QAM.

One of the major research concerns was linearity requirements imposed on the mobile terminal HPA due to the use of 16QAM. A careful optimization of transmitter design parameters, including power amplifier back-off, modulator shaping filter and signal constellation allowed the optimization of power amplifier efficiency while limiting receiver performance loss and side-lobe re-growth, thereby attaining optimal power efficiency while minimizing interference effects on adjacent and co-channels.

Another concern, associated with Turbo-coding is implementation complexity, and in particular when considering the mobile terminal. This was a major element in the final, proof of concept phase of the research.

These and further design tradeoffs and optimizations resulted with a robust air-interface which is suitable for interactive and conversational tele-services.

Figure 1: Scatter plot of demodulated 16-QAM signal,



with HPA operating at minimum output back-off.

The Narrowband Technology has been incorporated in the family of Inmarsat Multimedia Services, which are currently being developed. These services are designed to utilize narrowband mobile satellite channels in providing 64 kbit/s circuit-mode and packet-mode data to land-mobile, and aeronautical terminals. The combination of 16QAM and Turbo-coding is planned to be scaled up to form the basis of the air-interface design for the next generations of the Inmarsat system, providing data rates above 64 kbit/s to the user.

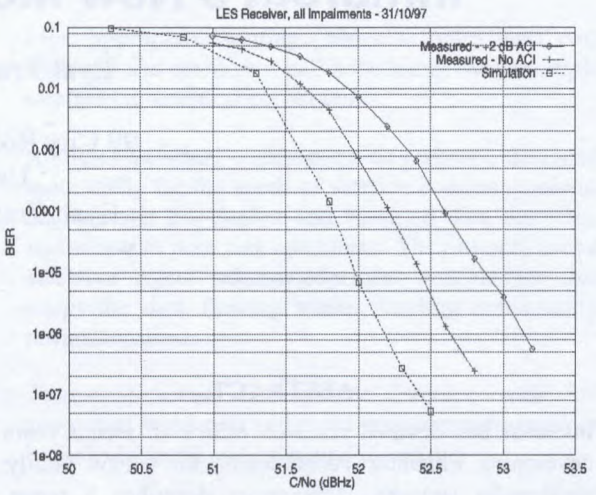


Figure 2: 16QAM and Turbo-coding, measured performance for Inmarsat-Phone M4 64 kbit/s channel, including all implementation impairments.

3. TRANSPARENT ISDN INTERFACE

In order to achieve seamless operation of multimedia services over the Inmarsat space segment, a transparent ISDN interface was defined, developed and verified for the Inmarsat 64/56kbit/s systems. This research effort paved the way for the introduction of a fully compatible ISDN interface in land-mobile, maritime and aeronautical terminals, providing circuit-switched Inmarsat multimedia services.

The interface extends the terrestrial ISDN network to the mobile node, providing mobile ISDN applications with a compatible interface to the one used on the terrestrial network, and presenting the mobile user with a compatible ISDN Network Terminator with S0 bus functionality.

The major effort was devoted to the development of a generic proof-of-concept testing tool, or a Mobile ISDN Network Simulator, for verifying the functionality and performance of the mobile ISDN interface. The network simulator has been used to verify the specifications for a 64/56 kbit/s ISDN interface for the Inmarsat 64/56kbit/s services, including protocol conversion between the Inmarsat signaling system and ISDN D-channel signaling, and space segment effects (physical, link and network layer) on end-to-end inter-networking performance. Simulated aspects of the space segment include effects of end-to-end delay; bit-error patterns resulting from channel fading characteristics; channel power saving algorithms;

clock synchronization aspects; mapping of signaling between ISDN and the Inmarsat system; and the support of basic ISDN Supplementary Services.

The network simulator may be used as a standalone tool, connecting CPE (customer premises equipment, like Terminal Adapters) to the network presentations at each end of the connection; or it may operate in "transparent mode" connecting the simulated terrestrial connection to a real terrestrial network. The simulator enables various configurations of Calling & Called Party Number for one or two ISDN-B channels. This facility enables simulations of calling ISDN CPE, which help the verification of varying combinations of Called and Calling party number.

A key element in the ISDN interfacing facility, which is also part of the mobile ISDN network simulator, is the provision of integrated 128 Kbit/s capability, interfacing a terrestrial Basic Rate Interconnect (BRI) ISDN connection, by using bonding of two 64 kbit/s channels. The network simulator supports independent call establishment set-up times for the two ISDN-B channels, for studying the effect of set-up delay variance between the two bonded satellite channels. This capability, along with the ability to dynamically block the "data" channel for forward and return channels enables investigation for channel bonding architectures. Other simulation elements include dual carrier set-up protocols.

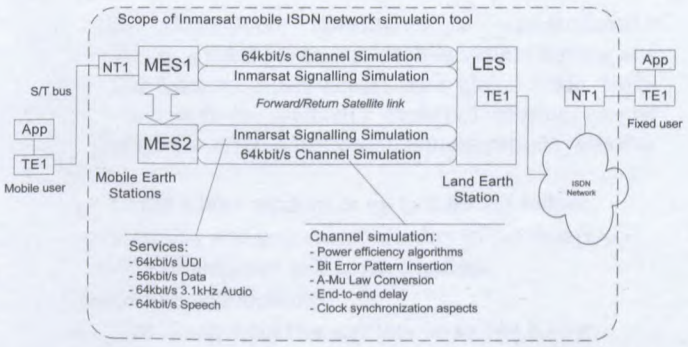


Figure 3: A block diagram of the Inmarsat mobile ISDN network simulator

4. INMARSAT GENERIC MEDIUM ACCESS CONTROL (MAC)

An Inmarsat generic Medium Access Control (MAC) layer has been developed, aiming to provide a suitable definition for the split between the satellite radio interface and higher layer protocols. The generic nature of the Inmarsat MAC layer allows a multitude of different air interfaces, be they packet or circuit switched based, to be incorporated beneath this interface.

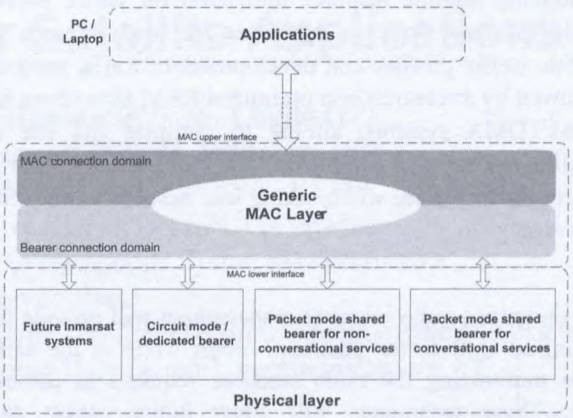


Figure 4: Functional architecture of the MAC layer

At its upper interface the MAC Layer provides an abstraction of Inmarsat services to higher protocol layers providing a single standardized interface to the application developer. The Inmarsat MAC layer also provides a QoS interface that maps user/application requirements to the most appropriate air interface. The MAC Layer allows the ATM like QoS characteristics of CBR, VBR etc. to be selected via the MAC control plane interface. Thus the MAC Layer provides a common low level interface that allows applications to select the most appropriate physical layer using QoS criteria. The different physical layers (e.g. packet mode shared bearer, circuit mode dedicated bearer) become a set of "drivers" to the application, differentiated only by the QoS they can support. Applications do not have to be rewritten to make use of different services. Only one interface is presented to the application developer.

5. RADIO RESOURCE MANAGEMENT FOR PLANNED AND FUTURE INMARSAT PACKET DATA SERVICES

Optimal use of the available radio resource, and in particular the use of return link capacity, is a challenge for packet-data services. The key problem is the definition of suitable Radio Resource Management (RRM) algorithms that can cope with capacity requests for varying message sizes and allocate capacity on multiple multi-rate TDMA channels, whilst maximizing channel occupancy and minimizing message transfer delay.

In order to define the algorithms, it is necessary to understand the requirements of the applications, which utilize the network. Different applications require a range of service attributes for satisfactory operation and produce a range of different traffic profiles. These requirements and traffic characteristics introduce a set of constraints on the underlying networks, and the accurate identification of these constraints is important to allow efficient resource management algorithms to be developed.

An Inmarsat research effort was initiated recently for optimizing satellite resource utilization for future packet mode Inmarsat systems. Activities included research of mobile traffic profiles and development of traffic models, followed by a research into optimized RRM algorithms for TDM/TDMA systems, aiming to optimize the use of satellite resources in a multi-rate and multi-channel, shared bearer environment, while taking into account complexity and cost.

An optimized radio resource management will provide the promised QoS to the user/application while at the same time minimizing the radio resource required to achieve this. These techniques will allow future packet data systems to have greater flexibility and efficient utilization of resources for Inmarsat, while at the same time providing a variety of QoS. Hence, the user will be charged at the QoS he can afford or the QoS his application requires.

6. MULTIMEDIA APPLICATIONS COMPATIBILITY

Finally in order to cement Inmarsat's position as the leading Multimedia Mobile Satellite Services provider, investigations into Multimedia applications compatibility have been initiated. The aim is to ensure compatibility of Inmarsat's new and future services with emerging terrestrial and mobile multimedia telecommunications applications. All aspects of the satellite channel characteristics including, fading, processing delay, satellite delay, BER, and packet error rate (PER) are incorporated in these investigations.

7. CONCLUSIONS

The expectations and communications needs of the mobile satellite user feed on the rapidly growing capabilities of the terrestrial fixed and mobile networks. This trend lead to a growing demand by mobile satellite users for higher data rates, supporting a wide range of tele-services, remote access, and mobile-office applications. This new market trend presented an opportunity as well as a technical challenge for Inmarsat.

Inmarsat took this challenge, and initiated research in key technology areas that have facilitated the provision of high data-rate and seamless connectivity of these emerging multimedia applications and services. These key technologies provided the foundation for high data rate multimedia Inmarsat services, that are power and spectrum efficient, are operating under optimized shared bearer resource management, are capable of seamless inter-networking with the terrestrial digital networks, and are compatible with the emerging multimedia applications and services.

ACKNOWLEDGEMENTS

The authors wish to thank the many colleagues in Inmarsat who have supported this research and development work, and in particular Dr L Christodoulides, Dr P Fines, and T McHugh for their significant part in this activity, and their contributions towards the preparation of this paper.

Development and Validation of Wideband-CDMA IMT-2000 Physical Layer For Satellite Applications

G. Caire (*I*), **R. De Gaudenzi** (*II*), **G. Gallinaro** (*III*), **M. Luglio** (*IV*),
R. Lyons (*V*), **M. Ruggieri** (*VI*), **A. Vernucci** (*III*), **H. Widmer** (*VII*)
(*I*) Politecnico di Torino (Italy) now with EURECOM (France), (*II*) ESA- ESTEC (Holland),
(*III*) Space Engineering (Italy), (*IV*) Coritel (Italy), (*V*) SPCI (Canada),
(*VI*) Universita' di Roma Tor Vergata (Italy), (*VII*) Ascom Systec (Switzerland)
For further information contact:
ESA-ESTEC (TOS-ETC), Keplerlaan 1, 2200 AG Noordwijk, Holland (e-mail: rdegaude@estec.esa.nl)
Space Engineering, via dei Berio 91, I-00155 Roma Italy (e-mail: vernucci@space.it)

ABSTRACT

This paper presents certain aspects relevant to the development and validation of a physical-layer concept dubbed SW-CDMA which was submitted[1] for evaluation to the International Telecommunications Union (ITU) by the European Space Agency (ESA) in the framework of the IMT-2000 satellite-component standardization. After briefly illustrating the proposed air interface solution, the main outcomes of the design trade-off which have led to selecting said access scheme are summarized. The main SW-CDMA differences with respect to the terrestrial W-CDMA counterpart are the permanent softer hand-off operations through satellite diversity, optional forward link CDMA interference mitigation capability and the integration of a radio-localization. To fully validate the proposed air interface in addition to detailed link and system simulations[2], a comprehensive Test Bed being implemented is briefly described. The results herein presented are based upon work[3] being performed by a team led by Space Engineering under ESA contract, encompassing study, analysis, computer simulation, and hardware / software realization activities.

1. INTRODUCTION

The ITU IMT-2000 constitutes a standardization framework for a third-generation mobile system aiming at a unified and integrated user-access on a global scale. IMT-2000, which will be gradually deployed starting around year 2000, will offer advanced communications services featuring:

- up to 384 Kbit/s outdoor or up to 2 Mbit/s indoor;
- good quality and guaranteed Quality of Service (QoS);
- connection-oriented and connectionless;
- bandwidth negotiation;
- satellite component (for services up to 144 Kb/s);
- seamless handoffs across multiple networks.

The European Telecommunications Standardization Institute (ETSI) is contributing to IMT-2000 standardization with his proposal for a Universal Mobile Telecommunications System (UMTS).

It is now well accepted that satellite systems will be necessary to complement terrestrial facilities to offer an *anytime anywhere* service. As a matter of fact satellite systems are mainly intended to cover rural and remote areas which may not be adequately served by terrestrial systems, which instead aim to provide service over high users concentrations. Besides, the targeted market is different due the higher cost of via-satellite services. This has led to consider the satellite component of UMTS (S-UMTS) as an essential element of the overall mobile system, to be fully integrated with the terrestrial component (T-UMTS).

Most of the air-interface proposals submitted to ITU for the terrestrial IMT-2000 are based on various flavors of Wideband Code Division Multiple Access (or W-CDMA), this fact constituting an important guideline for steering the development of the satellite-component access scheme, as access similarity will certainly contribute to making dual-mode Mobile Terminals (MTs) more cost-effective.

On the basis of the results achieved as part of the ESA-funded activity[3], ESA has recently submitted ITU two CDMA-based proposals for S-UMTS access, featuring a high level of commonality with the ARIB W-CDMA and the ETSI UTRA (Universal Transmission Radio Access) T-UMTS Radio Transmission Technology (RTT) proposals, though incorporating those modifications needed to best suit the specific requirements of the satellite environment. Two distinct S-UMTS proposals were submitted by ESA, respectively adopting Wideband CDMA (SW-CDMA) and Code/Time Division Multiple Access (SW-CTDMA). These were devised according to the ESTI UTRA FDD and TDD modes consistently with the main satellite systems constraints, namely:

- reduced power margin with respect to T-UMTS, especially in the Forward-Link (FL) due to on-board power constraints, this making the communications system also noise-limited and not only interference-limited as in T-UMTS;

- significant propagation delay, which tends to reduce the effectiveness of power-control schemes required for proper CDMA operation;
- existence of separate bands for transmission and reception in S-UMTS;
- high frequency error due to Doppler shift, especially for the Low Earth Orbit (LEO) case, which makes initial code and frequency acquisition more critical;
- different environment to implement a coherent Return Link[1] (RL), also caused by the comparatively lower bit-rates which may be expected for the satellite case[2];
- the need to adopt a satellite-diversity technique which requires a suitable support from the access scheme;
- the possible needs for introducing interference mitigation techniques, especially at the MT, to spare satellite power and improve quality of service.

Conversely, the satellite propagation environment is usually more benign, mainly because of its Rician characteristics (C/M typically > 10 dB), compared to the lack of a line-of-sight component, as normally occurring in terrestrial links.
In the following sections, reference is made to the W-CDMA S-UMTS proposal, which appears to be favorite for a global coverage system.

2. PHYSICAL LAYER

There is growing consensus about the CDMA advantages for wireless communication networks. This is particularly true for terrestrial systems whereby CDMA technology has been widely validated during the second-generation IS-95 cellular network deployment. The unique CDMA features for terrestrial networks have been recognized by IMT-2000 proposals[1] that are largely based on this technology. The situation is less clear for the satellite component, following the adoption of different access techniques by second-generation systems such as Iridium, Globalstar and ICO, which however aim at providing low-rate services. For IMT-2000 satellite applications, CDMA is being widely accepted as the most suitable access technique due to the following main reasons:

- full frequency-reuse which largely simplifies system operations and reduces on-board hardware complexity, as the network controller has to simply avoid that the average (and not the worst-case) interference level exceeds a given threshold. This also yields advantages in terms of satellite antenna design, due to the possible relaxation of the sidelobes-peak specification;
- low power flux-density emission. The spread spectrum CDMA nature certainly helps in overcoming regulatory constraints. Clearly the CDMA advantages tends to vanish when system loading increases. However, in general, CDMA provides an additional degree of freedom to control interference;
- graceful capacity degradation under heavy-load conditions,
- "softer" handoff and satellite-diversity exploitation (see Sect. 4);
- easy adaptation to variable-rate services;
- compatibility with the provision of a radio-localization service (see Sect. 5);
- suitability to adaptive interference mitigation techniques, both at the MT and the Gateway Station (GWS) side, to achieve higher capacity and improved quality of service. CDMA is suited much better than narrowband techniques to Multi User Detection (MUD) techniques; in particular the class of Linear Minimum Mean Square Error[3] (LMMSE) type of MUD detectors[4] represent a fairly mature technique which can be integrated in MTs or GWSs for circuit-switched services[5].

Among the often-quoted CDMA drawbacks we here just quote the main ones, that is:

- CDMA, when not complemented by MUD, is quite sensitive to power-control errors; besides the non-negligible satellite propagation delay makes closed loop power-control much slower than in terrestrial cases and dependent on the orbital height. However, detailed power-control simulations showed that power control is actually less critical than for terrestrial networks, because of the limited dynamic power variations due to fast (untrackable) fading;
- less suitable to Return-Link (RL) on-board regeneration, which however would anyway be scarcely attractive, it not permitting to fully exploit the important satellite-diversity advantage.

Passing to consider the proposed S-UMTS RTT, as already mentioned this is a derivation of the ETSI UTRA FDD mode[1]. A detailed presentation of the adaptations that were introduced to suit the satellite environment constraints is available in literature[2], so only the most essential features are here summarized.

3. MODULATION / SPREADING TRADE-OFFS

QPSK modulation with binary Walsh-Hadamard (WH) spreading and binary scrambling on the FL was traded off against other approaches, i.e.:

- dual BPSK with WH spreading and complex scrambling;
- BPSK modulation where each user carrier is spread by a WH code, half of the user carriers being transmitted on the in-phase (I) channel and the other half on the quadrature (Q) channel. I and Q channels are scrambled by two different scrambling codes.

Results of simulations with both the Single User Matched Filter (SUMF) and the LMMSE ideal coherent receivers are given in Figs. 1 and 2 where the Signal-to-Noise plus Interference ratio (SIR) obtained at the receiver output is shown under different loading scenarios. In these figures

[1] An approach being pursued in all recent terrestrial IMT-2000 proposals.

[2] Where only a subset of the terrestrial bearer services will typically be supported.

[3] More precisely Minimum Output Energy blind schemes.

"Q" denotes the QPSK system, "D" denotes the dual-BPSK system and "IQ" denotes the I&Q BPSK system. The nominal S/N (thermal) is 6 dB in all cases.

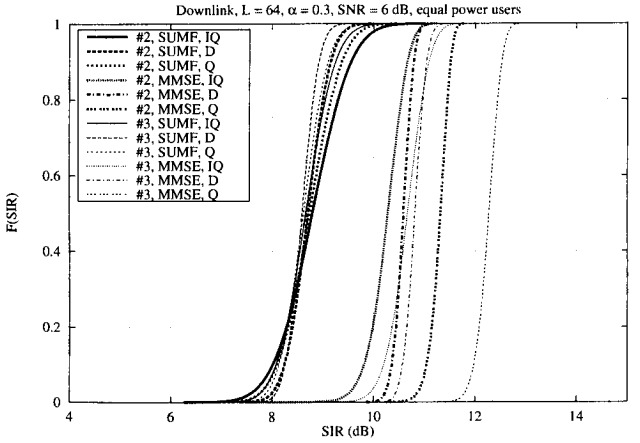


Fig. 1 FL SIR cumulative distribution function. Number of users/spreading factor=0.3

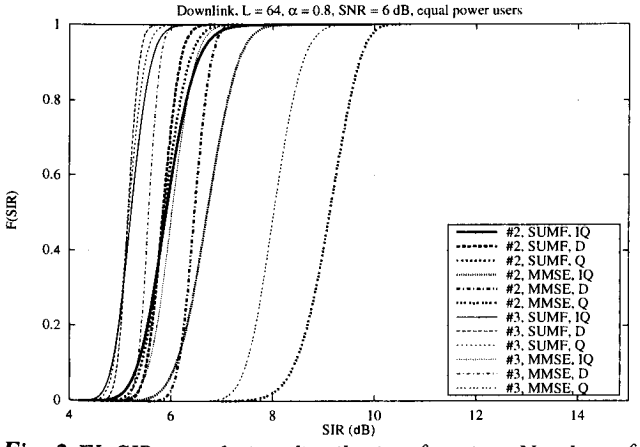


Fig. 2 FL SIR cumulative distribution function. Number of users/spreading factor=0.8

Both double-diversity and triple-diversity with maximal ratio combining were assumed in the simulations, with each satellite carrying the same K number of users (all at equal level). The spreading factor considered for I&Q BPSK was 64. For QPSK and dual-BPSK, the spreading code length is actually doubled, due to the shorter symbol interval. Thick lines are used for the case of Number of Satellites (N_{SAT}) = 2, whilst dotted lines are used for the case of $N_{SAT} = 3$.
With conventional SUMF receivers the three schemes achieve the same average SIR. However, the cumulative SIR distribution for dual-BPSK has slightly shorter tails than that for QPSK. I&Q BPSK has the longest cumulative distribution tails. With an LMMSE receiver QPSK performs significantly better than dual-BPSK and I&Q BPSK, the advantage increasing with the number of users. Clearly the reason of this fact is that the QPSK system uses a spreading code length which is twice that used by I&Q BPSK. Dual-BPSK, which also uses the same code length, requires however two codes instead of one. Another remarkable effect that we can observe from the shown results is that triple satellite-diversity provides better SIR under light loading conditions, whilst in high loading conditions the best SIR is achieved with double-diversity.
With practical receivers, particularly if channel estimation is performed on reference symbols time multiplexed within the traffic channel, at the lower bit-rates QPSK modulation may suffer from effects deriving from frequency errors, phase noise, imperfect channel estimation and to the higher overhead represented by the reference symbols. Fig. 3 shows that, at the low bit-rate of 2,600 bit/s, BPSK has lower overhead with respect to QPSK. Use of common reference symbols for the lower-rate carriers may however reduce the performance gap between QPSK and BPSK.

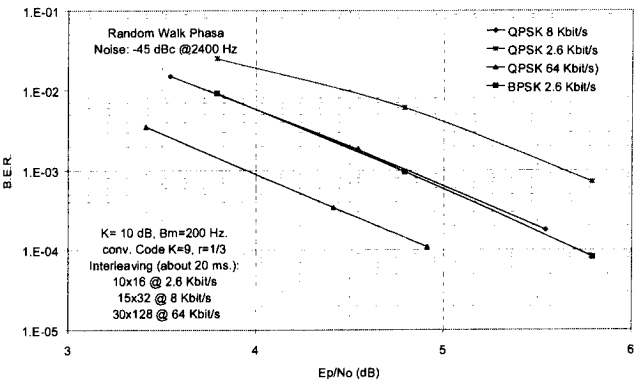


Fig. 3 FL performance in presence of frequency jitter, phase noise and non-ideal channel estimation. AFC loop bandwidth 6 Hz

On the RL (see Fig. 4) the coherent demodulation approach was found to be preferable to differential demodulation or non-coherent multi-level orthogonal modulations like the 64-ary Walsh Keying one used in current CDMA systems (IS-95 and Globalstar), notwithstanding the overhead represented by a pilot signal associated with the carrier.

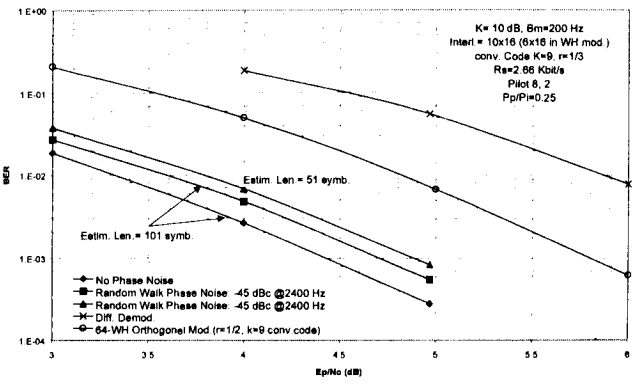


Fig. 4. RL BER performance at low bit rate (2.66 Kbit/s)

4. DIVERSITY

With non-geostationary constellations, beam- and satellite-handoff rates can be rather frequent, and this translates into the need for reliable handoff procedures not to be perceived by users. The so-called "softer" handoff procedure can

serve the purpose, having the receiver to coherently combine signal replicas transmitted over multiple satellites or beams. Softer handoffs are facilitated by the fact that the various signal replicas can operate on the same frequency, this being allowed by the full frequency reuse scheme typically adopted in CDMA systems. Differently from terrestrial systems, where hand-offs duration is intentionally limited in time to reduce the consequent system capacity loss, for satellite systems softer hand-off may well be permanently exploited. Main reasons are:

- satellite systems are often power limited; therefore counteracting shadowing and blockage effects by means of additional link margins may not be viable;
- the S-band mobile fading channel can be considered non-selective for chip-rates up to some 4 Mchip/s. No natural multipath exploitation is thus possible;
- satellite diversity can then be regarded as a sort of artificial multipath, it offering spatial diversity which is very effective in counteracting effects of link obstruction and signal shadowing;

In particular, with regard to FL the following can be said for satellite systems:

- differently from terrestrial, the non-selective satellite fading channel preserves the CDMA orthogonality thus minimizing the intra-beam interference effects.
- by splitting the allocated channel power among different non-collocated satellites the amount of infra-beam interference is somewhat increased. However, in-depth FL system analyses[6] showed that in practice, for a reasonable probability (e.g. 20%) of single satellite blocking (indicated as p_b=0.2 in Fig. 5), the overall system capacity (assuming that at least one satellite is in view) is almost independent of the number of satellites providing path-diversity.

For a larger blockage probability i.e. p_b=0.4, satellite diversity provides even larger overall system capacity.

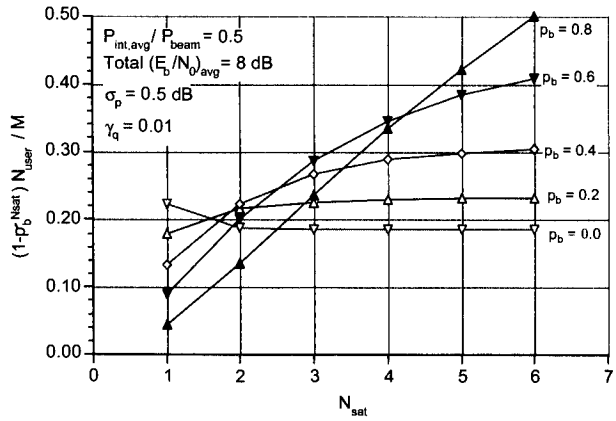


***Fig. 5** From reference[7]. Product of normalized FL capacity times the probability of at least one clear link versus the number of satellites in visibility [N_{sat}], with the single path blockage probability [p_b] as a parameter. 50% interfering power from other beams (same satellite), power control error standard deviation σ_p=0.5 dB, outage probability γ_b=10^{-2}.*

This is a major finding considering that the probability that all satellites are blocked greatly decreases with the number of satellites in simultaneous view, as recently shown by experimental campaigns[8] from which the empirical formulas upon which Fig. 5 is based were derived.

A software tool has been developed[9] to analyze satellite constellations and evaluate performance. Fig. 6 shows, for a Globalstar-like constellation, the probability to have at least one satellite in visibility, from both MT and GWS, for no diversity (one satellite) and diversity order 2 and 3. The minimum elevation angle, defined as the minimum required angle to be covered, is fixed (5 deg.) for the GWS and variable for the MT being supposed to stay in a suburban area. The assumed GWS is located at 42 deg. latitude and 13 deg. longitude, and the MT is located within an about 800 Km-radius ring around the GWS.

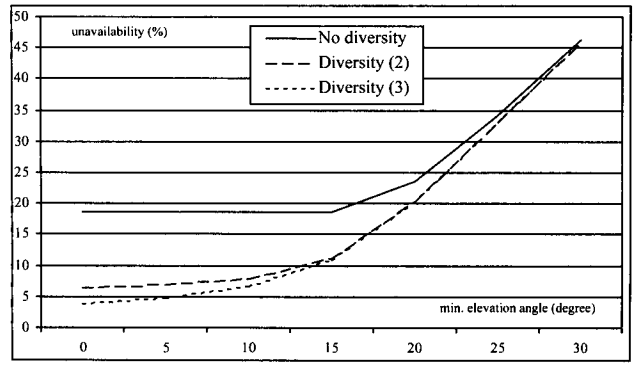


***Fig. 6** Diversity advantage in suburban environment*

In SW-CDMA, multiple satellite-diversity provides a practical way to reduce blockage effects with little or no impact on capacity compared to a system based on line-of-sight operations (thus not truly mobile). This allows achieving superior quality of service without affecting system capacity as it is the case for line-of-sight systems providing great link margins to shadowed users. CDMA also allows to flexibly allocating different degrees of diversity to the different classes of MTs. In fact, fixed or transportable MTs enjoying low blockage probability can be operated with almost no satellite-diversity, thus optimizing network resources exploitation.

Passing to consider the RL, assuming that the MT is equipped with an omni-directional antenna, its radiated signal is picked-up by all satellites in view. The CDMA full frequency reuse permits *satellite antenna arraying* (similar to Deep Space ground reception techniques) whereby the different replicas of the same signal transponded by the various satellites are independently demodulated, time aligned and coherently combined at the GWS. This more sophisticated detection technique requires some additional GWS hardware (essentially a Rake receiver as channel unit demodulator), but results in a drastic reduction in the user terminal Effective Isotropically Radiated Power (EIRP) even under non-blocked conditions.

A last important advantage related to path-diversity consists in the E_b/N_0 reduction achieved by spatial-diversity signal combining in presence of slow fading. This is quite

important as, differently from terrestrial systems, slow fading is neither counteracted by power-control (having slow dynamic capabilities) nor by the finite-size interleaver. Slow fading, characterizing low-mobility users, represents the most power demanding link condition. With path-diversity it is possible to largely counteract the adverse slow fading effects at no additional cost.
The Orthogonal Variable Rate (OVS) spreading technique selected[10] to accommodate different data rates, efficiently support frame-to-frame variable bit-rates, without requiring an increase in demodulator hardware complexity (no need for multi-code correlators, for higher data rate services). Furthermore, FL intra-beam orthogonality is kept also for channels supporting different data rates, thus minimizing the intra-beam interference effects. In the RL the code-division multiplexed signaling channel can also be kept orthogonal to the variable-rate traffic channel by using the very same spreading technique.

5. USER LOCALIZATION

The possibility to provide user positioning as an add-on service to customers is considered to be very important, especially for billing and emergency purposes. By using signals internal to the communication system, i.e. without the support of an external navigation system (e.g. GPS), the MT cost and power consumption can be significantly lowered (GPS operates at a different band than IMT-2000 thus requiring an additional RF front-end and a dedicated demodulator).
The most suitable positioning technique (Time of Arrival, TOA) is that based upon measuring the propagation time between the MT and several satellites. Most TOA schemes are "real time", in the sense that measurements are taken at a given instant in time and are therefore only dependent upon the location of the MT and the satellites, but not on their respective velocities.
Other techniques are possible but not recommended for the subject application. Direction of Arrival (DOA) methods imply too high a payload complexity (if direction has to be measured at the satellite) or too poor a precision (if direction is measured at the MT). Frequency of Arrival (FOA) methods, based upon measuring the Doppler shift between MT and satellite, imply the resolution of six unknowns (position and velocity of both terminal and satellite) instead of three, and would then imply doubling the number of satellites which need to be in view.
Assuming that MTs, for cost reasons, will have no accurate clocks, two main TOA varieties can be considered namely:

- Absolute TOA (ATOA). This is the well known technique adopted by GPS, whereby satellites transmit signals that are synchronized to specific time epochs. Position determination, which must be performed by the MT, requires four (three-dimensional fix) or three satellites (bi-dimensional fix). This is typically a "passive" service which may hardly be charged to customers.
- Loop-back TOA (LTOA). Conceptually identical to ATOA, except for the range which is measured by looping the signal back to the transmitter (i.e. MT-satellite-MT or, conceivably, satellite-MT-satellite). Positioning determination, which can be performed either by the MT or a central processing site where a sophisticated terrain model can more easily be available, requires three (three-dimensional fix) or two satellites (bi-dimensional fix). This can be mechanized as an "active" service, which may be charged to customers.

Other approaches (e.g. Differential TOA or SARSAT-type) are not considered for lack of space although they are appealing when a single satellite is in view.
If *three satellites are used with LTOA*, the position determination algorithm is the same as for the case of four satellites in view, except that the MT clock offset is not relevant. The error budget for this case is shown in Tab. 1.

Tab. 1. *LTOA with 3 satellites - Error budget*

Error Source	**1σ Error (m)**
Round Trip Time Estimation	4.5
Ephemeris	2.1
Multipath	1.4
Receiver Noise and Resolution	0.7
Ionospheric	5.0
Tropospheric	1.4
UERE (two way)	7.4
UERE (one way)	3.7

The one-way User Equivalent Range Error (UERE) is half that of the two-way UERE. In this case the ionospheric error is the largest contributor to UERE. As to the Global Dilution Of Precision (GDOP), in this case we no longer have a time parameter, and thus TDOP (Time DOP) is not relevant. Based on a preliminary guess one can assume that the PDOP (Position DOP) factors are 5 times larger for a communications constellation than for GPS, i.e. not exceeding about 10 for most of the time. Taking the product of the one way UERE and the PDOP factor, the performance of LTOA with 3 satellites can be estimated to be of the order of 40 m (1σ).

If *three satellites are used with ATOA*, external means (e.g. a terrain map in case of land vehicles) may be used for determining the MT altitude. It can be easily demonstrated that the four-satellite ATOA algorithm also applies to the three-satellite ATOA case, the only change being that the missing pseudorange measurement is interpreted as the distance of the MT from the center of the Earth, this not depending on the MT clock offset. A PDOP factor of about 10 can reasonably be assumed for most of the time, if an accurate terrain map or altimeter is available. Overall, we can estimate that the performance of ATOA with three satellites is in the order of 200 m (1σ).
If *two satellites are used with LTOA*, again external means of determining terminal altitude are needed. With an estimated PDOP of 10, the performance of the LTOA system with 2 satellites should be on the order of 37 m (1σ). Further work is needed to achieve more reliable estimates of the PDOP factor; as a general conclusion it appears that a

medium-accuracy positioning service (of the order of 100 – 300 m) may be feasible in most cases. This however may not be achieved with three-satellite ATOA or two-satellite LTOA if no altitude information is available, and only a crude Earth model can then be used. In these two last cases, a low accuracy positioning service (of the order of 1 – 10 km) may only be possible.

6. HARDWARE VALIDATION

The subject project[3] is currently proceeding with the implementation of a complete hardware simulator (the "Test Bed") having the main purpose to validate the SW-CDMA design already been assessed by computer simulations. The Test Bed is intended to reproduce SW-CDMA physical-level functions (including some auxiliary ones, such as power control, diversity combining, beam and satellite hand-off etc.); nevertheless some basic network-level functions are also offered for a more realistic service simulation. The Test Bed is highly re-programmable both to accommodate any adjustments & refinements which SW-CDMA, as all non-yet-mature standards, may be subject to, and to allow its adaptation to also validating the terrestrial ETSI UTRA.

The Test Bed comprises physical devices which generate signals according to the SW-CDMA standard, modify them for reproducing the effects experienced in a real operating environment, demodulate them and finally evaluate the overall transmission performance under programmable true CDMA interference loading conditions. The Test Bed allows to demonstrate in real-time a live service between two Personal Computers running standard applications (voice, video, file transfer, etc,), interconnected by a SW-CDMA physical layer, the parameters of which evolve in time consistently with the selected satellite constellation. One application (the "MT-side application") simulates that running at the MT; the other application (the "GWS-side application") simulates the application operated by the fixed-network user communicating with the MT via a GWS. As shown in Fig. 7, the physical layer, interconnecting the MT-side with the GWS-side application, comprises two distinct paths, the FL and the RL.

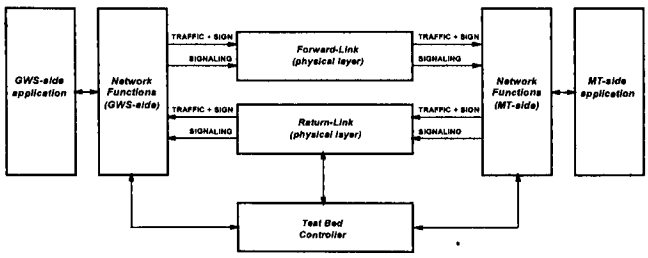


Fig. 7 Overall Test Bed Architecture

The Test Bed configuration resembles that of typical transparent satellite systems (multiple-beam in the mobile link, single-beam in the feeder link) in conjunction with LEO, MEO and GEO constellations. The Test Bed architecture was designed upon the requirements of the most complex orbital constellation, thus implicitly offering compatibility with simpler constellations (i.e. MEO and GEO). The parameters excursion range (e.g. delay, Doppler, path attenuation, etc.) is wide enough to accommodate the extreme values which occur in any foreseeable satellite system. Hand-held MTs using omni-directional antennas are assumed.

The Test Bed incorporates three "satellite paths", each in turn having several "beam paths". This configuration allows to support permanent dual satellite-diversity for both FL and RL), and to simultaneously permit satellite- and beam-handoffs. "Softer" handoffs (see Sect. 4) are possible, as an enhancement of the more common "soft" handoffs whereby the receiver simply selects the better signal replica. The Test Bed allows to reproduce all the main effects and parameters occurring in real via-satellite links, namely:

- path attenuation and delay;
- arbitrary satellite antenna beam patterns;
- thermal noise;
- Doppler;
- propagation channel. This is fully programmable to reproduce different operating environments;
- non-linearity;
- interference caused by other users due to CDMA channel sharing. Proper CDMA codes, and not just thermal noise, are used to simulate interference.

As far as auxiliary functions are concerned, the Test Bed supports the following ones:

- power control;
- frequency control;
- interleaving;
- MUD: the LMMSE scheme is implemented on the MT demodulator.

The Test Bed modems support the full-rate SW-CDMA mode (around 4 Mchip/s), this also allowing inherent compatibility with future terrestrial W-CDMA validation. As to service rates, various rates are supported including 64 Kbit/s and around 8 Kbit/s (tailored on G.729 and G.723-I applications). Higher user rates are possible by the multicode approach (i.e. utilizing multiple 64-Kbit/s channels). An inner Forward Error Correction code (rate either 1/2 or 1/3, and constraint length up to $K = 9$) is implemented, while any outer code is provided on part of the application to be demonstrated.

As already mentioned, the Test Bed also incorporates some basic network-level functions; these include the protocols for setting-up the satellite link when user activity is detected at either the GWS- or the MT-side, and appropriately setting the Test Bed configuration & parameters such as to suit service needs. Also handoff protocols are implemented which, on the basis of FL signal level measurements made by the MT demodulator, trigger and manage satellite- and beam-handoffs. To this end, in addition to the traffic channels, the Test Bed also supports in-band and out-of band signaling channels, with a configuration very close to the operational one.

A Controller, acting as the unique Test Bed control interface for all required functions, manages the configuration and all the parameters of the various Test Bed elements. Said control equipment interprets high-level commands, either generated manually (i.e. by the Test Bed operator via Man Machine Interface, MMI) or by the Dynamic Constellation Simulator, and translates them into low-level setting instructions for the individual elements forming the Test Bed. The MMI also offers full data display (parameters, settings, levels, etc.) & logging facilities for the duration of the communications session.
Purpose of the Dynamic Constellation Simulator is to control, in real time, the Test Bed configuration & parameters, this permitting to demonstrate in real-time a live service while the constellation satellites position evolve.

7. CONCLUSIONS

With the expected convergence of all mobile services into the common IMT-2000 framework, harmonisation of terrestrial and satellite services is becoming more and more an issue. This was one of the main drivers for the activity[3] upon which the present paper is based, which was intended to develop and assess an efficient *open* S-UMTS air interface proposal closely derived from the emerging T-UMTS W-CDMA UTRA standard. A key role at this regard will be played by the availability of a comprehensive Test Bed (see Sect. 6), which, by thorough air interface validation and parameters tune-up, will help in further qualifying the proposed S-UMTS scheme with regard to the on-going IMT-2000 standardization process.

REFERENCES

[1] http://www.itu.int/imt/ Radio/Proposals.

[2] G. Caire, R. De Gaudenzi, G. Gallinaro, R. Lyons, M. Luglio, M. Ruggieri, A. Vernucci, H. Widmer, *ESA Satellite Wideband CDMA Radio Transmission Technology for the IMT-2000/UMTS Satellite Component: Features & Performance*, submitted to IEEE GLOBECOM '99, Rio De Janeiro, Brazil, 5-9 December 1999.

[3] ESA contract no. 12497/NL/97/NB, *Robust Modulation and Coding for Personal Communication Systems*.

[4] E. Biglieri, G. Caire, G. Taricco, *Further Results on MMSE Applicability to Satellite Personal Communications Systems*, Technical note of ESA contract no. 12497/NL/97/NB[3].

[5] R. De Gaudenzi, J. Romero-Garcia, F. Giannetti, M. Luise, *A Frequency Error Resistant Blind Interference Mitigating CDMA Detector*, IEEE 1998 Fifth International Symposium on Spread-Spectrum Techniques and Applications, Sun City, South Africa, September 1998.

[6] G. E. Corazza, *Evaluation for the CDMA Option Definition*, Technical note of ESA contract no. 12497/NL/97/NB[3].

[7] G. E. Corazza, C. Caini, *Satellite Diversity Exploitation in Mobile Satellite CDMA Systems*, submitted to the IEEE Wireless Communications and Networking Conference, WCNC '99, New Orleans, Sep. 21-24, 1999.

[8] Y. Karasawa et. alii, *Analysis of Availability Improvement in LMSS by Means of Satellite Diversity Based on Three-State Propagation Channel Model*, IEEE Trans. on Vehicular Technology, Nov. 1997.

[9] P. Loreti, M. Luglio, *Time Domain Interference/Capacity Analysis for non-GEO Satellite Constellations*, 3rd European Workshop on Mobile Personal Satcoms, EMPS '98, Venice, Italy, November 4-5, 1998, pp. 63-72.

[10] F. Adachi et al., *Tree Structured Generation of Orthogonal Spreading Codes with Different lengths for the Forward Link of DS-CDMA*, Electronics Letters, Vol. 33, No.1, pp.27-28.

A Simulation of Audio and Video Telephony Services in a Satellite UMTS Environment

Daniel Boudreau[1], Robert Lyons[2], Gennaro Gallinaro[3] and Riccardo De Gaudenzi[4]

(1) Communications Research Centre, 3701 Carling Av., P.O. Box 11490, Station "H", Ottawa, Ontario, Canada, K2H 8S2, Tel.: (613) 990-6278, email: Dan.Boudreau@crc.ca
(2) Square Peg Communications Inc., 3701 Carling Av., P.O. Box 11490, Station "H", Ottawa, Ontario, Canada, K2H 8S2, Tel.: (613) 820-7817, email: Lyons@squarepeg.ca
(3) Space Engineering S.p.A., Via dei Berio, 91, I-00155, Rome, Italy, Tel.: 39 622595, email: gallinar@space.it
(4) European Space Agency, European Space Research and Technology Centre, ESTEC/TST, RF Systems Division, P.O. Box 299, Noordwijk, NL 2200 AG The Netherlands, Tel.: +31 71 565 4227, email: rdegaude@xrsun0.estec.esa.nl

ABSTRACT

This paper presents the simulated performance of several joint source and channel coding designs, for a satellite UMTS application involving audio and video telephony. For these services, source coding standards of the International Telecommunication Union (ITU) are combined with matched channel coding techniques. Two speech coding scenarios are proposed; a high quality design with a raw data rate of 8 kbits/s, using the ITU-T G.729 standard, and a slightly lower quality design, using the ITU-T G.723.1 standard at a rate of 6.3 kbits/s. In both designs, voice activation is assumed and simulated. The aggregate source coding bit rate for video telephony (voice plus video) is 64 kbits/s. In this multimedia service, the G.723.1 standard is used at 6.3 kbits/s for the voice channel, while the ITU-T H.263 standard is selected for video coding at 51.2 kbits/s. The three mobile telephony services are studied through computer simulations, in slow and fast fading conditions typical of mobile satellite communications, with and without satellite diversity. The performance is expressed in terms of bit error rate (BER) before and after channel decoding, and in terms of subjective and objective quality measures.

1. INTRODUCTION

As many mobile satellite communication providers are in the process of investigating future services involving low-rate speech and video transmission, the European Space Agency (ESA) is researching source coding, channel coding/modulation and multiple access alternatives for the third generation European mobile satellite system. This paper presents the results of a performance study, for a satellite UMTS channel scenario, of joint source and channel coding designs for audio and video telephony services. Two scenarios for digital speech coding are investigated. High quality voice is considered by using the ITU-T G.729 standard at 8 kbits/s [1]. This standard produces toll quality speech, with an algorithmic delay of only 15 msec [2]. The use of a lower quality speech coding standard, the ITU-T G.723.1 at 6.3 kbits/s, is also simulated [3]. With both of these cases, a silence compression scheme is used to lower the bit rate during silence segments. The video telephone uses the ITU-T H.324 [4] multimedia standard to combine the G.723.1 speech and the ITU-T H.263 video [5] at an overall rate of 64 kbits/s. The video telephone image format is QCIF (144 lines x 176 pixels), updated at 10 frames/s.

The joint source and channel coding strategies involve the use of unequal error protection in the audio telephone, and the use of a powerful Reed-Solomon code in the video telephone. The specific channel coding design is performed by assuming that the transmission system provides the audio and video services with a bit error rate (BER) of 10^{-3}.

In this paper, the channel scenarios implemented in the simulations are first described. The joint source and channel coding designs are treated in Section 3. The error burst statistics at the channel output are discussed in Section 4, along with the corresponding interleaver designs. The results of the simulation of the overall audio and video telephony services are then presented and discussed in Section 5.

2. CHANNEL CONFIGURATIONS

To realistically test the joint source and channel coding schemes devised for satellite UMTS, a comprehensive physical layer simulator was developed. This simulator was conceived in the frame of an ESA study [6] aiming at defining the physical layer of an S-UMTS system. In particular the simulator was developed to test one of the alternative access techniques proposed in the frame of the study. That access is based on a Wideband CDMA scheme [7], which is a close derivative of the terrestrial UMTS ETSI UTRA [8] access.

The simulator actually covers the complete physical layer of the proposed satellite W-CDMA access, including the inner convolutional coding and exact physical frame format of

traffic and control channels. Full support of satellite diversity operation is included for both the forward and reverse links.

For the simulations of this paper, a non-frequency selective Ricean fading channel is assumed, with a Ricean factor of 10 dB. Two different user speeds are considered: 70 Km/h and 3 Km/h, corresponding respectively to Doppler spreads of 140 Hz and 6 Hz (assuming operation in the 2 GHz band). The two cases are referred to as fast fading and slow fading respectively. All the simulations are run using the forward link channel scenario.

Two levels of channel coding are provided in the telephony services. Inner coding is performed with a constraint length 9 convolutional code, with rate 1/3 in all cases except in the reverse link videotelephony, where a rate of 1/2 is used. For the forward channel link considered in this paper, binary spreading is done before QPSK modulation. In the receiver, coherent demodulation (using reference symbols) and Viterbi decoding are accomplished. More information about the channel model can be found in [7].

3. JOINT SOURCE AND CHANNEL CODING FOR AUDIO AND VIDEO TELEPHONY

In order to better protect the different source coding schemes, an outer channel coding level specific to each standard is used. The choice of this second coding level is done by carefully studying the effects of the channel errors on the source decoder quality, and by establishing specific error protection levels. It is assumed that the inner Viterbi decoder delivers a BER of 10^{-3}.

3.1 Speech Coding with the G.729 Standard and Silence Compression.

The G.729 speech coder operates on 10 msec frames. When using the silence compression scheme of its Annex B, each voice frame is represented by 80 bits, while a silence frame is either represented by 15 bits, or is not transmitted. This scheme attempts to continuously model the background noise by transmitting 15 bits of updating information, if the noise statistics have changed, or by not sending information if the noise is stationary. Two extra bits are transmitted in each frame, in order to indicate the frame type. An analysis similar to that of [9] was performed. It is established that the two frame identification bits are extremely sensitive to channel errors. Among the 80 bits of a voice frame, 21 are significantly sensitive to channel errors and cause a noticeable degradation in the synthesized speech, while 59 were relatively insensitive. The sensitive bit numbers, using the numbering of the standard [1], are given in Table 1.

	Bit number
Sensitive class (21 bits)	2, 4, 5, 8, 19 to 25, 27, 46, 47, 51 to 54, 75, 76 ,80

TABLE 1: The ITU-T G.729 bits requiring extra protection in a voice frame.

The threshold BER, for the sensitive class, was established in order to produce a low subjective speech degradation. It was determined that a maximum BER of $5x10^{-5}$ was required on the sensitive bits, along with a BER of $5x10^{-3}$ on the less sensitive bits. The 15 bits modeling the noise in a silence frame also require a BER of $5x10^{-3}$. This level of protection is already given by the inner code, and no extra coding is required.

A shortened (12,2) BCH code was selected for the protection of the two frame identification bits, while a double-error correcting (31,21) BCH code was retained as the outer code for the sensitive bits of a voice frame. On a binary symmetric channel, the (12,2) and the (31,21) BCH decoders deliver BERs lower than $6x10^{-10}$ and 10^{-6} respectively, for an inner BER of 10^{-3}. This outer coding scheme results in a 102-bit coded voice frame, a 27-bit transmitted silence frame and a 12-bit untransmitted silence frame. This produces a maximum coded bit rate of 10.2 kbits/s.

3.2 Speech Coding with the G.723.1 Standard and Silence Compression.

The G.723.1 speech coder operates on 30 msec frames, and has an algorithmic delay of 37.5 msec. The corresponding silence compression scheme is defined by its Annex A, and is similar to the one defined for the G.729 standard. A voice frame is represented by 192 bits, while a silence frame with noise information is represented by 32 bits. Two of these bits, in both types of frame, correspond to information about the frame type. In the voice frame, the sensitivity analysis shows that 84 bits require a protection of $5x10^{-5}$, or better, and that 106 bits require $5x10^{-3}$. In the silence frame, 30 bits require a protection of $5x10^{-3}$. In both these types of frame, the two frame identification bits must be heavily protected. The sensitive bit numbers, using the numbering of the standard [3], are given in Table 2.

	Bit number
Sensitive class (84 bits)	9-11, 17-21, 23-42, 45-62, 65-66, 69-86, 93-98, 100-103, 107-112, 157,175

TABLE 2: The ITU-T G.723.1 bits requiring extra protection in a voice frame.

The two frame identification bits are again protected by a shortened (12,2) BCH code. The bits requiring $5x10^{-3}$ are left protected by the inner code only, while two different coding designs were tested to protect the 84 sensitive bits. In the first design, a 2/3 truncated convolutional code, using the tailbiting technique, was selected to give a (126,84) code. This code was selected for its good error correction characteristics on a binary symmetric channel, for its inherent time diversity, and for the possibility of the Viterbi decoder to produce soft outputs. This last characteristic could be used to apply error concealment in the G.723.1 decoder. In a second design, four parallel (31,21) BCH codes were used, to produce a (124,84) code. This choice was governed by the facts that the (31,21) BCH code produces good results

on bursty channels, and that a more straightforward comparison could be done with the G.729 design. On a binary symmetric channel, both the truncated convolutional code and the parallel BCH approach result in BERs lower than 10^{-6}, for an inner BER of 10^{-3}. These schemes result in coded bit rates of 8.13 and 8.07 kbits/s respectively.

3.3 *Video Coding with the H.263 Standard.*

The H.263 standard is a hybrid of inter-picture prediction to utilize temporal redundancy, and transform coding of the remaining signal to reduce spatial redundancy. It also has motion compensation capability. The bit rate at the coder output is variable, depending on the image information. A rate control algorithm is used, in order to obtain an average bit rate of 51.2 kbits/s. Annexes D, F, J, S and T are used in the coder [5]. The results of the sensitivity analysis performed on the H.263 video standard have indicated that a good strategy is to protect all the coded bits evenly, at an error rate of 10^{-5} or better. An 8-bit (255,223) Reed-Solomon code was selected to provide this threshold, from the inner BER of 10^{-3}. On a binary symmetric channel, this Reed-Solomon code results in a BER lower than 10^{-10} for an inner BER of 10^{-3}. It is also very efficient when long error bursts are encountered.

An extra level of joint source and channel coding is present in the video telephone design. The video coder is operated such that portions of the image are updated using only transform coding, without any reference to previous images (no prediction). This *forced updating* is performed at the image macroblock (16x16 pixels) level at least once every 20 frames. At 10 frames/s, this allows the image to be completely refreshed every 2 seconds, and prevents prediction error accumulation. Image reception is possible at a BER of 10^{-4}, although with a noticeable degradation.

4. ERROR STATISTICS AND INTERLEAVER DESIGNS

The combination of a Viterbi decoder with a multipath fading channel typically produces error bursts at the input of the outer channel decoder. Some form of bit or symbol interleaving is therefore necessary to spread out the effects of these bursts, and to allow the outer decoder to approach the performance predicted in Section 3, for a binary symmetric channel.

A typical error burst length distribution at the Viterbi decoder output is shown in Fig. 1, for the AWGN channel, when the inner Viterbi decoder delivers an average BER of 10^{-3}. An error burst is defined here as a group of bits in which two successive erroneous bits are separated by less than 12 correct bits [10]. The length of the error bursts is concentrated around multiples of 8 bits, i.e. multiples of the number of bits necessary to return the decoder to the correct decoding path, after an error has occurred. This figure shows that the Viterbi decoder, when operating in a static environment, typically produces error bursts with a maximal length of around 56 bits.

A simulated error burst distribution, for speech telelphony on a slow fading channel, is shown in Figure 2. As expected, large burst lengths are common, although the distribution is dominated by the bursts smaller than 56 bits. This indicates that, although it is undesirable (due to the constraints on the transmission delay) to use an interleaver that spreads out the longest bursts, selecting one that controls the bursts typical of the static channel minimizes the degradation in the subjective quality of the telephony services.

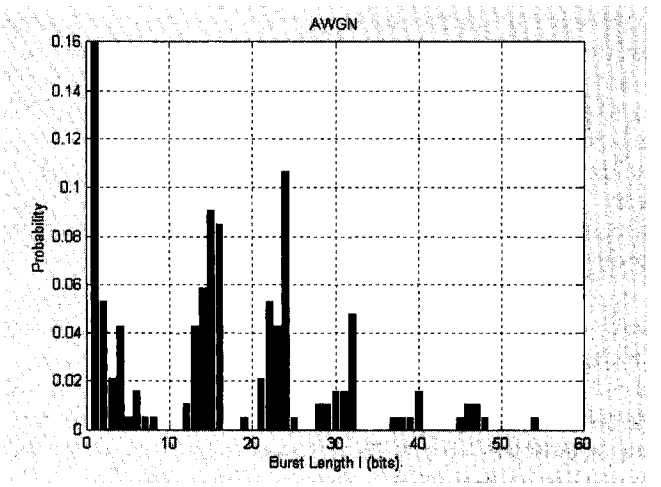


FIG. 1: Typical burst lengths distribution on the AWGN channel.

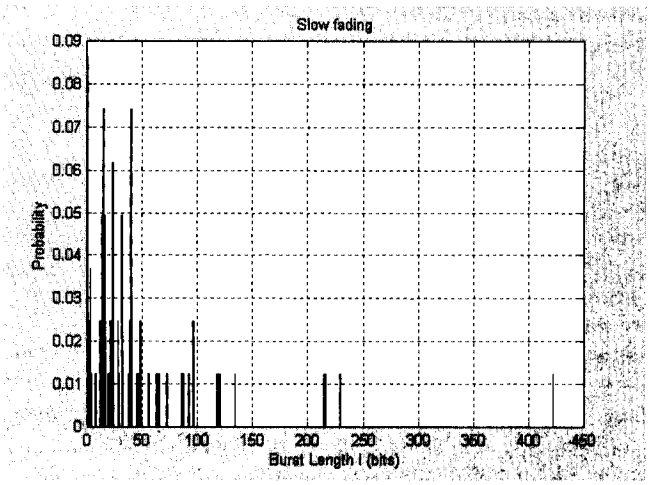


FIG. 2: Typical burst lengths distribution on the hand-held urban channel, for speech telephony.

In order to combat the effects of error bursts, as described in the previous section, specific interleavers were designed for the different types of services and outer coding schemes.

4.1 Interleaver Design for G.729 Speech Telephony

In the G.729 audio telephony service, the 102 bits of each outer coded frame are first rearranged in such a way that the

12 frame identification bits are spread out as much as possible by inserting a combination of sensitive bits and insensitive bits between them. The sensitive bits are themselves spread out by the insensitive bits. Denoting the frame identification bits as "X_i" bits, the sensitive bits as "Y" bits and the insensitive bits as "Z" bits, the frame reordering is

$X_1YZZYZZYZX_2YZZYZZYZX_3\ldots X_4\ldots X_5\ldots X_6\ldots X_7\ldots X_8\ldots$
$X_9\ldots X_{10}ZZYZZYZZX_{11}ZZYZZYZZX_{12}ZZ.$

For interleaving over N speech frames, the reordered bits are stacked in the rows of an Nx102 matrix, and are read column by column. This type of interleaver spreads the X_i bits by at least (9N-1) bits and the Y bits by at least (3N-1) bits. This arrangement is such that a single error burst of 6N bits or less does not exceed the double-error correcting capability of the (31,21) BCH code. This interleaver introduces an extra delay of 10x(N-1) msec. In order to counteract all of the error bursts of length 56 or less, N must be at least 19, implying an extra delay of 180 msec. In practice, the system delay constraints limit N to 3.

4.2 Interleaver Design for G.723.1 Speech Telephony and Convolutional Outer Coding

In this design, 244 bits are generated per 30 msec frame (12 frame identification bits, 126 convolutional coded bits, and 106 insensitive bits). First, the 126 coded bits of each frame are interleaved in a block inner interleaver of the form shown in Fig. 3. This scheme spreads an error burst of 10 errors over groups of length equal to the code constraint length (12 channel bits). Since the code has a d_{free} equal to 6, most of the resulting incorrect paths can be corrected by a Viterbi decoder.

R								
E	1	2	3	4	...	10	11	12
A	13	14	15	16	...	22	23	24
D	x	-	-	x	...	x	-	x
↓	x	-	-	x	...	x	-	x
	109	110	111	112	...	118	119	120
	121	122	...	126				

FIG. 3: Inner interleaver table for the G.723.1 audio telephony service using convolutional outer coding.

The inner interleaved bits are then reordered, in each frame, with the other bits in the following way (xZ refers to "x" consecutive Z bits):

$X_1\underline{Y}_1 9Z\ X_2\ \underline{Y}_2 9Z\ X_3\ldots\ X_7\ \underline{Y}_7 10ZX_8\underline{Y}_8 10Z\ldots.\ X_{11}\underline{Y}_{11} 10ZX_{12}\ \underline{Y}_{12} 2Z$

where the "Xi" bits are the frame identification bits, the "$\underline{Y}_j$" vectors represent the columns of the matrix of Fig. 3, and the Z bits are the insensitive bits.

As for the G.729 case, interleaving over N speech frames is performed by stacking the reordered frames in the rows of an Nx244 matrix, and by reading column by column. This type of interleaver spreads the X_i bits and the Y bits (the components of the $\underline{Y}$ vectors) by at least (21N-1) bits. It introduces an extra delay of 30x(N-1) msec. In order to counteract all of the error bursts of length 56, N should be at least 3, implying an extra delay of 60 msec. In practice N is limited to 1.

4.3 Interleaver Design for G.723.1 Speech Telephony and Block Outer Coding

The block outer coding scheme generates 242 bits per 30 msec frames (12 frame identification bits, 124 block coded bits, and 106 insensitive bits). The 124 coded bits are interleaved in a 4x31 block interleaver, where the 4 rows are the outputs of the four parallel (31,21) BCH codewords. The bits are read column by column, defined as columns $\underline{A}_1$ to $\underline{A}_{31}$. Then these bits are interleaved with the frame identification bits (the X_i bits) and the insensitive bits (the Z bits) as

$X_1\underline{A}_1\underline{A}_2 7Z\underline{A}_3X_2\underline{A}_4 7Z\underline{A}_5\underline{A}_6X_3 7Z\underline{A}_7\underline{A}_8 4ZX_4 3Z\underline{A}_9\underline{A}_{10} 7ZX_5\underline{A}_{11}$
$\underline{A}_{12} 7Z\underline{A}_{13}X_6\underline{A}_{14} 7Z\underline{A}_{15}\underline{A}_{16}X_7 7Z\underline{A}_{17}\underline{A}_{18} 4ZX_8 3Z\underline{A}_{19}\underline{A}_{20} 7ZX_9\underline{A}_{21}$
$\underline{A}_{22} 7Z\underline{A}_{23}X_{10}\underline{A}_{24} 7Z\underline{A}_{25}\underline{A}_{26}X_{11} 7Z\underline{A}_{27}\underline{A}_{28} 4ZX_{12} 3Z\underline{A}_{29}\underline{A}_{30} 7Z\underline{A}_{31}$
Z

This arrangement spreads the BCH coded bits such that a single error bursts of up to 15 bits does not produce more than two errors in any of the double-error correcting BCH codewords. Interleaving over N speech frames is performed by stacking the reordered frames in the rows of an Nx242 matrix, and then reading column by column. This type of interleaver spreads the X_i bits by at least (19N-1) bits. It can tolerate error bursts of 15N bits, without exceeding the error correction capability of the BCH codes. It introduces an extra delay of 30x(N-1) msec. In order to counteract all of the error bursts of length 56, N should be at least 4, implying an extra delay of 90 msec. In practice N is limited to 1.

4.3 Interleaver Design for Video Telephony

The powerful 8-bit (255,223) Reed-Solomon code selected for the video telephony service is by nature a burst error correcting code. It has 32 parity symbols for each codeword, and can correct error bursts of up to 121 bits [11]. On the slowest fading channel configurations, this error-correcting capability can be easily exceeded, leading to visually annoying artifacts. This implies that a fairly large symbol interleaver is required in these cases. It was determined through simulations that a 255x6 block symbol interleaver is a good compromise between delay and error spreading capability, at least for the static and faster fading channels.

The multiplexing of the audio and video bits is performed using a 10 msec time base, at a rate of 64 kbits/sec. Each multiplexed frame contains 64 G.723.1 bits (audio) and 512 H.263 bits (video). A 255x6 block symbol interleaver therefore spreads the error effects over 6 codewords, and introduces a delay of 16.8 multiplexed frames (168 msec).

5. SIMULATED PERFORMANCE

The simulated performance of the different source coding scenarios is evaluated by using a combination of objective and subjective measurements. The BER at the output of the outer decoder is first measured and gives an indication of the interleaver efficiency. In the case of the speech services, the segmental SNR (SEGSNR) is computed, and a subjective listening evaluation is conducted. For the videotelephony service, a subjective evaluation is performed.

5.1 G.729 Speech Telephony

The BER, as measured at the output of the (31,21) BCH decoder is given in Table 3, when the system is operating at threshold, i.e. when the inner Viterbi decoder delivers an average BER of around 10^{-3}. For all channel cases, the outer BER is about one order of magnitude higher than the desired threshold of $5x10^{-5}$. This performance is largely due to the fact that the system delay constraints limit the interleaving to three frames only (N=3). The higher outer BER does not necessarily lead to a large subjective voice degradation, as will be demonstrated below. This is the case because the value of $5x10^{-5}$ was determined when the 59 less sensitive bits were subject to an error rate of $5x10^{-3}$. Here, these bits experience a BER of 10^{-3}, which gives some margin for the subjectively tolerable BER on the sensitive bits.

Channel	Eb/No (dB)	Outer BER on BCH Decoded Bits
AWGN	4	$7x10^{-4}$
Fast fading (140 Hz)	6	$6x10^{-4}$
Slow fading (6 Hz)	9	$4x10^{-4}$
Slow fading with double satellite diversity	4	$5x10^{-4}$

TABLE 3: The measured BER at the output of the (31,21) BCH decoder (sensitive bits) in the G.729 speech telephony service operated at threshold.

The received voice quality corresponding to the conditions of Table 3 has been evaluated by using the segmental SNR measure (SEGSNR). This measure is an average, and tends to mask the localized effects of the error bursts. The results, for a one minute audio passage, are given in Table 4. It is noted that the SEGSNR is always close to its largest possible value of 1.5. The degradation in voice quality, as evaluated subjectively (in informal tests), is also indicated in this table. This degradation is always small and is dominated by the burst of errors still present in the slowest fading cases.

Between these error bursts, the subjective quality is high. The speech intelligibility is high at all times.

5.2 G.723.1 Speech Telephony and Convolutional Outer Coding

For the 6.3 kbits/s codec using a truncated convolutional code, the threshold outer BER is always approximately equal or worse than the input BER of 10^{-3}. These results indicate that the combination of the outer code and interleaver is extremely sensitive to error bursts, and that the outer BER levels are unacceptable. The resulting voice output is very poor and has not been evaluated further. The main cause of this poor performance is the system delay constraint, which limits the outer interleaving over a single speech frame (N=1).

Channel	SEGSNR (dB)	Subjective Degradation	Intelligibility
AWGN	1.41	small	high
Fast fading (140 Hz)	1.42	small	high
Slow fading (6 Hz)	1.44	small	high
Slow fading with double satellite diversity	1.45	small	high

TABLE 4: G.729 objective voice quality (SEGSNR) and the subjective degradation for the cases of Table 3. The error-free SEGSNR is 1.5 dB. The degradation scale is: none, small, medium and high.

5.3 G.723.1 Speech Telephony and Block Outer Coding

For the inner BER threshold, the measured BER at the output of the four parallel (31,21) BCH decoders is given in Table 5. Note that the results are very similar to those of Table 3 for the G.729 case, at least for the AWGN and the fast fading cases. The outer interleaving being limited to N=1, as in Subsection 5.2, these results indicate that the parallel BCH approach is more tolerant to the channel error bursts than the truncated convolutional coding approach.

Channel	Eb/No (dB)	Outer BER on BCH Decoded Bits
AWGN	3.5	$7x10^{-4}$
Fast fading (140 Hz)	5.75	$5x10^{-4}$
Slow fading (6 Hz)	8.5	$1x10^{-3}$
Slow fading with double satellite diversity	4	$1x10^{-3}$

TABLE 5: The measured BER at the output of the parallel BCH decoders (sensitive bits) in the G.723.1 speech telephony service operated at threshold.

Channel	SEGSNR (dB)	Subjective Degr.	Intelligibility
AWGN	9.16	high	medium
Fast fading (140 Hz)	10.18	high	medium
Slow fading (6 Hz)	10.38	high	medium
Slow fading with double satellite diversity	10.5	high	medium

TABLE 6: G.723.1 objective voice quality (SEGSNR) and the subjective degradation for the cases of Table 5. The error-free SEGSNR is 10.97 dB. The degradation scale is: none, small, medium and high.

The voice quality corresponding to the conditions of Table 5 are given in Table 6. Note that despite the fact that the BER performance is similar to that encountered in the G.729 scenario, the voice degradation is always high, and the speech intelligibility is deteriorated. This is an indication of

the superiority of the G.729 standard over that of the G.723.1 standard, on a bursty channel.

5.3 Video Telephony

The video telephony service was evaluated for one minute sequences. The BER measured at the output of the (255,223) Reed-Solomon decoder is indicated in Table 8. These results are better than the BER subjective threshold of 10^{-5} for the AWGN and the fast fading channel, but are poor for the slow fading cases.

Channel	Eb/No (dB)	Outer BER on R-S Decoded Bits
AWGN	3	$< 1x10^{-10}$
Fast fading (140 Hz)	4.5	$< 1x10^{-10}$
Slow fading (6 Hz)	8	$8x10^{-4}$
Slow fading with double satellite diversity	4	$4x10^{-4}$

TABLE 8: The measured BER at the output of the (255,223) Reed-Solomon decoder in the video telephony service operated at threshold.

These results indicate that the combination of the outer code and the outer interleaver is not powerful enough to deal with the error burst distribution typical of the slow S-UMTS channel. The subjective degradation corresponding to the cases of Table 8 is indicated in Table 9. As expected, the subjective quality is degraded in the slowest fading cases. This is particularly true for the video portion of the communications, in which even the smallest artifact is annoying. The reproduction of the audio sequence could benefit from using the G.729 standard instead of the G.723.1, although this would not comply with the H.324 multimedia standard.

Channel	*Audio* Subjective Degradation	*Video* Subjective Degradation
AWGN	none	none
Fast fading (140 Hz)	none	none
Slow fading (6 Hz)	high	high
Slow fading with double satellite diversity	high	high

TABLE 9: The subjective degradation for the cases of Table 8. The degradation scale is: none, small, medium and high.

6. CONCLUSIONS

The simulated performance of audio and video telephony over a satellite UMTS communications link was presented in this paper. The joint source and channel coding techniques selected for these services were discussed, as well as the resulting objective and subjective quality obtained over an AWGN channel, as well as over a fast and a slow fading channel. Satellite diversity, with two satellites, was also simulated. The results show that speech telephony is possible with good quality, over all the channel scenarios at a coded bit rate of 10.2 kbits/s, by using the ITU G.729 standard. The design based on the G.723.1 standard, and operating at a coded bit rate of 8.07 kbits/s, is not very good. In order to increase the quality of this latter design, either more channel resources are required, to increase the channel coding redundancy, or more delay needs to be incorporated in the system, to increase the interleaver length.

Despite a powerful outer coding scheme, and a long outer interleaver, the quality of the video telephony service is acceptable only in the AWGN and the fast fading cases. Extending the operation to the slow fading scenarios would require some combination of lower rate channel coding and error concealment in the video decoder.

REFERENCES

[1] International Telecommunication Union, *Coding of Speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP)*, ITU-T Recommendation G.729 (03/96), March 1996.

[2] R. V. Cox, "Three new speech coders from the ITU cover a range of applications," IEEE Communications Magazine, vol. 35, no. 9, September 1997, pp. 40-47.

[3] International Telecommunication Union, *Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s*, ITU-T Recommendation G.723.1 (03/96), March 1996.

[4] International Telecommunication Union, *Terminal for low bit-rate Multimedia Communication*, ITU-T Recommendation H.324 (02/98), February 1998.

[5] International Telecommunication Union, *Video Coding for Low Bitrate Communication*, ITU-T Recommendation H.263 (02/98), February 1998.

[6] European Space Agency Contract No. 12497/NL/97/NB, *Robust Modulation and Coding for Personal Communication Systems.*

[7] G. Caire, R. De Gaudenzi, G. Gallinaro, M. Luglio, R. Lyons, M. Ruggieri, A. Vernucci, H. Widmer, "Development and Validation of Wideband-CDMA IMT-2000 Physical Layer For Satellite Applications," 1999 International Mobile Satellite Conference IMSC'99, Ottawa, Ontario, June 16-18, 1999.

[8] http://www.itu.int/imt/2-radio-dev/proposals.

[9] D. Boudreau, C. Dubuc, J. Lodge, P. Guinand, R. Lyons and L. Erup, "High-quality video and audio telephony for mobile satellite applications," Proceedings of the Japan-Canada International Workshop on Multimedia Wireless Communica-tions and Computing WMWCC'96, Victoria, British Columbia, September 17, 1996, pp. 49-50.

[10] CCITT *Blue Book*, vol. 1, Fascicle I.3, Terms and Definitions, ITU 1988.

[11] R. Gallager, *Information Theory and Reliable Communication*, Wiley, 1968.

An Introduction to Inmarsat's New Mobile Multimedia Service

by

Howard Feldman
and
D.V. Ramana

Inmarsat
99 City Road
London EC1Y 1AX
Howard_Feldman@Inmarsat.org
D.V._Ramana@Inmarsat.org

ABSTRACT

Inmarsat's new multimedia service is based on the commercially successful mini-M platform. This paper describes how the new offering uniquely incorporates 16QAM (quadrature amplitude modulation) and turbo coding into an existing mobile satellite communications system. These and other new technologies have allowed us to create a laptop-sized multimedia communications tool which extends the terrestrial ISDN and IP networks to remote users via existing Inmarsat-3 spot-beam satellites.

1. BACKGROUND

Since its founding in 1979, Inmarsat has continually introduced revolutionary mobile satellite telecommunication systems. All systems have been designed for personal use, i.e. by individuals, ships, or small groups, offering voice, fax, and data communications to remote and mobile users. They vary from the palm-sized Inmarsat-D, a data system operating at some hundreds of bits per second, to suitcase-sized Inmarsat-B, which offers toll quality voice and data at 64 kbit/s. In between those extremes fall Inmarsat-M, mini-M (also known as Inmarsat-phone) and Inmarsat-C, as well as a variety of systems tailored to the requirements of the Aeronautical community. Each of the Inmarsat systems is targeted to a specific market, and hence makes a different set of trade-offs with regard to voice quality, data rate, equipment parameters (size/weight/cost) and call charges (tariffs). Naturally, the systems introduced most recently, such as mini-M, have taken advantage of the newest advances in technology. Mini-M is able to offer features such as a near toll-quality voice service, using a codec which requires a

channel rate of merely 3.6kbit/s. Improvements in technologies which are crucial to mobile satellite systems, such as antennas, DSP modems, and batteries have allowed the overall size and weight of mobile satellite systems to shrink with each new generation, even as the features are expanded and extended.

In spite of the exciting advances in many areas, the modulation and coding systems used in mobile satellite systems have changed little over the past decade. All of the existing systems rely on class-C RF power amplifiers, and therefore require compatible modulation methods such as offset QPSK or filtered BPSK. Inmarsat systems have used Viterbi or sequential coding, exclusively, to improve the bit error rate. Until recently, Viterbi coding has been the best choice to reduce satellite power requirements, at an acceptable price of slightly increased bandwidth, latency and complexity.

2. SUMMARY

The latest in the family of Inmarsat systems is a fourth-generation product, offering Mobile ISDN service. It uses many of the advanced technologies developed for the mini-M voice system, and reaches out to more state-of-the-art technologies for the data service. Significant enhancements have been made in the modulation and coding methods, as well as in the power control subsystem. The combination of these and other new concepts have allowed us to halve the power and bandwidth requirements, as compared to previous systems.

Features

Built on the successful mini-M platform, the Mobile ISDN service offers good quality, low-cost voice using the low-rate AMBE voice codec. In addition, full-rate Group-3 facsimile and ISDN-compatible 64kbit/s data communications are available. GSM-like SIM (subscriber identity module) is also a standard feature in the mobile terminals. The overall concept, which will be enhanced with additional types data services in future, is designed to allow the extension of corporate LANs and WANs into true Global Area Networks.

3. TECHNOLOGY

Inmarsat's Mobile ISDN offering comes from an engineering programme of development referred to as "M4". This section discusses several specific technologies which are part of the M4 programme: signalling, modulation and coding, power conservation, and the ISDN interface.

Signalling System

The M4 programme is founded on the mini-M access control and signalling system (ACSE). The signalling channels used for access control are based on a combination of out-of-band and in-band channels. The out-of-band signalling uses a TDM in the forward direction (fixed to mobile) and slotted Aloha in the return direction (mobile to fixed). The forward and return channels are both BPSK with Viterbi coding. The in-band signalling is a virtual carrier, and can disappear when not required.

Modulation and Coding

This section describes the modulation and coding used on the channels which carry customer traffic. The voice service uses the same modulation and coding methods as were developed for mini-M. AMBE voice coding is supplemented by block codes on selected bits to increase robustness in high BER environments. This is used in conjunction with Offset-QPSK modulation.

For the 64kbit/s services, we chose 16QAM combined with turbo coding, after evaluating several different modulation and coding combinations.

Power Conservation Features

There are a number of features developed or enhanced in the M4 programme to reduce power requirements. Reduction in power gives the service a significant market advantage in that it reduces the price, size, and weight of the mobile equipment, while minimising the cost of the calls. The main power-saving feature is power control. Power control was developed initially for use in the mini-M system, and operates in both the forward and return directions. The forward power control is useful to optimise use of the satellite L-band power; the return power control will conserve the mobile terminal battery as well as the satellite C-band power.

A second power conservation feature is called Sleep Mode. Sleep Mode is designed primarily to conserve the mobile terminal's battery power. In the idle mode, the mobile terminal lies dormant; it awakes when a call is received or initiated. The parameters of sleep mode are centrally controlled by Inmarsat, and broadcast by the network control station.

Another power saving feature is carrier activation. Carrier activation was originally developed decades ago for terrestrial telephony systems, where the measured activity rate (i.e. time when a user is talking) is approximately 40%. In mobile satellite systems, the transmit carrier is similarly turned off when the local speaker is quiet. This improves satellite power efficiency in the forward direction, and saves battery life in the return direction. The term for this method, when used during voice calls, is "Voice Activation". In the mini-M system, this technique was extended to cover facsimile traffic and the more general term "Carrier Activation" applied. The concept of carrier activation has been further expanded to include data calls as well, to save satellite power and reduce battery consumption.

ISDN Interface

The M4 programme included a development plan toward adapting the ISDN B and D channels for use in mobile satellite systems. ISDN was chosen as a standard interface to ease interworking with applications. Four ISDN bearer services will be offered: Speech, 3.1kHz Audio, 64kbit/s Unrestricted Data Interchange (UDI), and 56kbit/s data (using V.110 rate adaption). The international "standardised" ISDN interface will be present on all mobile terminals, supporting ISDN connectivity into terrestrial networks. This is critical since the new system is designed for a global market, with the same mobile earth station models being used world-wide. This causes two intrinsic conflicts resulting from lack of complete standards at the international level: differences in A/D coding algorithms, and differences in the user interface points.

In the majority of countries around the globe, the 64 kbit/s bitstream is derived from incoming analogue waveforms using A-law companding; notable exceptions are USA, Canada, and Japan where mu-law is used. To further complicate matters, the ISDN does not signal which companding system is used, rather it relies on implicit knowledge based on the ISDN location. Inmarsat mobile earth stations are transported around the globe and are allowed to interface to the terrestrial world via any of a network of land earth stations (LESs). In the context of M4, LESs will assume that the terminal adapters used at the mobile end are A-law. A-to-mu conversion will take place at the LESs when required. The consequence is that mobile users, even those based in countries which use mu-law encoding, will need to configure their mobile terminals with A-law ISDN terminal adapters.

To further complicate the picture, most countries specify the user connection point as the S-interface; again, ISDN operators in the USA and Canada tend to specify the U-interface. Mobile Earth Stations (MESs) will all offer the S-interface as a standard; some may offer U-interface as an optional feature.

4. CONCLUSION

Inmarsat's newest service is based on a combination of old and new technologies, designed to reach the target compromise between development time, feature set, satellite efficiency, and portability. Inmarsat has been a provider of mobile satellite services since 1979, whose commercial successes have been based on continuing improvements in our technology combined with a thorough understanding of market requirements. This is the latest development in the continuing improvement and enhancement of Inmarsat's mobile satellite systems.

ABOUT THE AUTHORS

Howard Feldman is Programme Manager, and as such has overall responsibility for the design, testing and roll-out of the new system based on M4 technologies. In his 12 year tenure at Inmarsat he has been responsible for various design and development projects on the Inmarsat-M, Inmarsat-B, and mini-M services, including significant contributions toward system design, facsimile and data interfaces, interfaces to secure communication systems, and mobile terminals. He holds numerous patents in the fields of satellite systems, facsimile and data communications. His Master's Degree in

Electrical Engineering, with a speciality in telecommunications, is from the University of Texas at Arlington. He received his BSEE degree is from the University of Michigan. He was previously with Motorola.

D.V.Ramana as the Manager of the Inmarsat M/B Engineering Dept is responsible for the overall Engineering issues related to System design, development, verification and service introduction of Inmarsat M/B/mini-M/M4 and Mobility Management Systems. Prior to joining Inmarsat in 1983, he was with the Indian Space Research Organization where he was responsible for the design and implementation of Telemetry, Tracking and Command systems for Indian satellites. He holds a PhD from Indian Institute of Science, Bangalore in the area of Satellite Based Packet Data Systems.

THE ASTROMESH DEPLOYABLE REFLECTOR

Mark W. Thomson
TRW Astro Aerospace
6384 Via Real
Carpinteria, CA, USA
93013-2920

1.0 Abstract

The AstroMesh is a mesh reflector for large aperture space antenna systems. It embodies a new concept for deployable space structures: a pair of ring-stiffened, geodesic truss domes, in which the ring is a truss deployed by a single cable. Compared to other mesh reflectors, the AstroMesh achieves uncharacteristically low levels of total mass, stowed volume, surface distortion, cost, and program schedule duration.

Several large AstroMesh reflectors are being produced for a series of U.S. built commercial geosynchronous mobile communications spacecraft; first flight is planned for early 2000. Figure 1 depicts the AstroMesh on the Euro-African Satellite Telecommunications (EAST) System, an example of a large aperture mobile communications satellite conceived by Matra Marconi Space.

Two large models have been qualified for space flight with as-manufactured shape errors of less than $D(6 \times 10^{-5})$ RMS, and repeatability errors below $D(5 \times 10^{-6})$ RMS (where D is the aperture diameter). With current materials and manufacturing techniques, the AstroMesh design can achieve a surface accuracy from all error sources, including in-orbit environments, of $D(2.5 \times 10^{-5})$ RMS. Large margins on a very demanding passive intermodulation (PIM) specification have been achieved consistently.

2.0 Background

Since the late 1960's, mesh reflectors have been favored for their potential to fill large apertures with extremely lightweight hardware. Development of this new genre continued with impressive devices that were expensive, heavy, and bulky when stowed. They also proved to be difficult to deploy, test and characterize on the ground. Great effort was required to achieve tolerable surface precision with apertures larger than 6 meters.

Models that have flown are of a ribbed or umbrella nature, often with a network of cords for fine shaping of RF reflective mesh.[1,2] Many designs tended to impose manufacturing errors, products of kinematic and/or static indeterminacy with high redundancy. Thus, they were difficult to pre-stress uniformly throughout.[3] The inherent tendency of mesh to saddle back into the dish with a positive curvature and its highly anisotropic mechanical qualities exacerbated inherent structural problems. The presence of the ribs leads to large-scale systematic surface errors, particularly during thermal extremes between full sun and eclipse in high orbits. Such problems, combined with reliability and mesh management issues, led to the undoing of several past designs.

7510-220D-01

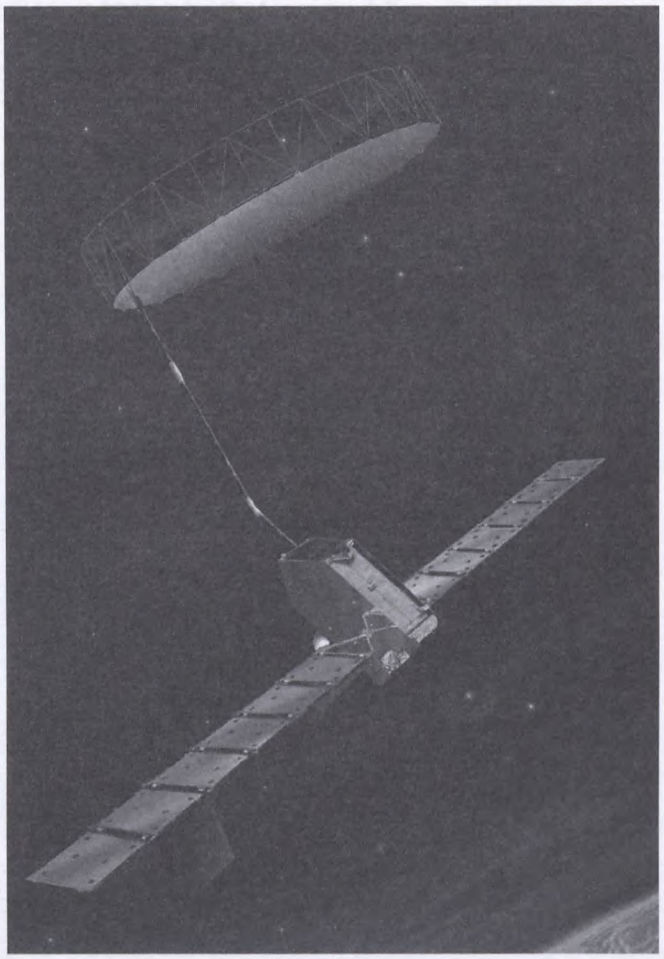


Figure 1.

Communications, imaging, and scientific satellite missions demand antennas with ever larger and more precise apertures. At the same time, satellite procurements are being streamlined to better match a commercial model. The programs require rapid, low-cost design and development turnarounds without sacrificing reliability. Ever larger reflectors with superior performance will always be in demand if competitive designs are available. Thus the goal of my project: to produce a better, faster, and cheaper deployable reflector.

7510-122-06

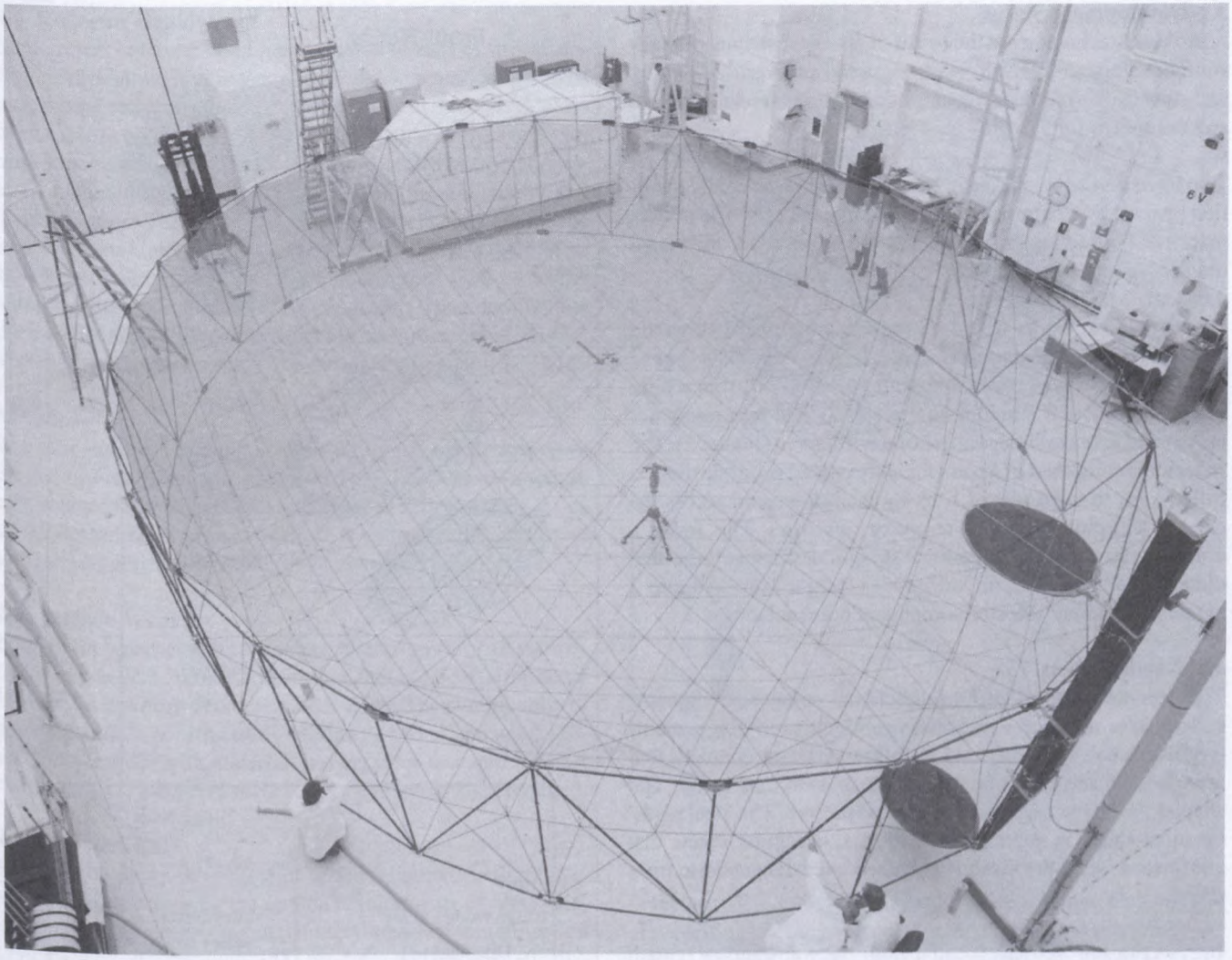


Figure 2.

3.0 Introduction

The patented AstroMesh technology is mature.[4] Both the 6-meter and the much larger flight models have been fully qualified for space flight through a comprehensive program of environmental testing. The flight hardware is shown in Figure 2. The design is readily scalable and supports low-cost, minimum schedule programs with rapid hardware delivery for commercial satellite competitiveness.

The AstroMesh utilizes a simple and efficient structural approach. When deployed, it is significantly stiffer, more precise, thermally stable, and lower in weight than existing mesh reflectors. When stowed, it requires less than usual volume. Apertures up to 25 meters can be integrated with large spacecraft buses and yet be launched aboard a number of available medium-size boosters. Reflector areal mass densities of 0.37 kg/m^2 have been demonstrated with flight hardware. For larger reflectors the areal density drops quickly due to an economy of scale in structural efficiency. For passive intermodulation (PIM) sensitive antenna systems, significant margins on the most rigorous PIM specifications have been achieved.

The AstroMesh is practical to build. With existing materials and technology, a maximum surface distortion of $(2.5 \times 10^{-5})D$ from all sources (including worst-case orbital thermal influences) can be achieved. For example, a 40 GHz antenna with an aperture of 12 meters can have a surface accuracy of 0.3 mm or l/25 RMS. Current knitted wire mesh designs can be used up to about 40 GHz with adequate reflectivity. For lower frequencies such as L band, aperture size is limited by existing launch vehicles to about 150 meters. The basic design of the reflector is the same regardless of aperture size, so it is inherently scalable.

Compared with other mesh reflectors, the simplicity and structural efficacy of the AstroMesh yields low fabrication and ground test costs. Although compatible with axisymmetric antenna geometry, the design is inherently compatible with off-axis antenna systems, as shown in Figures 1 and 2. The AstroMesh is appropriate for the majority of applications where very large apertures are required. It is being considered for missions that include communications, synthetic aperture radar, radiometry, radio astronomy, and interferometry.

3.1 Development Status

AstroMesh technology is the result of five generations of hardware development over 8 years. A record of over 220 deploy and stow cycles has been accumulated by three models without any failures to fully deploy unaided.

Two large flight versions have been built and successfully qualified by a suite of electrical and environmental tests. The performance of both reflector subsystems met or exceeded all design requirements.

4.0 AstroMesh Concept

Deployable reflectors based within circumferential rings are recognized as superior, in principle, to historic ribbed types.[5] The AstroMesh was conceived with this in mind; it is a light and inherently stiff structure that precisely and repeatedly deploys reflective mesh, regardless of environment (Figure 3). The reflector is composed of a pair of doubly curved geodesic trusses, called *nets*, that are placed back-to-back in tension across the rims of a deployable graphite-epoxy ring truss. This forms a drum-like structure of exceptional structural efficiency, thermal dimensional stability, and stiffness-to-weight ratios. Figure 2 reveals additional reflector component nomenclature.

4.1 Geodesic Nets

The nets are made of stiff unidirectional composite filaments called *webs*, attached in a pseudo-geodesic truss dome geometry. Webs can be configured to follow reflector contours that are spherical and parabolic with single or double curvature, and shaped, as long as negative curvature is present. The ideal mathematical shape is approximated by flat, triangular facets that are formed when the mesh is stretched over the geodesic truss dome.

Intersections of webs are located along the web lengths, L, with an accuracy better than (3×10^{-7})L RMS using specialized equipment and material conditioning techniques. The intersections, or net "nodes," become hard points in the deployed, tensed nets. A state of uniform tension averaging at least 50 N per meter between the mirrored net nodes is established by the action of tension tie assemblies. The nets are not adjustable, nor do they need to be to achieve the required surface accuracy.

4.2 Ring Truss

The articulated graphite composite and aluminum ring truss is deep and extremely stiff normal to the plane of symmetry between nets. Its structural depth and regularity provides high thermal stability and ease of manufacturing. Though the reflector is extraordinarily lightweight, overall stiffness allows its direct mounting from the edge of the ring truss to a boom (Figures 1 and 2).

During deployment, the deployable truss is a cable-actuated, synchronized parallelogram. During all phases of deployment, the truss retains sufficient stiffness in all directions to sustain significant attitude control, system-induced loads while remaining tightly coupled kinematically.

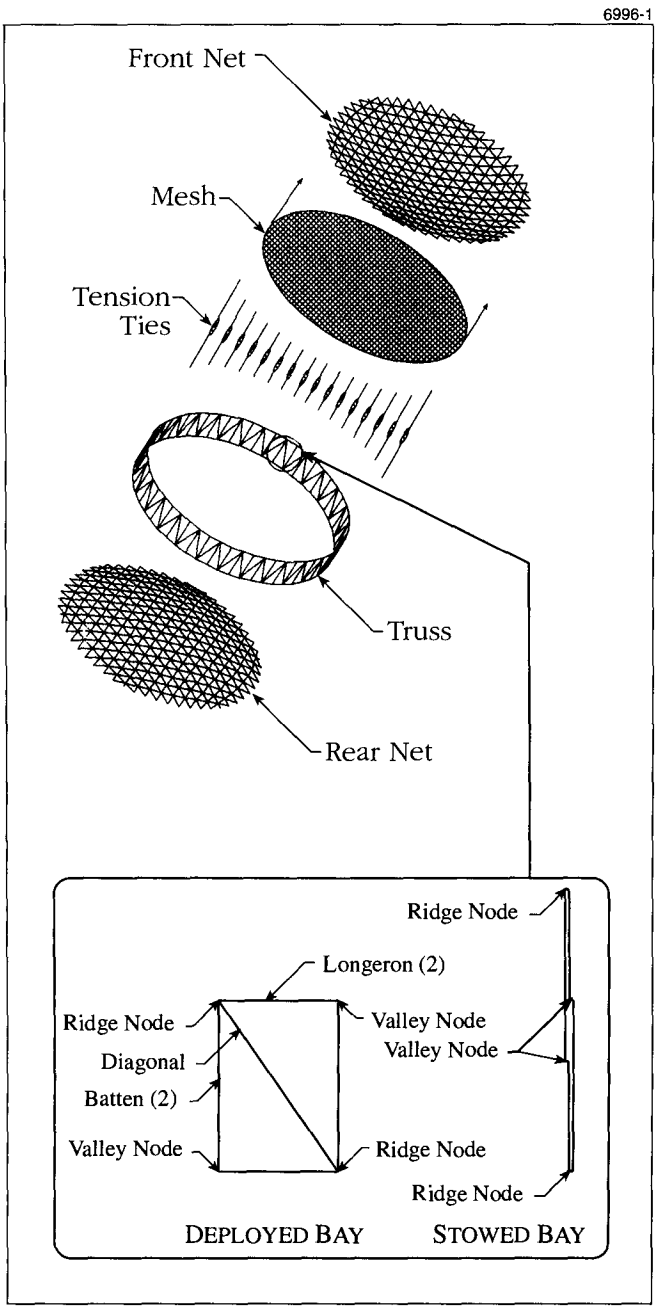


Figure 3.

4.3 Stowed Reflector

The AstroMesh is compactly stowed. Figure 4 shows an optimum interface arrangement. Fully stowed, the truss package is a narrow, hollow cylinder, deployable truss members adjacent to each other. The end-most members are preloaded against lightweight hoops that act as stiffening and debris-shielding end-caps for the stowed package. This forms a barrel-like structure that is strapped into lightweight cradles for launch. Certain members in the package are intentionally over-stowed so that the truss gently expands or "blooms" when the tie-downs are released.

4.4 Reflector Deployment

Deployment behavior combines an initial release of stowed energy with a slow, highly controlled and kinematically robust motor-powered phase to complete deployment. A single deployment cable is given powerful geometric authority over reflector truss kinematics. The articulated truss deploys with synchronous, kinematically-coupled behavior and predominantly low levels of strain. The cable enjoys a large mechanical advantage over reflector kinematics; at the end of deployment it transfers significant work-energy to the dome and ring systems to develop the high overall pre-stress condition. Upon deployment the truss latches solidly, thus requiring no further cable tension. The mesh and nets are snag-free during deployment.

4.5 Mass

The efficient truss and diaphanous mesh and net design gives the AstroMesh a dramatic downward trend in mass-to-aperture area ratios as diameter increases. Mass to reflector area ratios of 0.37kg/m^2 have been proven. It is estimated that ratios as low as 0.2 kg/m^2 are feasible for 25-meter class reflectors.

4.6 Surface Accuracy

The lowest worst-case total contour distortion from all sources believed practical for the AstroMesh is about (2.5 x 10^{-5})D, about 1/4 of what has been demonstrated thus far. The combination of denser geodesic net structures, tailoring of truss and net CTEs, and thinner webs with materials having ultra-low strain scatter characteristics, will boost accuracy to these levels when required.

4.7 PIM and ESD

Every stride in the development of the reflector for flight systems has been driven by the unique requirements of offset-fed, extremely PIM-sensitive geostationary mobile communication systems. The elimination of PIM products and electrostatic discharge (ESD) events has been achieved in every component of the deployed reflector, regardless of location. Simple but specialized design and fabrication measures insure consistent PIM performance that is well below -150 dBm at the feed point.

5.0 Summary

Easy achievement of high net tension is a key feature of the performance and economy of the AstroMesh. This capability is in part due to the statically determinate structure of the deployed system. This eliminates the need to characterize or control the inherently nonlinear characteristics of mesh tension and highly redundant shaping networks. It also contributes to the repeatability and uniformity of the structural shape during thermal extremes and after multiple deployments.

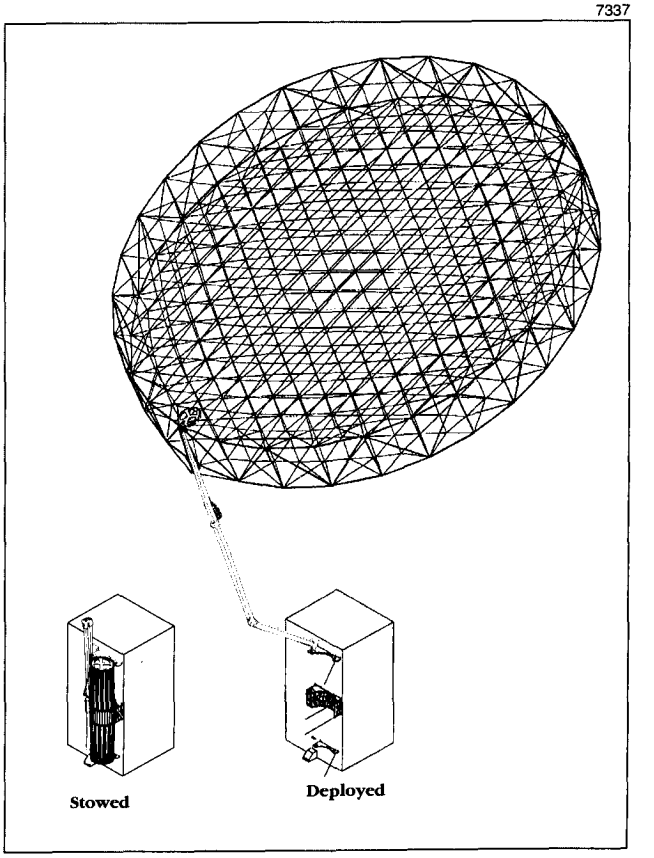


Figure 4.

[1] Freeland, R.E., "Survey of Deployable Antenna Concepts," Proceedings of Large Space Antenna Systems Technology—1982, NASA-CP-2269, Nov.-Dec., 1982, pp. 381-421.

[2] Roederer, A.G. and Rahmat-Samii, Y., "Unfurlable Satellite Antennas: A Review," Annales des Telecommunications, Vol. 44 No. 9, 10, 1989, pp. 475-488.

[3] You, Z. and Pellegrino, S., "Cable-Stiffened Pantographic Deployable Structures Part 2: Mesh Reflector," AIAA Journal, Vol. 35, No. 8, 1996, pp. 1348-1355.

[4] U.S. Patent No. 5,680,145.

[5] Hedgepeth, J. M., "Influence of Fabrication Tolerances on the Surface Accuracy of Large Antenna Structures," AIAA Journal, Vol. 20, No. 5, May 1982, pp.680-686.

The Modular Mesh Reflector developed at NTT

Kazuhide ANDO, Akihiro MIYASAKA, Hironori ISHIKAWA, Mitsunobu WATANABE
NTT Network Innovation Laboratories
1-1 Hikarinooka, Yokosuka-shi, Kanagawa-ken, 239-0847, Japan
Tel: +81-468-59-3508
Fax: +81-468-59-3351
e-mail: ando@maria.wslab.ntt.co.jp

ABSTRACT

Large deployable on-board antenna structure will be necessary in future space missions such as advanced satellite communications. These large reflector structures must be lightweight and easy to stow, and they must have high surface accuracy. In order to meet these requirements, the authors have developed a large deployable modular antenna for the S-band that has mesh surface reflector. This modular mesh reflector consists of 14 modules and has an electrical aperture of more than 13 m (15 m in mechanical diameter). Seven modules were fabricated as prototypes., The mesh reflector surface accuracy, deployment characteristics, natural frequency, and thermal properties were studied regarding seven-module assembly. The authors also suggest simplified ground tests using only one module on deployment, stiffness and so on which can reduce the costs and the period of development.

INTRODUCTION

Large on-board antenna reflectors will be key devices for advanced satellite communications and other space missions [1]. These large reflector structures must be lightweight and easy to stow, and they must have surface accuracy in accordance with radio frequency (RF).

Several proposals have been put forward for geostationary satellites that have large onboard deployable antenna reflectors. Figure 1 shows the recent technological trends in terms of reflector mass as a function of reflector aperture. From the figure, one can see that deployable structures fit into four categories : 1) solid folding plate-based structures, 2) cable/mesh deployable structures, which use a cable structure to form a metalic mesh surface into a pseudo-parabola and have a deployable support structure, 3) graphite mesh structures, which use carbon fiber triaxial textiles to form a mesh-like reflector plane that can be stowed by wadding it into a cylindrical shape, and 4) inflatable structures, which use an inflatable polymer-based film as the reflector plane structure. However, from the standpoint of weight, stowability and cost, the solid folding plate-based structures are unusable at the present moment for deployable antenna structures that measure over 10 meters in diameter. Inflatable structures have been under development for many years and there are variety of design possibilities [2]. However, it is still not clear whether such a structure can ever be successfully implemented with high surface accuracy. Cable/mesh type reflector must have support structures, such as truss structures because the tension of the cable and mesh membrane structures must be supported. Truss structures have high structural accuracy, but have, until now, major disadvantages with respect to cost and weight. In a effort to reduce the cost and weight, NTT Network innovation Laboratories have been developing a large deployable antenna reflector with a mesh surface and a deployable truss structure, which consists of several modules [3,4] (Fig. 2). Figure 1 shows that NTT's reflector is sophisticated among cable/mesh type reflectors. Large structures based on the module concept are high-extensibility in designs, in other words, structures can be enlarged by adding more modules. In addition, each module is small enough to be tested at a limited facility. The AstroMesh Deployable Reflector [5] has reached very high level in weight and diameter, however, it's structure consists only of "one module".

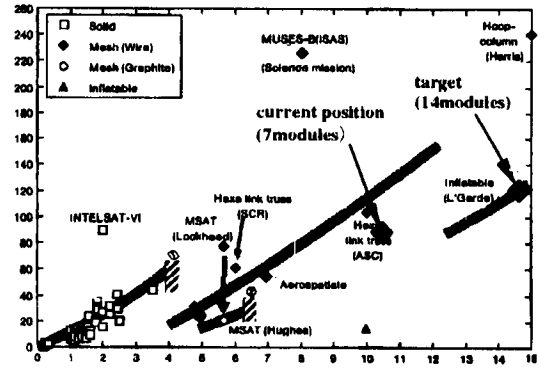


Fig.1 Trends in Large Reflectors

DESIGN and ANALYSIS PROGRAM

The design process for large deployable mesh reflector structures, involves a lot of re-design according to the results of various analyses that evaluate every aspect of the

design. In particular, the consistency among models with respect to various analyses, such as reflector surface shape analysis, truss deployment analysis, thermal deformation analysis, modal analysis, etc. has to be taken into account. In order to make sure the models are consistent, the authors have developed a software tool called the Object Oriented Coordinate Designer (OOCD) that makes it possible to store, reuse, modify and compose basic structural concepts and design know-how. It can shorten the design term even if the analyses have to be repeated several times.

OOCD has various structure classes that are described in C++, such as Node, Element (subclasses are Cable, Rod, Beam, Membrane), and Hinge (subclasses are Revolute, Slide, Cylindrical, Weld), as can be seen in Fig. 3. These classes are derived from the primary abstract class, which is called ModuleBase. ModuleBase is the most abstract class and has virtual functions for basic instructions (for example, graphical expressions, output for various analyses, and modification). A structure model made in this manner can be transformed into a manufacturing

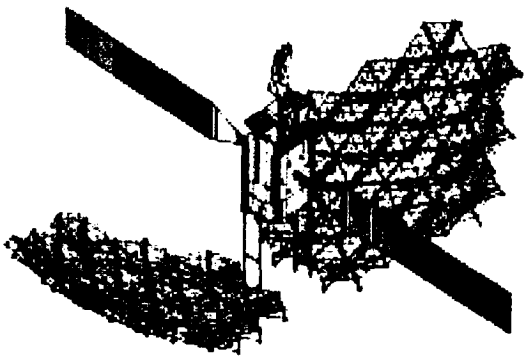


Fig.2 Modular Mesh Reflector Image

[6]. SPADE makes it possible to estimate the elastic distortion of structural parts in a sequence of motions. Furthermore, we can precisely predict structural behavior before manufacturing the structural hardware. A mesh reflector's surface can be considered to be a membrane structure. SPADE can also treat tension structures, such as

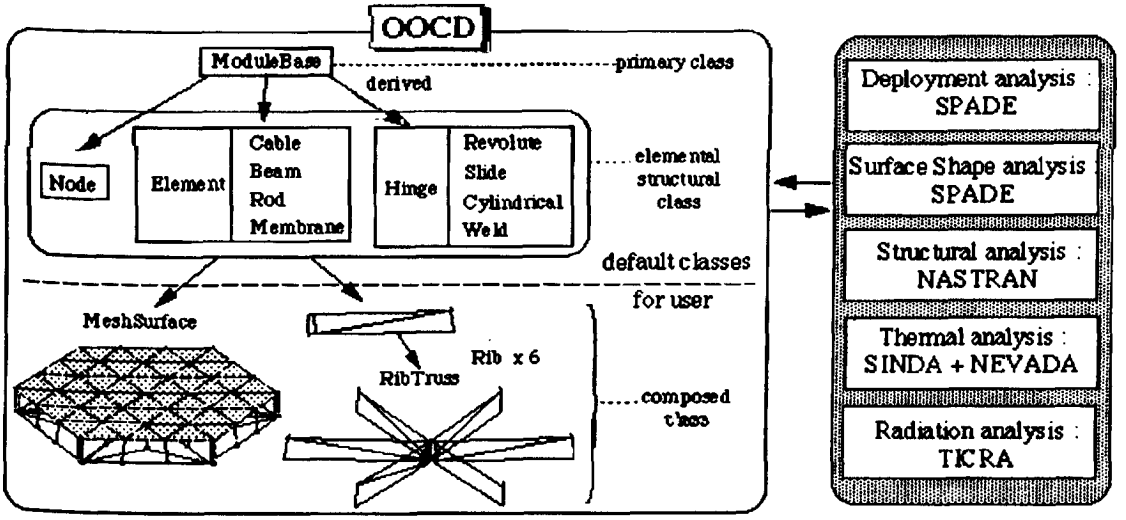


Fig.3 Software Environment for Design

design format, NASTRAN format, thermal analysis format, and SPADE (Simple Partitioning Algorithm based Dynamics of finite Element) format (Fig.3). Needless to say, OOCD ensures consistency among formats of various analyses. Detailed sizes in the model are defined with the OOCD parameter resource file and they can be changed very easily. When the elemental structure must be changed (for example, by adding or reducing beams in the partial base structure), the analysis model of the whole structure can be changed in the same way.

SPADE is a computer program developed at NTT for flexible multibody dynamics. It is based on finite element method using linear kinematics and a corotational frame

cables [7,8] and membranes [9].

MODULE DESIGN

The authors designed using OOCD and SPADE a large deployable antenna reflector with a mesh surface and a deployable truss structure that consists of several modules. The complete antenna consists of 14 modules and has an electrical aperture of more than 13 m (15m x 19m in mechanical size). Seven module have been fabricated as prototypes. The basic modules are about 4.8 m in diameter. Each module weights 10.7 kg, which includes the weight of the deployment motor, and can be stowed within 30 cm x 30 cm x 3.5 m envelope. The envelope of stowed 14

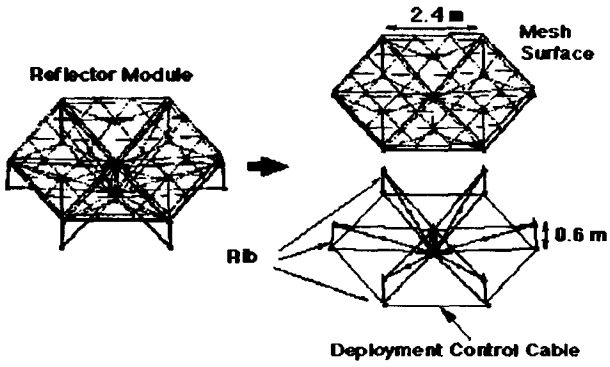


Fig.4 Module Composition

module assembly measures 104 cm x 120 cm x 3.5 m.

Each module is composed of metallic mesh, a cable network, and a deployable truss structure (Fig.4) and has the following features. (1) The deployable truss structure comprises a center axis and six radial ribs. Each rib is a transformable frame that includes a stabilizing beam for smooth deployment. (2) The deployment force is provided by a translational spring connecting the slider and the center axis. The spring configuration stabilizes the module in the deployed state, which gains deployment reliability of the module. (3) The surface of the connected modules is configured to best fit sphere surface for the designed parabola. Therefore, each module can be constructed from beams of the same size. This simplication greatly reduces the cost and time needed to fabricate large antennas. The deployment control cable surrounds the module so that the deployment motion will proceed gradually.

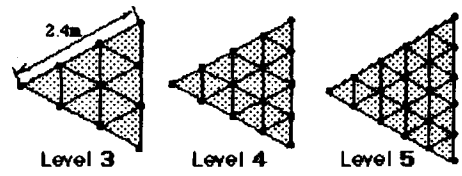
Two types of deployable structures (a slide-type truss structure [10,11] and an articulated-type truss structure [12,13]) were compared from the standpoint of deployment driving force needed to overcome the mesh surface tension. The both types were evaluated by SPADE [14]. The slide-type truss structure without articulated beams was selected because it had a lower peak mesh surface tension in accordance with deployment motion than did the articulated-type truss structure. This judgement was based on the idea that the lower the peak of mesh surface tension is, the higher the deployment reliability is.

MESH SURFACE DESIGN PROCEDURE

The surface of a mesh reflector can be regarded as a membrane structure, and therefore, it is subject to the kind of distortions found in membrane structures. Pillow distortion is a deformation mode unique to membranes, and it occurs in the direction opposite to the curvature of the reflector surface. This sort of distortion is a significant obstacle to achieving the desired RF performance. The ideal surface is approximately parabolic in shape and consists of many flat planes, but in reality, each plane becomes curved due to the pillow distortion. Pillow distortion depends on cable tension, mesh membrane tension, membrane size and surface curvature (related to focal length).

Table 1. Surface error due to approximation and pillow distortion

Surface cable network division



Focal length 10.4m

(1) Surface error due to plane approximation [mmRMS]

Level 3	Level 4	Level 5
0.64	0.4	0.3

(2) Surface error due to pillow distortion [mmRMS]

TensionRatio (Tc/Tm)	Level 3	Level 4	Level 5
1.0	0.97	0.42	0.22
1.2	0.82	0.36	0.19
1.5	0.67	0.29	0.15
2.0	0.52	0.22	0.11

RSS [mm] of (1) and (2)

TensionRatio (Tc/Tm)	Level 3	Level 4	Level 5
1.0	1.16	0.58	0.37
1.2	1.04	0.54	0.36
1.5	0.93	0.49	0.33
2.0	0.82	0.46	0.32

Tc : Surface cable tension
Tm : Mesh membrane tension

The SPADE results showed that the shape of the mesh membrane surface in 0G could be determined by the ratio of cable tension and mesh membrane tension and that the shape did not depend on the absolute tension of the cable and membrane. The higher the cable tension is relative to the mesh membrane tension, i.e., the higher the tension ratio, the lower the pillow distortion of the mesh surface is.

The topology of the cable network was selected according to the accuracy requirement. Even if the pillow distortion is zero, the mesh surface is not complete parabolic and consists of many flat planes. Therefore the surface error due to plane approximation is inevitable. This error depends on the focal length and topology of the cable network. Table 1 (1) shows the surface error due to plane approximation when focal length is 10.4 m, and Table 1 (2) shows the surface error due to pillow distortion in accordance with the tension ratio computed by SPADE. Table 1 can also be used for making this type of mesh surface for a C or Ku band reflector.

REFLECTOR ALIGNMENT MECHANISM

A large structure that consists of many module structures inevitably has shape error due to combining structures. The alignment error of the reflector due to the module connection was estimated by error analysis to be about 1 degree. An error as high as this would make it

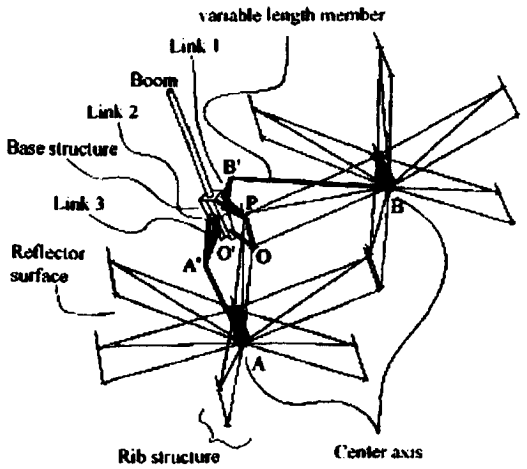


Fig.5 Reflector Alignment Mechanism

very difficult to assure the initial alignment of the reflector along the desired parabola axis. Therefore, a mechanism for alignment is indispensable at the attachment of the reflector to modify the initial pointing error. We have developed a support structure which compensates for alignment error by giving the reflector a little rotation. Figure 5 shows the structure. It is located between the reflector and the support boom and has four points connected to the reflector to get enough stiffness. This alignment concept has been verified with a prototype support structure. It turned out that the support structure is accurate enough to control the alignment of the.

HOLD-DOWN STRUCTURE

A hold-down structure (Fig.6) and a load path for the reflector protects the reflector in the stowed configuration against severe vibration and static load during launch. In orbit, the hold-down structure deploys and releases the stowed reflector. Figure 7 shows the load path that protects the mechanism of the reflector module and prevents the beams from breaking. The load paths are attached to both edges of each longitudinal beam of the reflector modules. In the stowed configuration of the reflector, load paths join with one another and form a rigid structure (load path assembly) as in Fig.6. The hold-down structure has some specific pins, which pre-stress the load path assembly and make it firm. The natural frequencies of the stowed reflector with hold-down structure are above 35 Hz in primal vibration modes. The hold-down structure and load path effectively protect the stowed reflector.

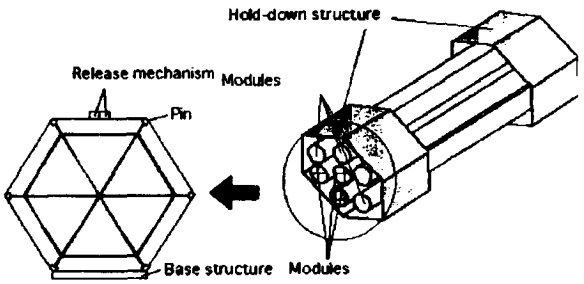


Fig.6 Hold-Down Structure

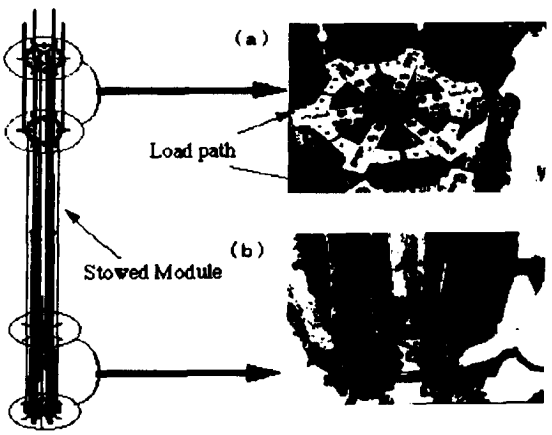


Fig.7 Load Path

NATURAL FREQUENCIES

The reflector has a lot of hinges and tension structures such as mesh surface and cable network. Therefore, it is necessary to consider the effects that hinge clearance and pre-stress of mesh surface have on the natural frequencies of the reflector.

One module configuration test and three module assembly configuration test were carried out to clear the effect of combining module on natural frequency. Modal analyses were performed for both configurations. It was found that the natural frequencies of both configurations can be predicted to within 10% accuracy. We also showed that it is possible to evaluate the natural frequencies for a larger number of modules using SPADE.

THERMAL TEST

To verify the thermal design of the reflector structure, a thermal balance test was performed on one module in a thermal vacuum chamber (13 m diameter) with solar simulator at NASDA (NAtional Space Development

Agency in Japan). The thermal mathematical model produced using OOCD was verified in this test. Thermal deformation measurements were also taken for this one module at another facility in which room temperature can be varied between –50℃ and +60℃. Under various thermal conditions, the shape of module structure was measured with V-STARS (photogrammetric instrument) and the results of the test were compared with the analysis. SPADE was shown to be able to predict thermal deformations accurately.

DEPLOYMENT TEST and ANALYSIS

Successful reflector deployment is critical to most missions. Therefore, it is equally important to be able to evaluate the reliability the deployment mechanism. The barriers to deployment are the tension of the deployment control cable, mesh surface tension, friction at the slider of the center axis and deformation of the beams. The effect that combining the modules has on the driving force margin for deployment must also be investigated. We performed deployment tests on one module, a three-module assembly and a seven-module assembly. The driving force margin for deployment can be monitored by measuring the tension of deployment control cable of each module even though the effect of friction can not be identified. The tests on the seven-module assembly were completed in August 1998. The reflector model was suspended from the ceiling at a height of 15 m by suspending cables in a cup-down configuration. A special control unit installed in the ceiling automatically maintained the desired tension of the suspending cable. Strains in the truss beams were measured during deployment. In the ground test, weight of module works against the deployment motion of module, while the tension in the suspending cables helps the deployment motion.

Figure 8. shows the deployment sequence. Seven modules were connected together, and their deployment characteristics were tested on the ground. The total diameter of the seven-module assembly is more than 10 m in a deployed configuration and less than 1 m in diameter in a stowed configuration. The tension reaches a maximum near the completion of deployment. It was found that the driving force margin deduced from the tension of deployment control cable does not depend on the number of modules. In other words, combining modules does not change the effects that the beam inner force and the friction at the center slider have on the deployment motion.

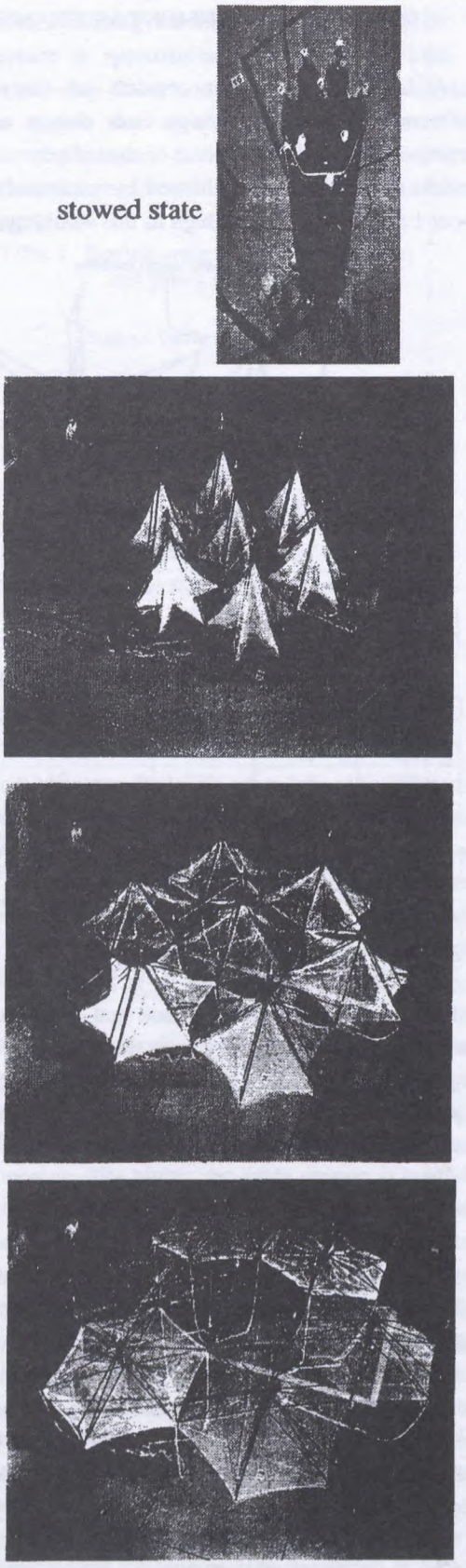


Fig.8 Deployment Sequence of 7-Module Assembly

Even though the effect of the slide friction can not be considered in analysis, the analytical results using SPADE indicate the same tendency regarding the combination of

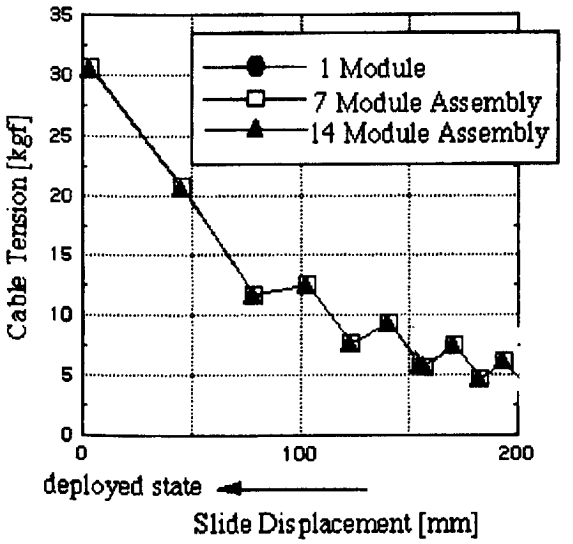


Fig.9 Tension of Deployment Control Cable

modules. Figure 9 shows the plot of the tension in deployment control cable of 1 module, 3-module, 7-module and 14-module assemblies as a function of the slide position along the center axis. The analytical results show that the tension of deployment control cable does not depend on the number of modules.

The above results of experiments and analyses show that the driving force margin of a multiple-module assembly can be deduced from the properties of one module. The deployment behavior of a multiple-module assembly can be estimated by SPADE. Finally, we suggest the simplified test on the ground, namely, only one module test is necessary to evaluate the driving force for deployment and the effects of friction that may hinder deployment.

CONCLUSION

NTT has developed a modular mesh reflector that consists of several module structures as well as a design program for such large space structures. The results of our experiments and analyses suggest simplified ground tests with one module making use of module concept, which can reduce the costs and the period of the development.

REFERENCES

[1] M. Minomo, H. Kazama and T. Itanami ., "Global Mobile N-ISDN Satellite Communication System", IAF-97-M.3.04, 1997.

[2] Artur B. C. and Robert F., "Low Cost Large Space Antennas", International Mobile Satellite Conference, 1998, pp.375-380.

[3] J. Mitsugi and T. Yasaka,"Deployable Modular Mesh Antenna - Concept and Feasibility", Proceedings of 17th Int. Symp. on Space Technologyand Science, Vol. 1, , 1990, pp. 599-604.

[4] M. Watanabe, J. Mitsugi, A. Miyasaka and M. Shimizu, "Large Space Antenna Structural Design Technology Status", IAF-98-I.1.01, 1998.

[5] Mark W. T., "The AstroMesh Deployable Reflector", International Mobile Satellite Conference, 1998, pp.393-399.

[6] J. Mitsugi, "Advanced in Simulation and Interaction Techniques", CIVIL-COMP Press, 1994, pp. 237-243.

[7] J. Mitsugi and T. Yasaka, "Nonlinear Static and Dynamic Analysis Method of Cable Structures", AIAA J., Vol. 29, 1991, pp. 150-152.

[8] J. Mitsugi, "Static Analysis of Cable Networks and their Supporting Structures",
Computers and Structures, Vol.51, No.1, 1994, pp. 47-56.

[9] K. Ando, J. Mitsugi, "Numerical analyses for the accurate design of large satellite on-board antennas", Computational Mechanics '95.,1995, pp. 296-301.

[10] M. Watanabe, A. Meguro, J. Mitsugi, and H. Tsunoda, "Module Composition and Deployment Method on Deployable Modular-Mesh Antenna Structures", Acta Astronautica, Vol.39, No.7, 1996, pp 497-505.

[11] J. Onoda , D.Y. Fu, and K. Minesugi, "Two-Dimensionally Deployable Hexapod truss", AIAA 95-1297, 1995.

[12] Y.Horiuchi, M. Inoue, K. Miyoshi, T. Sugimoto, T. Okamoto, and K. Hariu, "Deployment of Magnetically Suspended Sliders for Deployable Antenna Test Facility", 5th International Symposium on Magnetic Bearings, 5B, Japan Society of Mechanical Engineers, Aug. 1996.

[13] A. Flechais, P. Picard, C. Dauviau, and C. Truchi, "Dynamics of Large Reflectors AEROSPATIALE Concepts", 43rd Congress of the International Astronautical Federation, IAF Paper 92-0307, Washington DC, Sept.1992.

[14] J. Mitsugi, "Comparative Analysis of Deployable Truss Structures for Mesh Antenna Reflector", AIAA J., Vol.36, .No.8, TECHNICAL NOTES, 1998, pp.1546-1548.

A Real-time Dynamic Space Segment Emulator

P. Taaghol, H. M. Aziz, K. Narenthiran, R. Tafazolli, B. G. Evans
Centre for Communication Systems Research (CCSR)
University of Surrey, Guildford, Surrey GU2 5XH, UK
Tel: +44-1483-25 9810, Fax: +44-1483-25 9504
Email: P.Taaghol@ee.surrey.ac.uk

ABSTRACT

In this paper, an advanced Real-time Dynamic Space Segment Emulator (RDSSE), capable of emulating many satellite system characteristics is presented. The paper address the requirements for any such emulator and reports on the results of the work carried out on development of the most comprehensive space segment emulation system to date.

I. INTRODUCTION

Design and development of any satellite system requires regressive testing and optimisation of the satellite payload, User Terminal (UT), and the Land Earth Station (LES). In practice, most of the UT and the LES development and testing would have to take place in parallel with the development of the satellites under laboratory conditions. This implies that the impact of constellation dynamics, propagation effects, user mobility and many other characteristics of the eventual system on the performance of various system components cannot be fully investigated and measured until the system is in place. Furthermore, as various terminal manufactures around the world would be designing UT and LES units or components, it is of vital importance to be able to type-approve, test and optimise performance under realistic conditions through the use of Real-time Dynamic Space Segment Emulator (RDSSE) platforms.

Design of any such unit is constrained by several requirements, namely the real-time emulation of the propagation channel, dynamics of the constellation (changing delay, Doppler, etc.) and various other system characteristics demanding a fast yet accurate implementation at a designated IF/RF. The research and development work presented here has been partly carried out within SINUS (Satellite Integration into UMTS) and SUMO (Satellite-UMTS Multimedia Demonstrations), two European ACTS project dedicated to development and demonstration of a W-CDMA based satellite UMTS system. The presented RDSSE consists of a hardware and a software module. What makes the developed RDSSE unique is the advanced controller software. Consequently, the major part of the presented work is dedicated to describing various software modules.

The paper is organised as follows, the emulator requirements are identified in section-2. This leads to selection of the appropriate hardware configuration in section-3. Section-4 provides a comprehensive treatment of the advanced control software. Finally, the specification, performance, limitations and possible improvements to RDSSE are described.

II. REQUIREMENTS

The real-time emulation of propagation characteristics in the land-mobile satellite systems (exploiting dynamic satellite constellations) initiated this development. However, the developed emulator has come a long way since it was initially specified and is now much more capable than any commercially available system. As shown in *Figure 1*, the developed system is capable of real-time emulation of the propagation characteristics, payload non-linearities, antenna pattern, user mobility and many other system characteristics.

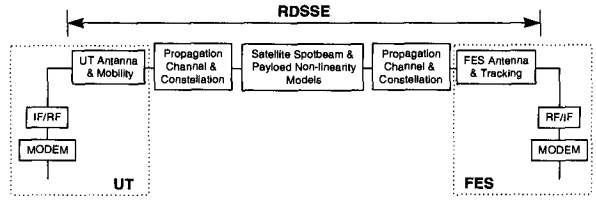


Figure 1: Functional model of a single emulator link

Dynamic satellite constellations are capable of providing multiple satellite visibility (satellite diversity). It is therefore important to emulate several links simultaneously. Different configurations of the emulator links can provide emulation of a wide range of scenarios. Amongst these configurations, the spotbeam and satellite handover and combined/switched diversity are the main configurations of interest.

Operational environment	*Open/highway, lightly wooded, heavily wooded, suburban and urban*
Frequency band	*L, S, X, Ka Ku and EHF*
Constellation	*LEO66, LEO48, MEO10, GEO or other user defined constellations*
Initial orbital phase	*Allowing user the choice of the constellation starting phase to enable the trials at worst case and best case elevation angles, etc.*
Satellite spotbeam configuration	*A complete set of satellite spotbeam configurations as well as antenna patterns in addition to possibility of having user defined configurations*
FES antenna and tracking error model	*A configurable feeder link tracking error model*
UT and FES co-ordinates	*Longitude and latitude*
UT speed	*Stationary to 300km/h*
UT route, direction	*Direction of the mobile movement and provisions for possible operational environment changes during a run.*
Satellite payload characteristics	*Payload characteristics such as the amplifier non-linearities*
Traffic distribution	*User defined traffic distributions enabling realistic real-time interference emulation*
Forward-link Doppler and delay pre-compensation	*Doppler and residual delay between two diversity links can be pre-compensated by the FES in the forward-link*

Table-1:Configuribility requirements

Through the use of a graphic user interface (GUI), the user can define a set of parameters some of which are listed above in Table-1.

III. HARDWARE EMULATION PLATFORM

There are a few hardware platforms capable of such real-time emulation, but their capability is largely constrained by the absence of a comprehensive software which generates the emulation sample points for any given desired real life scenario. The available hardware platforms can be categorised into two main families,

- *Real-time DSP-based emulation:* generates fast phase, amplitude, frequency changes internally based on a pre-defined set of statistical distributions.
- *Real-time play-back emulation:* create the desired impairments by pre-loading the fading sample points and FIR filtering of the input signal.

The former is the preferred platform as pre-loading the sample points would require large amounts of on-board memory which imposes limited run-time. On that note the NoiseCom emulation platform was selected, since it has combined characteristic of both the families.

The selected hardware consists of one NoiseCom Satellite Link Emulator (SLE-250) capable of emulating 2 full-duplex paths and two multipath fading emulators (MP-2700) each capable of emulating one full-duplex link as shown in *Figure 2*. The SLE-250 is capable of generating varying Doppler shift, propagation delay, path loss, and slow fading. Each half-duplex MP-2700 link emulates direct path component and two echoes with Ricean and Rayleigh fading statistic respectively. The complete RDSSE unit is hence capable of emulating two full duplex channels which represent the radio link between two satellites/spotbeams and the UT.

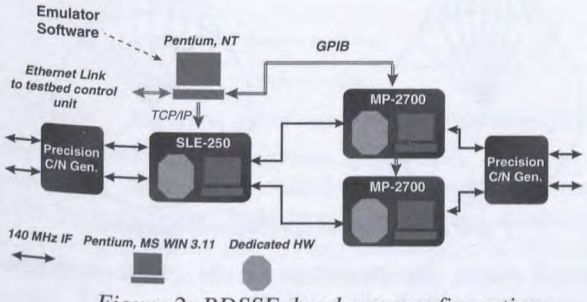


Figure 2: RDSSE hardware configuration

In order to be able to realistically emulate the relative characteristics of the two links within a given constellation, all the hardware units have been externally synchronised by the emulator controller unit. The two precision C/N generator units in *Figure 2*, enable introduction of dynamic interference and thermal noise during a run according to the configured scenario. The considered configuration is currently capable of coping with a maximum 10MHz input bandwidth, however, 20MHz upgrades are available now.

IV. DYNAMIC EMULATION SOFTWARE

a. Software Architecture

The RDSSE software consists of three main sections, the dynamic satellite constellation generator, the wideband channel model for all the environments and the elevation/azimuth angles followed by the interference generator module. Through the use of a graphic user interface (GUI), the user can define the desired set of parameters. The interaction between different software modules is depicted in *Figure 3*.

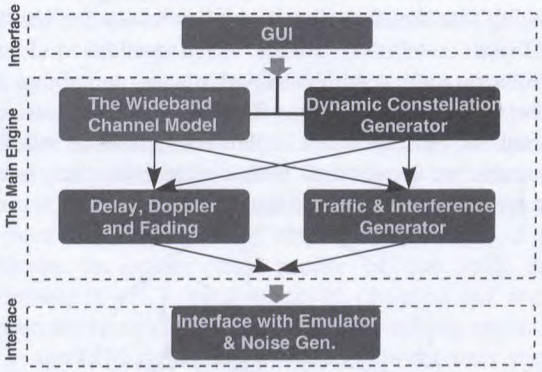


Figure 3: Software configuration

The dynamic constellation generator provides a series of relevant information such as the elevation angle of the two highest satellites/spotbeams, the azimuth separation angles, Doppler and delay (compensated or uncompensated), etc. to other software modules of the RDSSE controller. The RDSSE controller would then produce the necessary files for a given set of parameters. This file also includes other relevant information required by various hardware platform units to reflect the dynamic nature of constellation dependant changes in real-time.

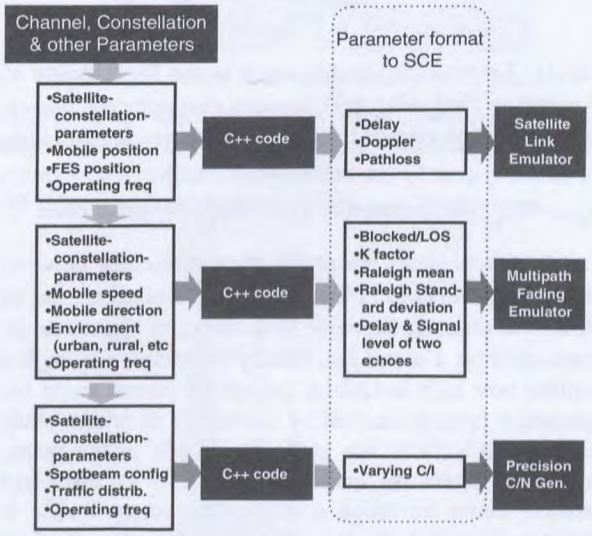


Figure 4: Sample point generation process

Figure 4, shows the logical input/output to and from the controller software unit more clearly. The format of the environment dependent parameters is that of a dynamic link library (DLL). By selection of the appropriate DLL (via the GUI), different operational environments can be emulated. This ensures complete flexibility for user defined

environmental DLLs to be simply plugged into the software if deemed necessary.

b. Constellation Generator

Within elliptical and circular orbits, only circular orbits are considered here. At present, there are two common methods, street of coverage method and spherical triangle method, for arranging satellites in a circular orbit constellation that result in efficient satellite coverage. Each method divides the satellites up into separate orbital planes containing equal numbers of satellites. *Figure 5*, shows the input parameters required to described a circular constellation. Additional parameters are required to define elliptical orbits, however these are not fully discussed in this paper.

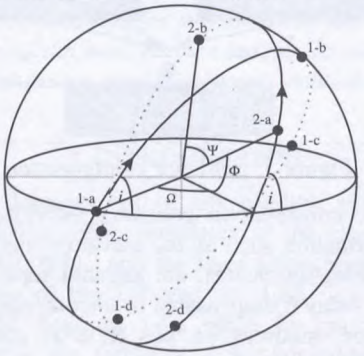


Figure 5: Satellite constellation structure

The angle Ψ represents the earth centred angle between satellites in the same plane. For a constellation plane containing N satellites Ψ is given by:

$$\Psi = 360/N$$

The angle Ω represents the difference in the Right Angle of the Ascending Node (RAAN) between two adjacent planes. For a constellation containing M orbital planes Ω can either be:

$$\Omega = 360/M \text{ for inclined or } \Omega = 180/M \text{ for polar orbit}$$

The angle Φ represents the phase angle difference between satellites adjacent planes. Consider plane 1 shifted around so that it lies on plane 2, the angle Φ is the earth centred angle between satellites 1-a and 2-a. Finally the inclination angle *i* determines how high in latitude the orbital planes travel i.e. the maximum latitude reached by satellite in an orbital plane corresponds directly to the inclination angle of the plane. With inclined orbits, the inclination angle can be optimised to provide better coverage over regions where traffic is expected to be very high. When the inclination angle of an orbit is less than 90° it is called a *pro-grade* orbit. Orbits with inclination angles greater than 90° are *retro-grade* orbits. When a satellite is inclined at 90° it is said to be in a *polar* orbit. Each orbital plane is defined separately in terms of the following parameters

- Altitude - this parameter specifies the height of the orbital plane above the earth's surface.
- Inclination angle - This inclination angle of the orbital plane as described above.
- Number of satellites - The number of satellites in the orbital plane. The satellites are distributed evenly within the 360° of the plane.
- Number of planes
- RAAN - Right Ascension of the Ascending Node. This angle is equivalent to Ω above. It is measured from the Vernal Equinox to the ascending node. The ascending node is the point where the satellite passes through the equatorial plane moving from south to north. Right ascension is measured as a right-handed notation about the pole.
- Mean Anomaly - represents the fraction of an orbit period which has elapsed since perigee. For a circular orbit the mean anomaly equals the true anomaly.
- Argument of Perigee (elliptical orbits)- The angle from the ascending node to the eccentricity vector measured in the direction of the satellite motion. The eccentricity vector points from the centre of the Earth to perigee with a magnitude equal to the eccentricity of the orbit.
- Eccentricity (elliptical orbits)- Defines the shape of the ellipse.

c. Spotbeam Antenna Models

The two main spotbeam configuration models commonly in use are variable beam width and equal beam width, which are shown in *Figure 6 (a) and (b)*, respectively.

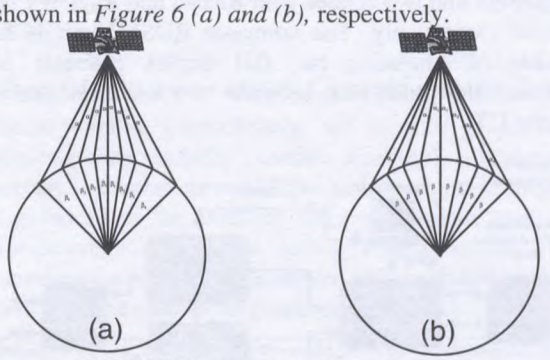


Figure 6: Antenna spotbeam arrangement

Figure 7 shows the projection of the above spotbeam configurations on the Earth's surface.

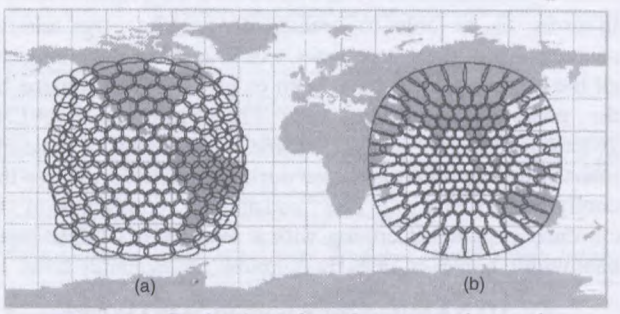


Figure 7: Projection of spotbeams on the earth

It can be seen that the in the equal beam width, case *(b)*, the distortion in the outer ring is relatively high. This can be compensated as shown in case *(a)* by varying the beam

widths of the satellite antennas. Both the above options have been coded into the software as described in [6].

d. The Propagation Channel Model

When a mobile user moves through the communication environment, the signal received by the user is blocked time to time due to LOS obstructions around the user. Furthermore, in non-geostationary satellite constellations the relative position of the satellite also changes causing similar shadowing impairments. This leads to a large drop in the signal strength commonly referred to as shadowing. Signal variation during shadowing is modelled using Loo's model [2]. In the proceeding sections of this paper, the shadowed and the non-shadowed cases are referred to as bad and good states, respectively. The bad and good duration are measured in terms of distance for a required environment and the measured values are incorporated within a two state Markov model for a single user-satellite link. Within any given states, there are signal strength variation due to the multipath effect. *Figure 8*, shows the narrowband representation of a typical received signal, highlighting various fading elements and parameters considered in the RDSSE software.

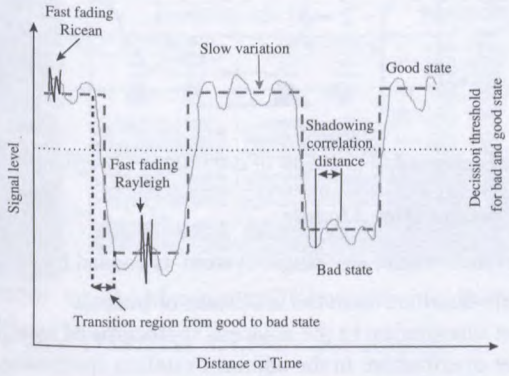


Figure 8: Received signal (narrowband representation)

The narrowband propagation parameters such as the shadowing statistics for various operational environments have been extracted from large experimental databases of University of Surrey [5], [7] and combined with that of the DLR [4] results in order to provide a harmonised European model. *Figure 9*, shows how the space and ground segment propagation effects can be treated separately.

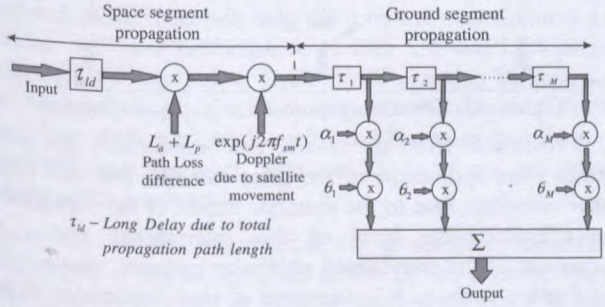


Figure 9: Complete Propagation Model

Note that as far as the path loss implementation is concerned, only the variation between the centre of the footprint and the terminal position are emulated due to the limited dynamic range of the attenuators within the hardware platform. However, external attenuators can be employed to increase the dynamic range if necessary. The ground segment propagation part represents the multipath effects and fading due to local reflections around the mobile. The wideband model of *Figure 9*, can be used for both shadowed and non-shadowed cases by just changing the parameters θ & α of each tape. The model parameters have been developed based on actual wideband recordings. Analysis of the results generally show delay spreads of 100ns or less. At higher elevation angles and open environments there are not many multipath components. As far as multipath is concerned, low elevation angles of the urban environment have been found as the most hostile environments. *Figure 10*, shows the power-delay profile of one such urban environment at 45°, L-band. It can be observed that even in the urban environment, all the resolvable echoes arrive very close to the LOS signal. *Figure 11*, shows the time aligned version of *Figure 10,* used in development of tap-delay-line models.

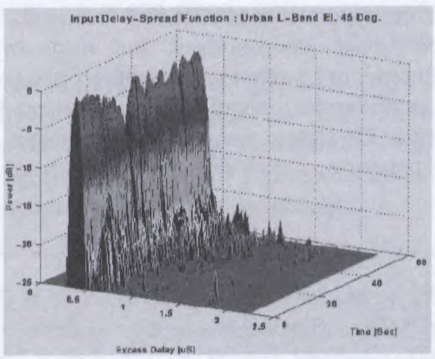


Figure 10: Power delay profile, urban, L-band, 45°

As shown in *Figure 11*, no more than 2 major reflections are encountered in this particular case. Furthermore, the average power in the reflected components are generally about 15-25 dB below the average power of the LOS component.

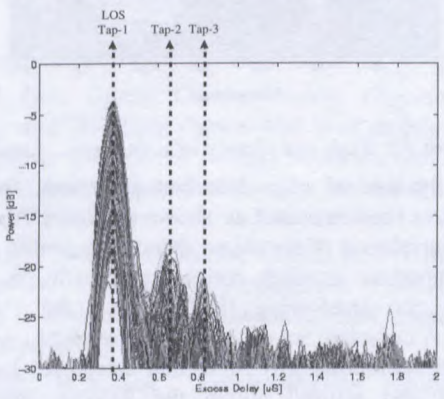


Figure 11:Time-aligned power delay profile, L-band, 45°

But nevertheless, at lower elevation angles the LOS is significantly attenuated resulting in much higher reflected powers (relative to the LOS). The LOS has been found to have a Rician and the reflected components a Rayleigh

distribution. It is important to point out that the relative delay of each tap varies in time as the distance between the mobile user and the reflector changes. Assuming conventional fixed tap-delay models one cannot hence test many system algorithms such as the performance of the RAKE receiver tracking algorithm. Therefore in tapped-delay-line model of *Figure 9*, the tap delays and their relative power does not stay constant for a duration of a run. Extensive analysis of the data has resulted in a definition of a new tap-delay-line models enabling realistic wideband channel representation [1]. A comprehensive set of tap-delay-line models for all the elevation angles and environments have been developed, enabling emulation of a wide range of operational environments.

e. Azimuth Correlation

Satellite diversity can be employed as an effective tool for combating shadowing and hence achieving higher service availability figures. However, in realistic diversity scenarios, some correlation between the shadowing characteristics of the diversity links exists. This correlation is not only dependent on type of constellation, azimuth separation angle, operational environments, but also on the UT location. In order to be able to represent this accurately, fish-eye pictures of various operational environments have been taken. *Figure 12*, shows one such picture taken in central London, representing a typical European urban environment.

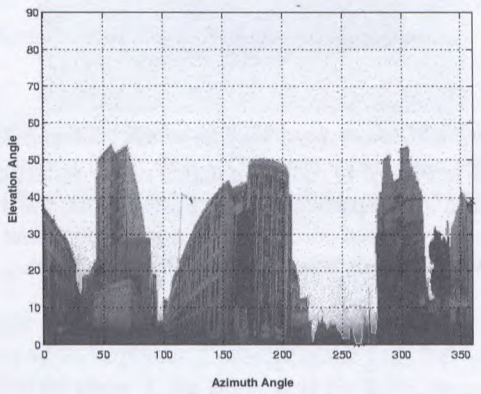


Figure 12: Fish-eye picture of urban env., London

Through the use of edge detection algorithm, shadowing profiles have been extracted as shown in *Figure 13*.

The autocorrelation of the above shadowing profile provides the all important azimuth correlation coefficient used to determine the shadowing likeliness of the diversity channels. A complete set of azimuth correlation coefficients for all the operational environmental categories have been developed and utilised within the RDSSE software to represent realistic diversity scenarios. In order to incorporate, the azimuth correlation, the four state model proposed by Lutz [3] was utilised. Full detail of the Morkov model can be found in this reference. *Figure 14* shows the correlation coefficient of two satellites at 45° elevation angle with variation in azimuth angle.

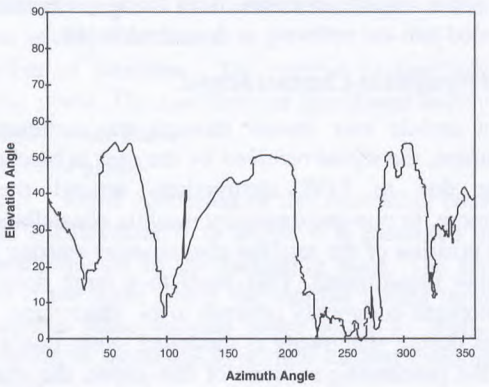


Figure 13: Shadow profile of urban env., London

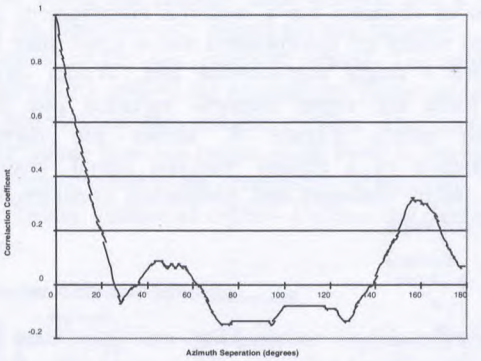


Figure 14: Variation of correlation coefficient

f. The Interference Module

Interference within any single system is caused by:

- users distribution in the spotbeam of interest
- user distribution in the adjacent spotbeams of satellite
- user distribution in the adjacent satellite spotbeams

as shown in *Figure 15*.

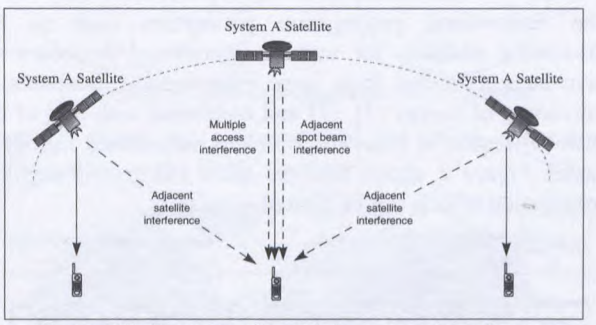


Figure 15: Sources of multiple access interference

It is realisable from the above, that interference not only comes from spotbeams of the same satellite, but also from other satellites. Due to the dynamic nature of the considered constellations, the level of this interference varies as formation of the overlapped spotbeam patterns move over the Earth's surface. It is therefore of vital importance to be able to test and evaluate the performance of system under such conditions. The considered interference module generates the real-time varying interference based on a

comprehensive set of input parameters. *Figure 16*, shows an example of the varying C/I for a LEO-based system.

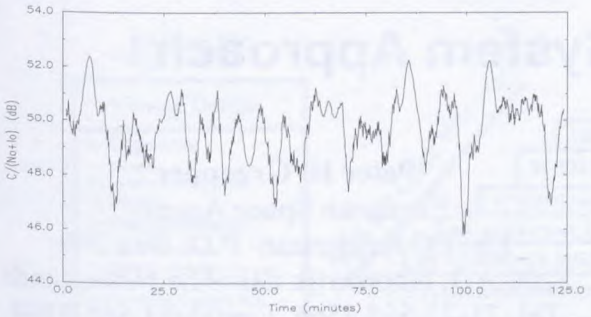
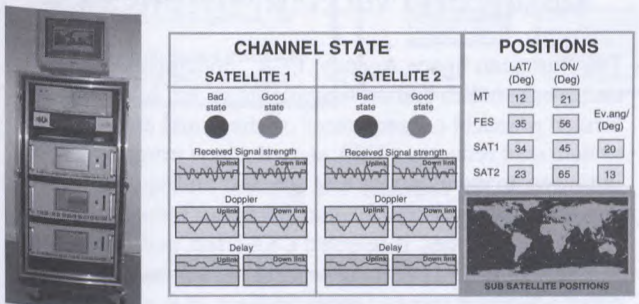


Figure 16: Real-time C/I variations in a typical LEO

V. PERFORMANCE SUMMARY

In this paper, we reported on the development of an advanced Real-time Dynamic Space Segment Emulator (RDSSE), capable of emulating any mobile satellite system in real time. The complete RDSSE is shown in *Figure 17(a)*. The controller PC displays the channel state during a run for monitoring purpose as shown in *Figure 17*(b).



(a) RDSSE (b) Monitoring GUI windows for RDSSE

Figure 17: RDSSE hardware and display unit

The software has a comprehensive wideband land-mobile satellite channel which includes a varying tapped-delay-line model (varying delay and power of echoes with time). The model also takes into account the azimuth correlation between the diversity channels, enabling realistic and unexaggerated satellite availability. The RDSSE is capable of testing the performance of existing and future mobile satellite system, User Terminal (UT), and the Land Earth Station (LES) before operation, without the need to have the system in place. This will help the terminal manufactures to test and optimise performance of their UT and LES units under realistic conditions. The RDSSE can also be used to test the signalling, call set-up procedures, power control algorithms, dual satellite diversity and handover procedures (inter-spotbeam handover, inter-satellite handover and even inter-segment handover).

Figure 18, shows an example of a satellite handover scenario. As it can be seen, the power in link-a (solid line) starts deteriorating almost halfway through the run. On the other hand, the upcoming satellite of link-b (dashed line) appears to be increasingly received at a higher level until the crossover point, when handover should take place. Note that both the links experience uncorrected shadowing in this environment.

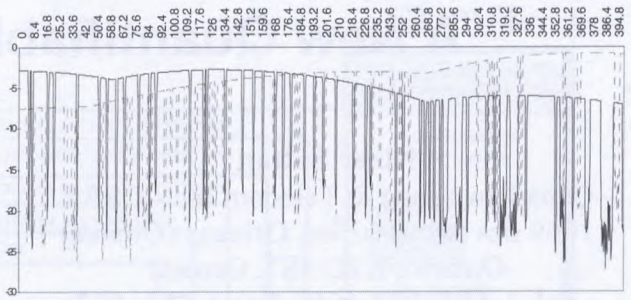


Figure 18: A typical handover scenario

Since, the unit currently consists of two full-duplex channels, the handover performance together with diversity cannot be tested simultaneously. A third full duplex link is therefore being acquired to increase the emulator capability.

Furthermore, the atmospheric propagation characteristics software module is currently being developed for emulation of a rain, scintillation and cloud attenuation at Ka, Ku and EHF bands.

The emulator can be further expanded to include a terrestrial cellular system emulation module, for real-time emulation of propagation and signalling between the satellite and terrestrial segments. This would enable execution of inter-segment handover under realistic conditions.

VI. REFERENCES

[1] A.Jahn, et. al., "Channel Characterisation for Spread Spectrum Satellite Communications", IEEE Fourth International Symposium on Spread Spectrum Techniques and Applications (ISSSTA'96), pages 1221-1226, 1996.

[2] Chun Loo; A statistical model for a land mobile satellite link; IEEE Trans. On Vehicular technology, Vol. VT-34, No. 3, Aug 1985.

[3] Erich Lutz; A Markov model for correlated land mobile satellite channel; International journal of satellite communications, Vol. 14, 333-339,1996.

[4] Erich Lutz, Daniel Cygan, Michael Dippold, Frank Dolainsky, and Wolfgang Papke; The land mobile satellite communication channel recording, statistics, and channel model; IEEE transactions on vehicular technology, Vol 40, No 2, May 1991.

[5] M A N Parks, B G Evans, G.Butt and S Buonomo; Simultaneous Wideband propagation measurement applicable to mobile satellite communication systems as L-band and S-band; AIAA96

[6] J.Wertz, W.Larson, "Space Mission Analysis and Design", Kluwer Academic Publishers, ISBN 0-7923-0971-5

[7] P.Taaghol, R. Tafazolli,, 'Correlation model for shadow fading in land-mobile satellite systems', IEE Electron. Lett., *1997, vol. 33, No. 15, pp. 1287-1289*

Mobile Satellite Life Cycle Cost Reduction: A New Quantifiable System Approach

Nizar Sultan
Canadian Space & Telecom Inc.- CSAT
1669 des Broussailles, Orleans (Ottawa)
Ontario, K1C 5S7, Canada
Tel. 1-613- 837 4548, Fax 1-613- 837
csat@sympatico.ca

Peter H. Groepper
European Space Agency
ESTEC Keplerlaan- P.O. Box 299
2200 AG, Noordwijk ZH- The Netherlands
Tel. 31-71 565 4566, Fax 31-71 565 5184
pgroeppe@estec.esa.nl

ABSTRACT

There has always been a need to predict potential cost reductions in communications satellites. As mobile satellite owners are becoming also the service providers and expect the best return, it is not adequate to reduce only the space segment cost, often using cost models. This paper addresses the potential cost reduction of the space mission life cycle, from requirement analysis to the disposal of spacecraft. Since cost effectiveness is a life cycle optimization issue, a system approach to trade off is developed here and applied to mobile satellite missions similar to Inmarsat 3. Such a tool, not available in the open literature, provides a powerful means to perform detailed design/cost trade–off analyses of the entire mobile system, as shown by the results. It can be adapted to any other communication satellite systems.

1- INTRODUCTION

Eighty per cent of the global space market expenditure is currently in the commercial mobile and fixed communications satellites, for medium and large size spacecraft, with increasing powers, complexities and international teaming partnerships. There is a need to reduce the traditional space programs schedules and avoid redesign due to inadequate system approach, both leading to delays and cost overruns. In addition, the high cost of satellite constellations, combined with the need to secure funding from potential investors, is leading to the necessity, not only to reduce the spacecraft cost, but also the life cycle cost and schedule of the whole mobile satellite system.

2- A NEED TO PREDICT THE SPACE MISSION LIFE CYCLE COST EFFECTIVENESS

The European Space Agency, ESA[1], concluded that the numerous studies and tools available so far, have had limited practical consequences on the actual end-to-end system cost reduction. ESA started a new program initiative, to develop new and specific findings for a more significant impact on ESA and European industry corporate practices. It selected CSAT Inc. and Microcosm Inc. to lead and implement the first of such studies. This paper addresses some of these findings. It is worth noting that NASA is setting up a 25-year program, involving space cost reduction and other integrated approaches.

There are a few "Cost Estimating Tools", developed for spacecraft and sometimes for limited types of payloads, but none that address the life cycle space program cost, from the initial mission requirements and definition, the ground segment, the operations, right to the disposal of the spacecraft.

Cost reduction in itself is meaningless, unless the issue of cost effectiveness and figures of merits are addressed. Typically, a business case is pursued only after acceptable returns on investment is clearly demonstrated by predicting the net benefits. In the mobile satellite case, where bandwidth at L-Band is a precious resource, one of the returns can be the maximum profit per MHz, as an example.

2- KEY COST DRIVERS

One of the results of our ESA study is the identification and grouping of the life cycle key cost drivers for space missions. These drivers are summed up in Fig. 2, in their order of importance.

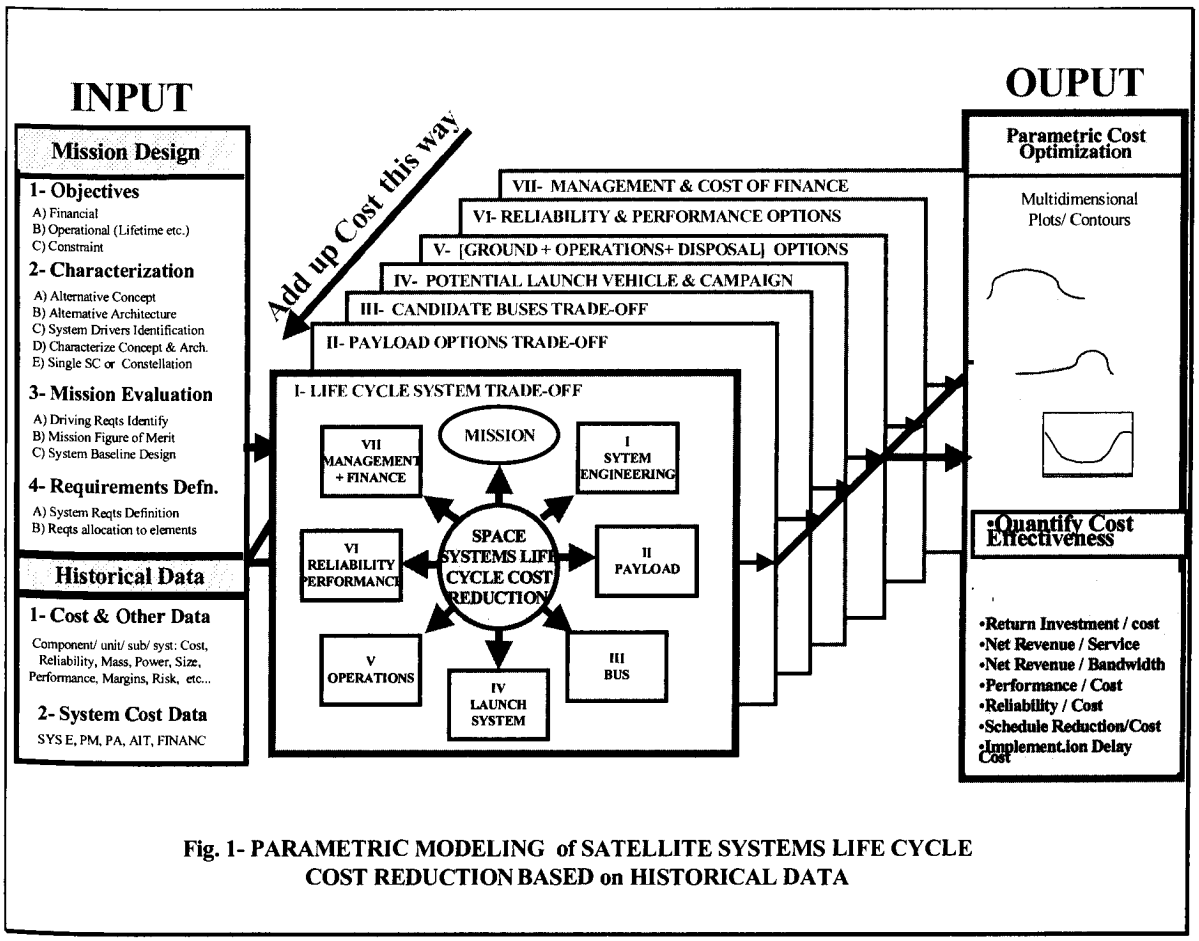


Fig. 1- PARAMETRIC MODELING of SATELLITE SYSTEMS LIFE CYCLE COST REDUCTION BASED on HISTORICAL DATA

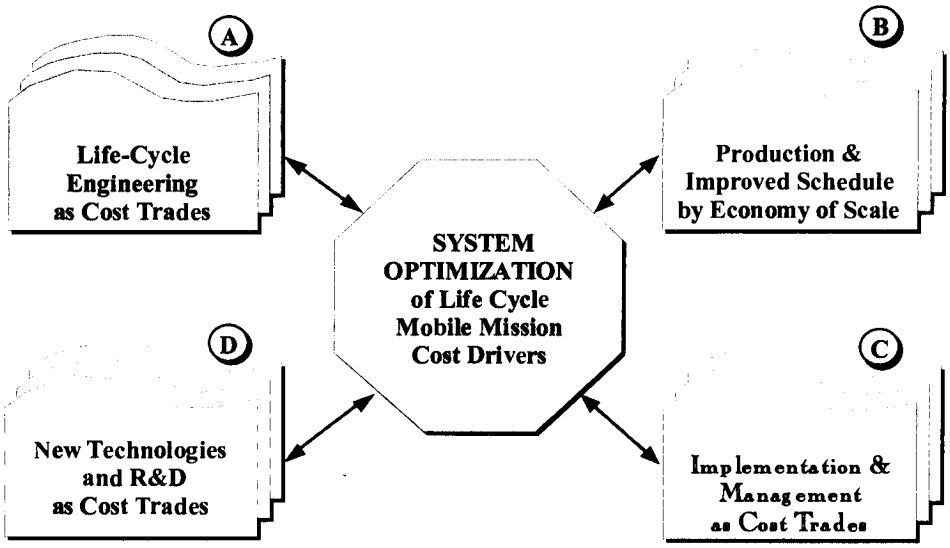


Fig. 2- Space Missions Life Cycle Key Cost Drivers.

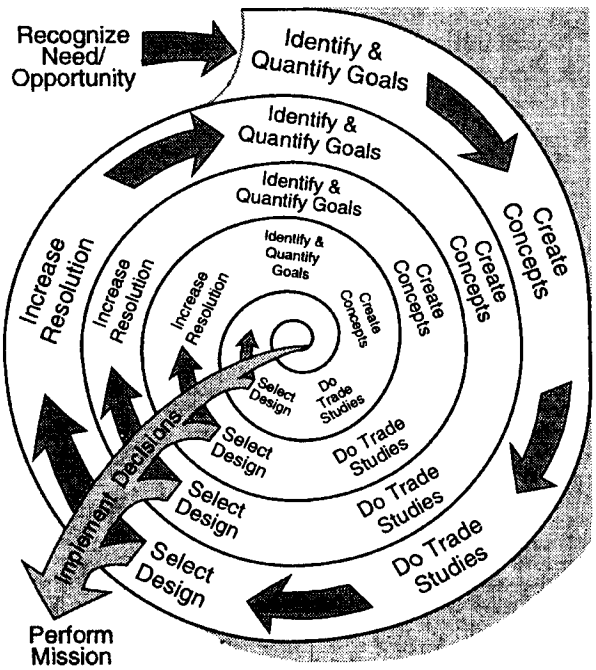


Fig. 3- The Doctrine of Successive Refinement (from R. Shishko[3])

Drivers A, B, C and D are described briefly below:

- 'A', the engineering trade off analysis is the key driver that system planners can control at the program onset, as is shown below

- 'B' can achieve the most significant cost reduction, if "quasi mass production" is possible, by a factor of 2 for quantities of fifteen spacecraft.

- 'C' can achieve up to 35% cost reduction in medium and large space missions, if the program is managed and implemented in the most efficient way. Of course, for small satellites, much higher cost reductions have been achieved[2], due to a better control over teams, resources and facilities.

- 'D' provides the risk and cost reduction benefits attributed to R&D and new technologies before a program is started. This involves initial investment and later benefits recovery. Again, experience has shown, that barring an outstanding new invention, most R&D in new technologies can result in up to 35% cost reduction and over a period of some ten years, excluding the cost of research.

3- PARAMETRIC MODELING OF MOBILE SATELLITE SYSTEMS LIFE CYCLE COST EFFECTIVENESS

This paper concentrates on an illustration of a system tool, applied to mobile satellites and that can achieve significant increase in systems cost effectiveness.

An excellent illustration by Shishko[3] in fig. 3 highlights the great importance of pre-Phase A and Phase-A continuous trade-off analyses to increase the program life cycle cost effectiveness. More and more spacecraft builders are becoming service providers. The traditional approach of reducing the space segment cost as much as possible without evaluating the impact on the end-to-end cost effectiveness is no longer acceptable.

Most available cost reduction approaches advocated so far are not quantifiable, and apply mostly to smallsats. As NASA recently discovered[4], Faster, Cheaper is not necessarily Better.

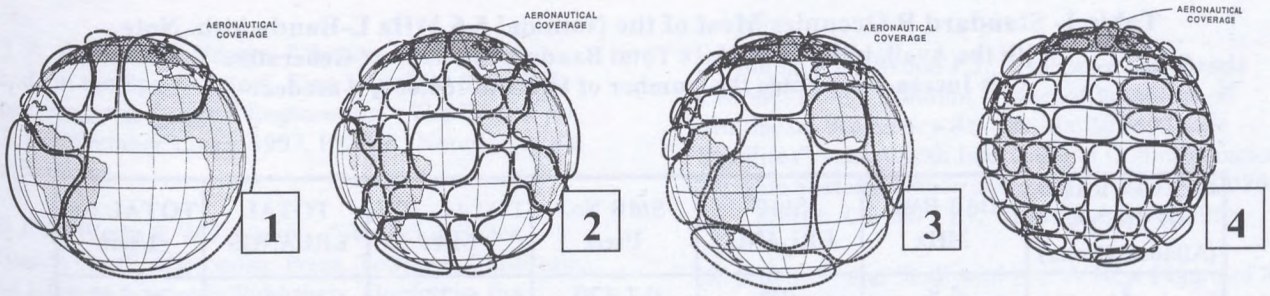
It is true that key cost reduction approaches, such as better planning, efficiency and motivation cannot be quantified. But, nevertheless, there is an urgent need to quantify and predict the complex life cycle space system cost, in order to convince potential investors to provide billion dollar funds.

In modeling, as in fig. 1 the timing of the trade off analysis is very important: extensive trades must occur at the earliest stages of the program. A lot of iterations are required in the mission design, such as characterization, evaluation and mostly the requirement redefinition.

Another key issue is that any quantification of space systems cost reduction, to be convincing, must be based on historical data of actual programs. CSAT is developing new modeling algorithms for commercial communications satellites, in spite of the limitations of predicting future costs based on past programs and the difficulty of finding sufficient historical data.

4- LIFE CYCLE MODELING APPLIED TOGLOBAL MOBILE MISSIONS, SUCH AS INMARSAT 3 or 4.

Previous results applied to Inmarsat life cycle optimization combine the author's contribution to Inmarsat 3 planning with the knowledge acquired since from published data on actual Inmarsat 3 implementation[5-6]. Therefore, the examples shown here and in previous references[7-9] do not represent Inmarsat 3 views. They are for the purpose of illustration of the possibility of predicting life cycle cost effectiveness.



6 Beams	19 Beams	18 Beams	52 Beams
4 Transponders	7 Transponders	11 Transponders	19 Transponders
7 Feeds	7 Feeds	19 Feeds	19 Feeds

Fig. 4- Four Configurations 1 to 4, for Global Coverage by 3 or 4 Identical Geostationary Mobile Satellites (Maritime, Land and Aeroronautical,)

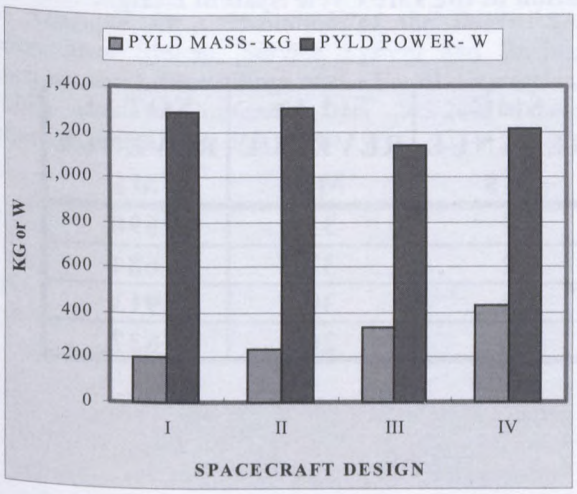


Fig. 5- Payload Mass and Power for the Four Designs

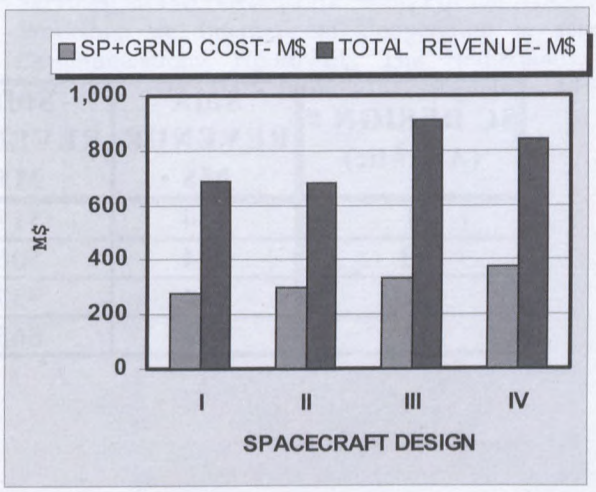


Fig. 6- Cost and Revenues over 10 years. Design III is the most cost effective configuration

The global mobile satellite system assumed here is shown in fig. 4. Essentially, spacecraft 1 design is, in general, to the best of our knowledge, the closest to the actual Inmartsat 3 configuration. A very large number of parameters have been traded off at length over wide ranges of values.

5- EXAMPLES OF RESULTS

For brevity, a couple of results are shown in figs 5 & 6, and in tables 1 & 2.They all point to the fact the least costly spacecraft is not necessarily the optimum design to choose. The issue is the cost effectiveness and optimization of the space, ground and operations that can predict the optimum solution[10]: clearly, configuration III is the most promising.

Table 1- Standard B Occupies Most of the Nominal 5.5 MHz L-Bandwidth. Note With the Availability of 34 MHz Total Bandwidth, the Next Generation Can Increase Six Folds the Number of Users to 150,000, if needed.

SPACECRAFT DESIGN # (Atlantic Ocean)	StdB BW-MHz	StdB ERLANG	StdB No. Users	TOTAL BW-MHz	TOTAL ERLANGS	TOTAL No. Users
I	4.8	305	17,429	5.73	454	25,943
II	4.5	302	17,257	5.38	447	25,543
III	4.7	440	25,143	5.50	581	33,200
IV	4.4	397	22,686	5.13	527	30,114

Table 2- 10 year Revenue Distribution Amongst all Services. Standard B is Dominant in the Cost effectiveness Optimization of the Life Cycle System Design.

SC DESIGN # (Atlantic)	StdA REVENUE-M$	StdB REVENUE-M$	StdC REVENUE-M$	Std Ae REVENUE-M$	TOTAL REVENUE-M$
I	144	511	2	33	**690**
II	144	506	2	32	**684**
III	144	737	2	30	**913**
IV	144	665	2	26	**837**

Since mobile standard B is the economically dominant service, design III can satisfy 25,000 standard B users with just under 5 MHz bandwidth. For a possible Inmarsat 4, 34 MHz total bandwidth can provide up to 150, 000 standard B, if it is still in use in 10 years' time.

6- CONCLUSION

A cost-estimating tool is necessary, but not sufficient, to analyze life cycle cost reduction. That is why a "parametric system" approach was developed to quantify and predict cost effectiveness before the space mission program is started. This should reduce significantly rerun and cost overruns. Such a methodology is extended elsewhere to demonstrate and illustrate some of these features

ACKNOWLEDGEMENTS

The opinions expressed in this paper do not necessarily represent those of the European Space Agency. Some of the data used came from earlier work performed by the author under the sponsorship of the International Maritime Organization (Inmarsat) and Canadian Astronautics Ltd, on Inmarsat Third Generation system trades.

REFERENCES

[1] A. Atzei & M. Novara, " System Engineering trends in the Space sector", First Joint ESA/INCOSE Conference on Systems Engineering"- The Future, pp. 1.3.1, November 11-13, 1997, ESTEC, Noordwijk, the Netherlands.

[2] J. R. Wertz & W.J. Larson, " Reducing Space Mission Cost", Microcosm Press, Torrance, California, and Kluwer Academic Publishers, Dordrecht, the Netherlands, 1996.

[3] R. Shishko, "NASA Systems Engineering Handbook", NASA publication SP-6105, 1995, p. 7.

[4] David, " Is Faster, cheaper Better?", Aerospace America, September 1998.

[5] A. Howell & G.V. Kinal, " Inmarsat's Third Generation Space segment", 14th AIAA International Comunications Satellite Systems Conference, Washington, DC, USA, March 1992.

[6] Sengupta J.R., "Evolution of the INMARSAT Aeronautical System, Service System and Business Consideration", Proceedings of the Fourth International Mobile Satellite Conference, pp. 245-249, Ottawa, Ontario, Canada, June 6-8, 1995.

[7] Sultan N. & Wood P.J., "Adaptive Sub-Bands Channelization: Solution to Reconfigurability of Multibeam Frequency Re-Use Maritime Mobile Satellites", AIAA 12th International Communication Satellite Systems Conference, Paper AIAA -88-0769, pp. 157-166, Crystal City, Virginia, March, 1988.

[8] Sultan N. and Shallwani A., "A New Figure of Merit for Modeling and Optimizing Advanced Mobile Satellite Systems", Fifth International Conference On Satellite Systems For Mobile Communications and Navigation, London, UK, May 1996.

[9] Sultan N., "Payload, Bus, and Launcher Compatibility, for Multibeam Mobile Communication Satellite Systems", AIAA Progress in Astronautics and Aeronautics Series, Volume 128, Space Commercialization: Satellite Technology, August 1990.

[10] D. Hernandez, T. Cussac, R. Ecoffet, J. Foliard & M. Thoby, " Constellations for Mobile Satellite Services, an Overview", Proceedings of the AIAA/ESA Workshop on International Cooperation in satellite Communications, Noordwijk, The Netherlands, pp. 219- 229, March 1995, ESA SP-372, September 1995.

Development of High-Power Laser Link for OISL Terminal Applied to Mobile Communication

Asoke Ghosh[1], Rupak Changkakoti[1], Peter Park[1],
Robert Larose[2], Jocelyn Lauzon[2] and Stefan Mohrdiek[3]

[1]MPB Technologies Inc., Pointe-Claire, Quebec, Canada H9R 1E9,
[2]Institut National D'Optique, 369 Franquet, Quebec, Canada G1P 4N8
[3]Uniphase Laser Enterprise, Binzstrasse 17, CH-8845, Zurich, Switzerland

E-mail address: ghosh@mpbtech.qc.ca

ABSTRACT

A high power laser transmitter module is being developed for Low Earth Obit (LEO) to Low Earth Orbit (LEO) Inter-satellite Link (ISL) application in collaboration with Institut National d'Optique (INO) and Uniphase Laser enterprise (ULE). It is based on the Semiconductor Laser Ytterbium Booster (SLYB) concept. SLYB uses a polarization maintaining double clad Yb-doped fiber amplifier (YDFA) with counter propagating 917 nm pump laser diode and directly modulated (1.5 Gbps in NRZ modulation format) 985 nm signal laser. The overall output power is predicted at greater than 600 mW which is sufficient for link distances of 7000 Kms assuming 10^{-9} BER and a low noise Si APD receiver. Preliminary space radiation susceptibility tests indicate an additional 1.5 dB end of life loss.

1 INTRODUCTION

In recent years there has been an escalating demand on mobile satellite communication due to an upsurge in data communication. To support this high data rate communication the market trends have placed satellite communication in the forefront for global connectivity. Several satellite communication system companies listed in Table 1., are launching satellites in tens or hundreds to provide cellular mobile radio and personal communication systems (PCS) to the user. It is predicted that many of these systems will provide voice, data, paging, Facsimile, Video and multimedia communication globally in the near future. However, due to the high bit rate demands for such applications, communications using the microwave bands puts high demands on the size and weight of the satellite systems. Optical inter-satellite link (OISL) with its various advantages over its microwave counterpart can relieve the industry of some of these major bottlenecks. From table 1 the potential market for OISL's can be assessed with regards to the quantity and invested capital. A comparative assessment of some of the key factors for microwave and optical communication systems is presented in table 2. OISL for satellite communication can offer several potentially valuable applications for GEO-to-GEO, LEO-to-GEO and Mobile Satellite Services (MSS). Figure 1 presents a conceptual view of a global communication network where the end user on ground is connected to the LEO satellite using L-band or K-band microwave link and to the rest of the globe via OISL.

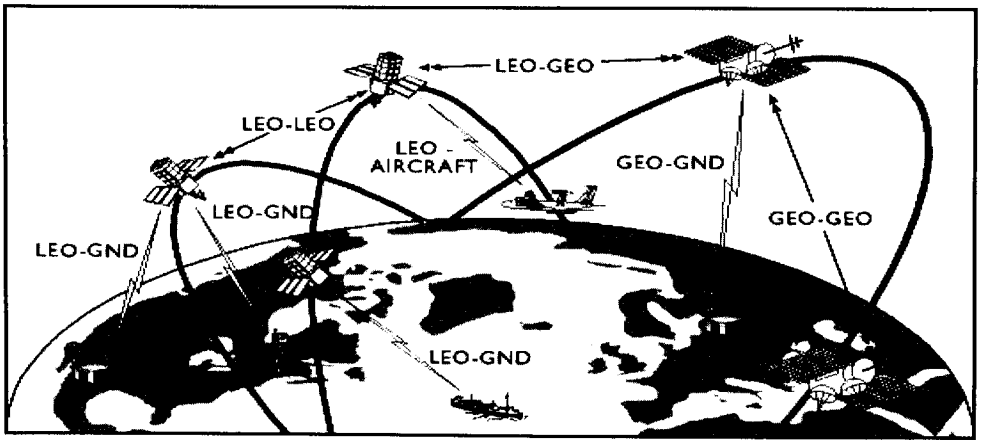


Figure 1: A mobile satellite communication architecture

Table 1: Comparison of features of various commercial satellite constellations for future global communication

Satellite Constellation	No. of Satellites	ISL/ No. per satellite	Link distance	Services	Transmitter Data rate	Total system cost (in billion US $)
Iridium	66 (+6 spares)	6	4000 Kms	Voice, Data, Paging , Facsimile	1.00 Gbps	4.4
Teledesic	288	8	2600 Kms	Voice, Data, Facsimile, Video and Multimedia	1.53 Gbps	9.0
Globalstar	56	None		Voice, Data, Paging, Facsimile		2.6
Spaceway	16	4 (+1spare)	84,000Km	Global broadcast, Video, Global multimedia services	1.00 Gbps	3.2
Astrolink	9	2		Data, Video and rural telephony	340 Mbps	4.0
Expressway	14	2				3.9

This paper presents the ongoing developmental work for a European Space Agency (ESA) funded program on High power laser transmitter for LEO-to-LEO OISL application and addresses future plans for LEO-to-GEO and GEO-to-GEO OISL's. Chosen System concept, expected performance characteristics and space qualifiability aspects of the existing system would be discussed in detail.

2 SYSTEM CONCEPT

2.1 Trade-off analysis

Based on the Laser transmitter requirements for a LEO-to-LEO OISL, a comparative assessment of the existing technologies was carried out. The technologies compared are: (a) 840 nm fiber pigtailed laser diode, (b) 1550 nm Erbium-doped fiber amplifier (EDFA) and (c) 985 nm SLYB. At one stage of this assessment MFA-MOPA was considered due to its high output power capability. But learning that the manufacturers SDL Inc., have spent significant time and money on the reliability of this product without achieving more than few thousand hours of median life this technology was dropped for our final technology assessment. In table 3 a summary of our assessment is presented where the technology trade-off analysis are carried in terms of maturity, space qualification and reliability, and commercial status of the technology. From our assessment, taking the above mentioned factors into consideration the 985 nm SLYB system concept was selected. It had no foreseen major fundamental limitations. Potential disadvantages would be high cost and custom made components associated with this approach. Other issues of concern can be the space qualification due to radiation induced loss of GRIN lenses and the Yb-fiber.

2.2 SLYB system design

The chosen SLYB system configuration is schematically presented in figure 2 where the signal and pump beam propagate counter to each other in the fiber. The 985 nm signal generates a directly modulated signal which is amplified in the Yb-doped fiber booster by the counterpropagating pump power generated from the 917 pump source. Polarization maintaining isolators are placed at the output of the signal laser, pump laser or the transmitter to avoid any back reflection to the lasers. The pump beam is coupled to the inner cladding of the double clad Yb-fiber using a custom made WDM coupler/ combiner.

3 EXPECTED SYSTEM PERFORMANCE CHARACTERISTICS

The scope of this program is to build a laser transmitter which can provide a modulated signal and be able to connect two LEO satellite systems at distances from 4000 –10,000 Kms with Bit error rate (BER) of 10^{-9}. Link power budget calculations for the system revealed that 600 mW would suffice to close the link [2]. The expected performance specifications for the SLYB laser transmitter are given in table 4 below.

Table 2: Comparison of the RF and Optical space communication system [1]

Factors	RF communication system	Laser Communication system
Type	60-GHz solid-state power amplifier	GaAlAs multiple diode
Transmit power	2 Watts	Approximately 1 watt
Antenna/telescope aperture	48" antennas	5.5" aperture gimbaled telescope
Weight per terminal	145 lb including boom (20 lb)	106 lb
Power required / terminal	140 W	126 W

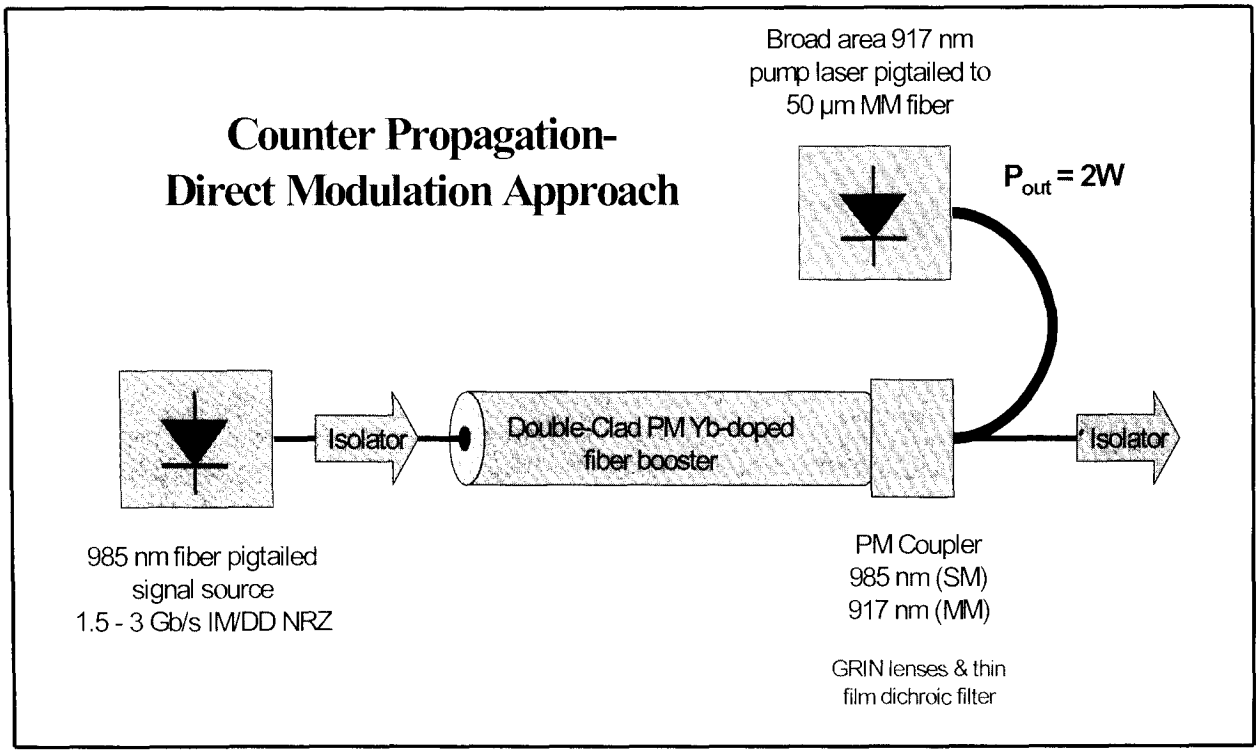


Figure 2: Schematic diagram of the SLYB laser transmitter

Table 3: Summary of technology trade-off analysis

System concepts	Status of technology	Space qualification and reliability	Commercialization
840nm fiber pigtailed laser diode	Very mature.	Very reliable. Space qualified or nearly space qualified	Only one known supplier of the high-power 840nm laser diodes (SDL inc.).
980nm SLYB	Available commercially. Some may be custom-made components.	No data on 917nm high-power pump laser reliability. Preliminary simulated space radiation tests of Yb-doped fiber and components with GRIN lenses do not show high radiation induced loss for a 10 Yr mission.[3]	No supplier of the 980nm SLYB. Potential suppliers (more than one per element) have been identified for all sub-components.
EDFA based system without Er fiber optical pre-amplifier	Er powers boosters with outputs more than 20dBm do not exist commercially to our knowledge. Largest power value published from such a component is 27dBm [4].	Except for the power boosters which is an issue for reliability and space qualification, the rest of the associated components are very mature and are close to being space qualified.	A multitude of suppliers for all the elements except the power booster. No supplier for the Er power booster at those power levels. Potential suppliers have been identified however.
EDFA based with (Er/Yb) booster system without optical pre-amplifier	Er/Yb power boosters with outputs of 30dBm are available commercially.	The commercially available power boosters are not tested for reliability. Phosphorus dopant in the Er/Yb fiber an issue for space radiation qualification.	A multitude of suppliers for all the elements except the power booster. As for the power booster, there are IPG, Pritel and maybe SDL Inc. as potential suppliers.
EDFA based system with optical preampli-fier	Very mature except for power booster.	Er power boosters an issue for reliability and space qualification, the rest of the associated components are very mature.	A multitude of suppliers for all the elements except the power booster. No supplier for the Er power booster at those power levels.

Table 4. Expected performance characteristic of the SLYB laser transmitter

Parameters	Expected characteristics
Output Power	>600mW
Wavelength	985 nm
Spectral Width	<2.0nm (FWHM FP laser)
Spectral Stability	<+/-1.0nm
Pointing Stabitity at the collimator	+/- 2μrad
Polarisation	Linear, extinction 20dB
Modulation	Intensity, NRZ direct modulation
Data Rate	>1500Mbps
Power Consumption	8W (no pump TEC)
Mass	0.9kg
Size of transmitter	15x12x4.3cm^3
Fiber Pigtail output	Singlemode PM fiber

4 SPACE RADIATION QUALIFICATION OF SLYB LASER TRANSMITTER

As one of the most important requirement for any space system is its space radiation hardness, a detailed assessment of individual component for their space readiness was carried out. Among all the constituent components (a) the Yb-doped fiber and (b) GRIN lenses used in the WDM combiner and Isolators were identified as being critical for the functional performance of the system over the 10 year mission life. Preliminary radiation testing of these critical components were carried out using ^{60}Co Gamma radiation source. The required radiation dose was calculated using Radmodls program considering all the metallic shields provided to encase the system. The expected radiation dosage for the chosen LEO orbit was determined as 110 Krads for a period of 10 yrs. As the lasers have been experimentally tested in earlier experimental verification for radiation hardness and were found to be resistant to radiation at such low dosages they were not tested in this program. A complete space radiation qualification of the Laser transmitter is planned for the future [5]. The results of the radiation test are presented in table 5 where samples of Yb doped fiber and WDM couplers containing GRIN lenses were irradiated to Gamma radiation from a ^{60} Co source. The induced loss due to the gamma irradiation are mentioned in table 5.

5 MECHANICAL DESIGN

The SLYB laser transmitter is housed in a compact aluminum module. The commercial package for the optical isolators meant for terrestrial applications was retained in our first prototype. In future versions the housing for the isolators will be made much lighter. Thus most of the weight of the complete transmitter is due to the optical isolators. Figure 3 presents the view of SLYB transmitter and the mechanical layout of the various components. The Yb-doped fiber is secured at the bottom of the housing and the bend radius is maintained at 50 mm with only one loop at 38 mm. The electrical connections to the transmitter are made through 2 DB15 and 1 SMA connectors for the RF modulation of the 985 nm signal diode. A finite element analysis of the transmitter module exhibited a resonance frequency of 2.5 KHz and good heat dissipation from the pump diode (0.1 degree increase of the pump diode compared to the whole module). The overall dimensions of the module are 15X12X4.3 cm^3 (LXWXH) and the total weight is 0.9 Kg.

Table 5: Induced attenuation due to Gamma irradiation at end of life (10 Years)

Component	Induce loss (dB)
Yb-doped fiber	1.5
WDM coupler	0.24
985 nm PM fiber isolator	0.12

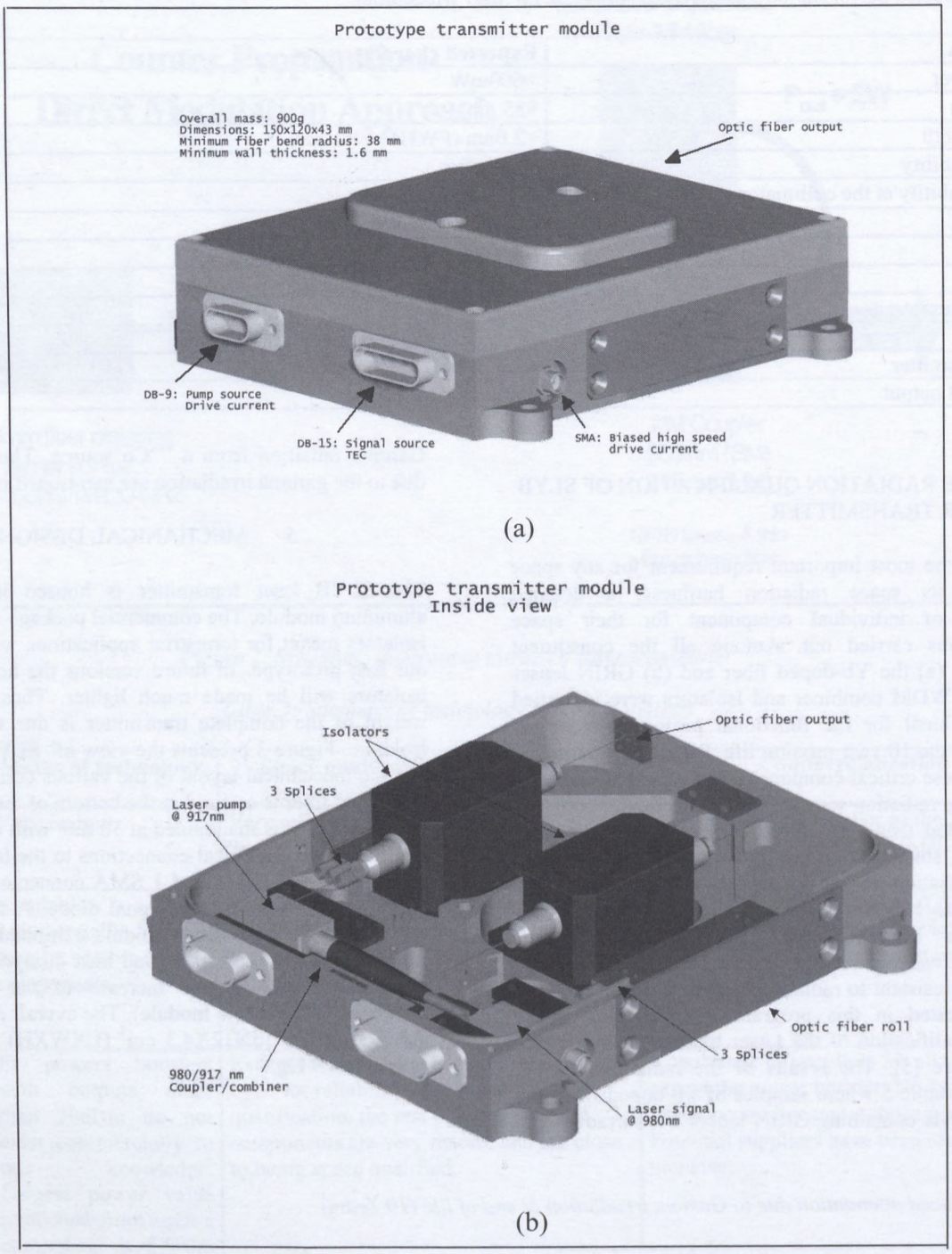


Figure 3: (a) External and (b) Inside view of the SLYB laser transmitter module

6 FUTURE POTENTIALS AND PLANS FOR HIGH POWER LASER TRANSMITTER

The commercial satellite communication market shows a trend towards Optical links for interconnection between various layers of satellite planes in LEO, MEO and GEO. With the existing satellite constellation (e.g., Teledesic) opting for Optical inter-satellite links for their high data transmission capacity, light weight and low power consumption future satellite programs will depend on OISL's for all their communication requirements. MPBT with its other partners have plans to space qualify the existing system to a certain degree and commercialize it for various satellite programs. The future plans of the existing High power laser transmitter activity is to build and space qualify a 3-4 watts laser transmitter for LEO to GEO link.

7 CONCLUSION

One of the major challenges at this point of technology development in lasers is the reliability and space radiation hardness of high power lasers components for long term (10 years) satellite programs. MPBT with its other partners are building a laser transmitter with output power of 600 mW for LEO-LEO Inter-satellite link. The first space qualified demonstrator module is expected to be ready towards the end part of 1999.

8 REFERENCES

[1] S.G. Lambert and W.L. Casey, Laser Communications in Space, Boston, Artech House, 1995.

[2] MPB Technologies Inc., Fibre-coupled High power Laser Transmitter for Optical Communication Terminal: Task 1 Technical report, Report No. MPBT-529S002, Submitted to ESTEC/ESA, April 9th 1998.

[3] MPB Technologies Inc., Fibre-coupled High power Laser Transmitter for Optical Communication Terminal: Preliminary design report, Report No. MPBT-529S003, Submitted to ESTEC/ESA, June 25th 1998.

[4] Y. Tashiro, H. Tachibana, A. Fujisai, H.Ogoshi, High power erbium-doped fiber amplifier pumped by wavelength multiplexed semiconductor laser diode unit, OFC'97 Technical digest, 1997, pp. 107-108.

[5] MPB Technologies Inc., Proposal for Space radiation qualification of SLYB Laser transmitter, MPBT Proposal no. DP98-026S, Submitted to ESTEC/ESA, 21st January 1999.

9 ACKNOWLEDGEMENTS

The authors like to thank the European Space Research and Technology Center (ESTEC) of the European Space Agency for providing the funding for this program under the ESA contract 12440/97/NL/SB.

Solid-state power amplifiers for the Japanese Engineering Test Satellite-Ⅷ

Hitoshi Ishida (ASC), Yoichi Kawakami (ASC), Haruzou Hirose (MITSUBISHI) and Masahumi Shigaki (FUJITSU)
Advanced Space Communications Research Laboratory (ASC)
HAYAKAWA TONAKAI BLDG, 2-12-5, IWAMOTO-CHO,
CHIYODA-KU, TOKYO 101-0032 JAPAN
E-mail: ishida@asc.co.jp

ABSTRACT

The Advanced Space Communications Research Laboratory (ASC) has been developing a space-qualified solid-state power amplifier (SSPA) and its electronic power conditioner (EPC) for 10W and 20W SSPAs. The performance of these SSPAs for multi-beam mobile satellite communications and satellite broadcasting sound systems is to be verified by the Japanese Engineering Test Satellite Ⅷ (ETS-Ⅷ). Not only the microwave performance specifications but also producibility and reliability are taken into account in the amplifier's design.
The 10W SSPA and 20W SSPA are designed to operate in the 2.5 to 2.54 GHz range. They must be also to withstand the rigors of launch and have a working life time of 3 years in space. They display an operating gain of 58 dB at 10W and 20W output power with an efficiency of 22.7% and 21.7% including EPC for multi carrier operation with more than 16 dB NPR. These SSPAs are fully gain-stabilized in the temperature range 0℃ to 50℃.

INTRODUCTION

The ETS- Ⅷ satellite is three-axis-stabilized geostationary satellite with a rectangular body, a 2500kg beginning-of-life mass and 5000W end-of-life power. In ETS-Ⅷ, single-hop mobile-to-mobile communications experiments at S-band with multiple beams frequencies using active phased array antennas and S-band digital broadcasting experiments providing compact-disk quality music for mobile terminals will be demonstrated.

In mobile satellite communications, it is necessary to use multiple beam satellite antenna in order to communicate with small mobile terminals. Although a high gain narrow beam satellite antenna enables use of small eirp mobile terminals, sub-dividing of coverage area to large number of small cells occurs. To alleviate these problems, we proposed multi port amplifier.

To accommodate the growing demand for such system, it is becoming increasingly important to transmit more than one carrier simultaneously through a common power amplifier. In these systems, the amplifier is required to suppress the intermodulation(IM) between carriers for obtaining low crosstalk between channels. Recently, to characterize the IM properties for multiple carriers, the noise power ratio (NPR) has been used. This is defined as the ratio of average power of all carriers to the average power falling within a certain notch bandwidth. In this measurement multiple carriers are dealt with as white noise with a narrow notch and intermodulated signal power is measured as the power falling within this frequency notch.
The 10W SSPA and 20W SSPA comprise 23 units and 8units, respectively. They are designed to operate in the 2.5 to 2.54 GHz range. They must also be able to withstand the rigors of launch and have a working life time of 3 years in space. They display an operating gain of 58 dB at 10W and 20W output power with an efficiency of 22.7% and 21.7% including EPC at multi carrier operation with more than 16 dB NPR. These SSPAs are fully gain-stabilized in the temperature range 0℃ to 50℃.

SSPA REQUIREMENTS

The requirements for SSPAs suited for ETS- Ⅷ applications are thus high NPR, high efficiency and very low mass. Gain and phase tracking over flight sets of large number of units is also need, therefore requiring use of technologies with repeatable performance, such as MMIC(monolithic microwave integrated circuit).
Table Ⅰ summarizes the SSPA performance specifications suitable for the ETS-Ⅷ applications.

Table I SSPA Performance Specifications

Requirements Parameter	Value
Frequency Range	2.5 to 2.54GHz
Return Loss (input & output)	20dB min
NPR	16dB min
Gain	55dB min (10W SSPA)
Gain	58dB min (20W SSPA)
Gain Flatness	0.5dB max
Gain Tracking	0.2dB p-p max
Phase Tracking	20 deg. p-p max
DC Input Voltage	100 V
DC Power Consumption	44W max (10W SSPA)
	92W max (20W SSPA)
Weight	660g. max (10W SSPA)
	2415g. max (20W SSPA)
Size	260x60x110 mm max (10W SSPA)
	220x210x64 mm max (20W SSPA)

The SSPA block diagram is illustrated in Figure 1.

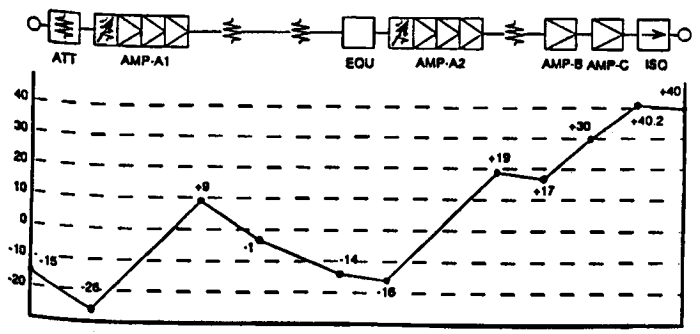


Figure 1. The 10W SSPA block and level diagram

SSPA DESIGN APPROACH

The variable gain amplifier (AMP-A1 and AMP-A2) has been designed using MMIC technology. In the gain amplifier pHEMTs are used, which provide very low noise figures and very high gain at S-band. With such devices, a total of 3 stages are needed to achieved the required gain. The packaging of these amplifiers is performed individually to mitigate risk and for reasons of modularity.

The driving(AMP-B) and final stage(AMP-C) amplifiers have been designed using Internally Matched Power GaAs FET modules. Temperature compensation of the intermodulation, output power and DC power consumption are achieved by adjusting the drain voltage of these modules to the temperature.

The DC/DC converter is located in a separate component of the SSPA. The housing is machined from solid aluminum, with two back-to-back compartments for the DC and RF circuits. RF and DC interfaces are through SMA and connectors respectively.

SSPA PERFORMANCE

The measured SSPA RF performance is illustrated in Figure 2 to 4, for RF gain and input and output characteristics. A temperature range of 0℃ to 50℃ is used.
Gain, gain flatness, gain tracking, and NPR specifications are all satisfied.

Comparison of gain and input-output measurement between the 10W SSPA and 20W SSPA provides the confidence that the tracking specifications can be satisfied without any form of active compensation, for large number of SSPAs.

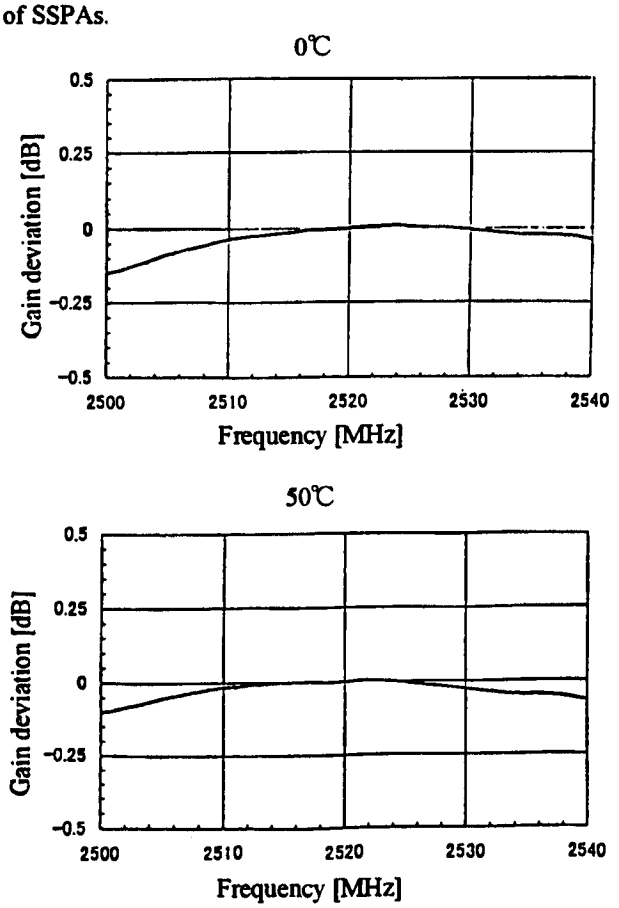


Figure 2. S-band Gain of 10W SSPA

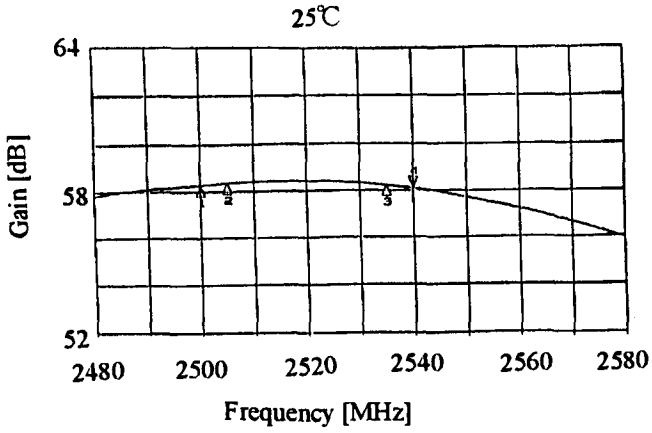


Figure 3. S-band Gain of 20W SSPA

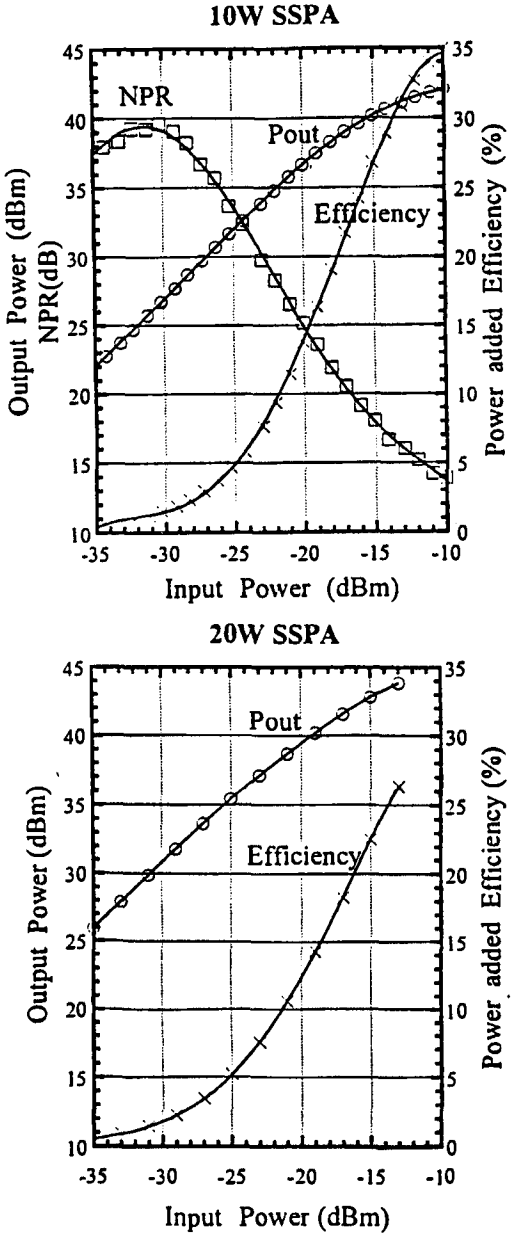
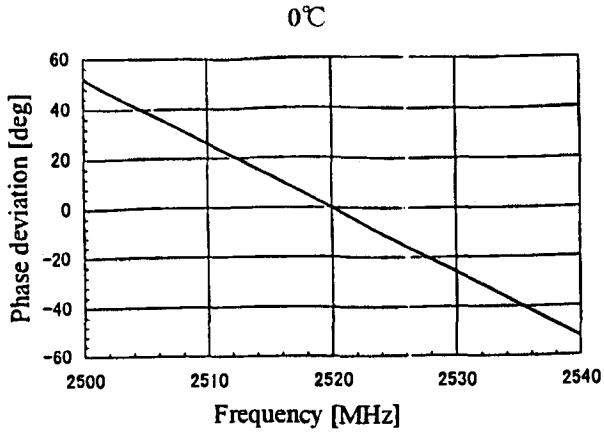


Figure 4. Input and Output characteristics of SSPA



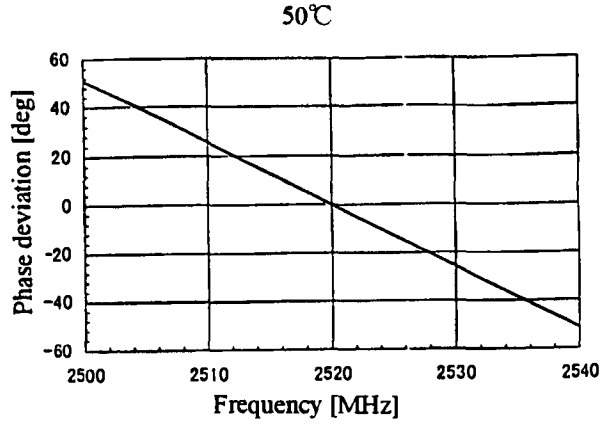


Figure 5. S-band Phase of 10W SSPA

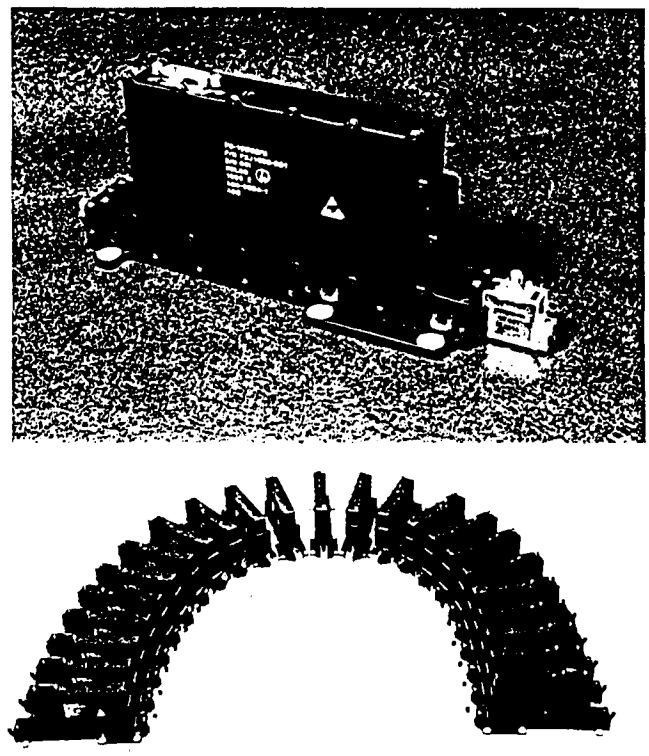


Figure 6. Photograph of 10W SSPAs

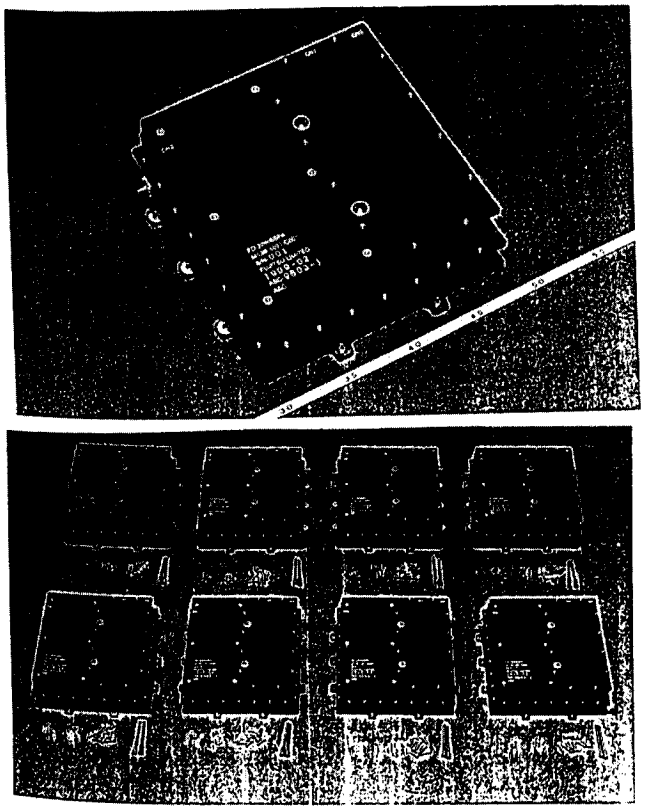


Figure 7. Photograph of 20W SSPAs

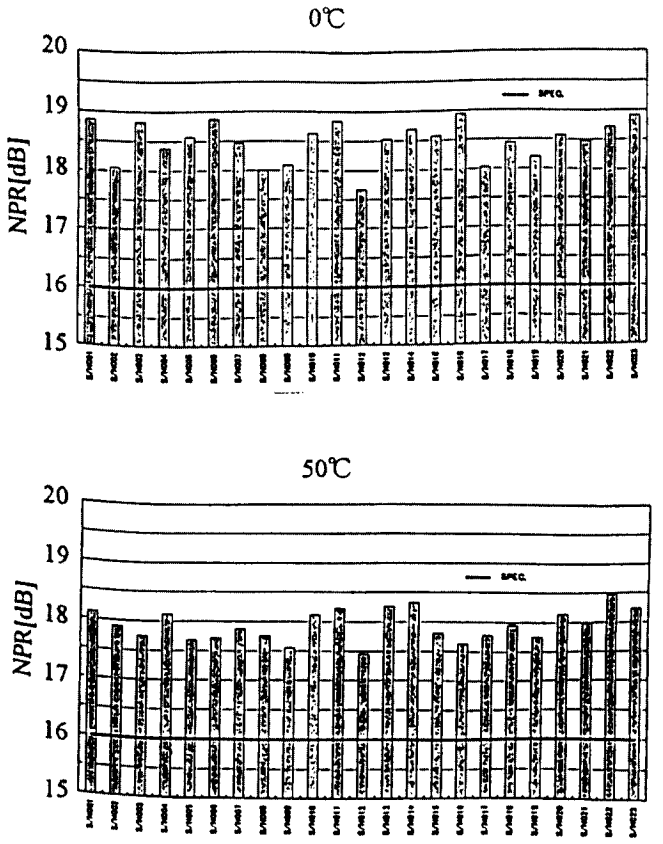


Figure 8. NPR histogram (10W SSPA)

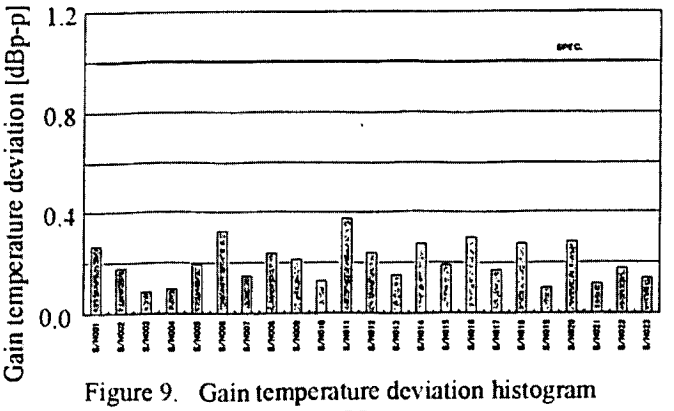


Figure 9. Gain temperature deviation histogram (10W SSPA)

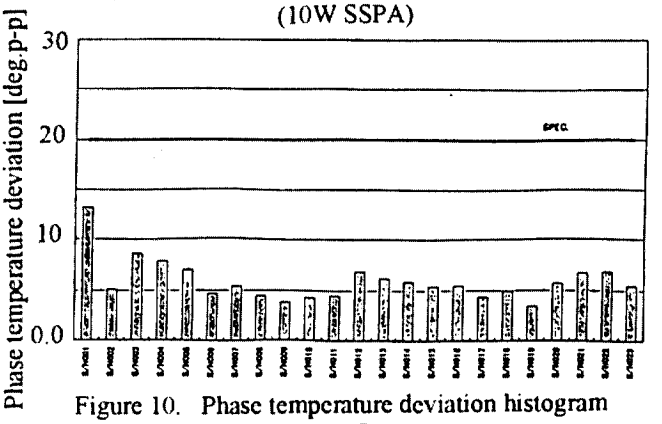


Figure 10. Phase temperature deviation histogram (10W SSPA)

These results show that all SSPA key performance specifications are satisfied by ETS-Ⅷ.

CONCLUSIONS

The 10W and 20W SSPA have been developed by ASC for S-band multi-beam phased array satellite systems, ETS-Ⅷ. The 10W SSPA and 20W SSPA comprise 23 units and 8units, respectively. They are designed to operate in the 2.5 to 2.54 GHz range. They must also be able to withstand the rigors of launch and have a working life time of 3 years in space. They display an operating gain of 58 dB at 10W and 20W output power with an efficiency of 22.7% and 21.7% including EPC at multi carrier operation with more than 16 dB NPR. These SSPAs are fully gain and phase-stabilized in the temperature range 0℃ to 50℃.

This combination of high NPR, high efficiency , very low mass, low gain and low phase deviations over SSPAs can provide the best system performance.

ACKNOWLEDGEMENTS

The authors wish to thank the many ETS-Ⅷ team members at MITSUBISHI and FUJITSU for their hard work in designing the ETS-Ⅷ system and their generous assistance in writing, reviewing, and publishing this paper. Special thanks goes to Shigetoshi Yoshimoto of ASC general manager for his support in producing this paper.

The SWAID Project: Deriving Powerful Modulation and Coding Schemes for Future Satellite Multimedia Systems

C. Valadon, Y. Rosmansyah, R. Tafazolli, B.G. Evans
Centre for Communication Systems Research, University of Surrey
Guildford, Surrey, GH2 5XH, England,
Tel: +44-1483-259808, Fax: +44-1483-259504
E-mail: C.Valadon@ee.surrey.ac.uk

ABSTRACT

In this paper, different modulation and coding schemes suitable for future satellite multimedia services in Ka band are proposed and their performance are assessed. The considered techniques include convolutional codes, trellis coded modulations as well as concatenated schemes and turbo codes.

I. INTRODUCTION

The anticipated demand for the provision of multimedia services such as videoconferencing, telemedicine, audio and video broadcasting to compact user terminals has prompted the emergence of a number of proposals for Ka band satellite systems such as WEST from Matra Marconi Space (MMS), EuroSkyWay by Alenia Aerospazio, ASTROLINK by Lockheed Martin, CYBERSTAR from Space Systems/Loral and SPACEWAY from Hughes Communications. On that note and following the Space Foresight exercise, the British National Space Center (BNSC) initiated the SWAID (Satellite Wideband Air Interface Design) project as a collaboration between MMS and the Centre for Communication Systems Research (CCSR) at the University of Surrey.

When compared to first-generation mobile satellite communication systems such as GLOBALSTAR and IRIDIUM, future multimedia satellite networks will have to provide significantly higher data rates while achieving extremely low BER values. In order to keep the user terminal complexity as low as possible, it is essential to reach the required quality of service (QoS) with very low values of the signal-to-noise ratio (SNR). Hence, the choice of the optimum channel coding technique with the best possible access scheme will be of utmost importance in realizing such performance limits.

In this paper, different modulation and Forward Error Correction (FEC) coding schemes suitable for these future multimedia satellite systems are proposed and their performance assessed. The considered techniques include conventional Convolutional Codes (CC), Trellis Coded Modulations (TCM) as well as concatenated schemes and Turbo Codes (TC). Issues such as interleaving and decoding complexity are addressed in order to provide accurate comparisons between the different schemes. Practical limitations such as the buffering and the processing required on board the satellite are most important and need to be considered before adopting a scheme for a real system.

II. SYSTEM DESCRIPTION

SWAID is aimed at designing and optimising the air interface of a future Ka band system delivering services with data rates ranging from 32 kbps up to 384 kbps to small easily installed inexpensive User Terminals (UT) with small dish type antennas. It is assumed that the UT antenna diameter is equal to 0.7 m and the RF output power is provided with a single SSPA with a power lower than 1 W.

The system comprises advanced geostationary satellite payloads providing an extended European coverage with a 53 spotbeam arrangement. Full onboard regeneration, including channel decoding in the uplink and coding in the downlink, is considered in order to ease the UT requirements and increase the system flexibility. The satellite *G/T* has been taken equal to 18 dB/K. Moreover, it was put as a system constraint at the beginning of the project that the payload power should not exceed 10 kW.

The satellite air interface packets are ATM encapsulated with a modified header to support the satellite specific protocols and Fast Packet Switching (FPS) is provided. By so doing, efficient use of the satellite resources as well as bandwidth can be achieved. It is assumed that the size of the ATM-like cell is equal to 55 bytes.

Rain attenuation can be particularly penalising at Ka band. It is therefore important to provide adequate link margins based on the required availability. The SWAID system is designed to provide a link availability higher than 99% of the year for a target end-to-end BER of 10^{-10}.

A nonuniform traffic distribution over the coverage area has been assumed. The traffic in the user link is equal to 1.8 Gbps in the uplink and 4.2 Gbps in the downlink.

In the initial phase of the project, a comparison between Synchronous CDMA (S-CDMA) [1], [2] and TDMA has been performed in order to choose the most suitable access scheme. The results of this comparison can be found in [3] and are summarised in Table 1.

Parameter	S-CDMA	TDMA
UT power (W)	0.34	0.29
Uplink system bandwidth (MHz)	136	195
Bandwidth to be processed in the uplink (MHz)	3060	1685
Satellite RF power (W)	1721	2210
Maximum RF power in one downlink spotbeam (W)	118	122.8
Downlink system bandwidth (MHz)	408	669
Bandwidth to be processed in the downlink (MHz)	10700	6891

Table 1: Access scheme comparison

It can be seen from Table 1 that both designs lead to similar values of the required UT power. S-CDMA presents a slight advantage in terms of required satellite power as well as system bandwidth. However the bandwidth to be processed on-board the satellite is higher for the CDMA case by 2.6 dB in the uplink and 1.9 dB in the downlink. Taking these results into account together with the fact that regenerative CDMA payloads are a less developed technology, TDMA has been chosen as the baseline access scheme.

III. CONVOLUTIONAL CODES

The first schemes to be considered for SWAID are based on CC. Over the past years, CC have been used for a large number of communication systems. Maximum Likelihood Sequence Estimation (MSLE) can easily be implemented with the *Viterbi* Algorithm (VA). Figure 1 presents the *Eb/No* required to achieve the target BER for half rate CC with different constraint length values. The plain line corresponds to simulation results whereas the results associated with the dotted line have been obtained with the following upper bound:

$$P_e^b \le Q\left(\sqrt{2 \cdot d_{free} \cdot \frac{E_c}{N_o}}\right) \times e^{d_{free} \cdot E_c/N_o} \times \left.\frac{dT(D,N)}{dN}\right|_{\substack{N=1 \\ D=e^{-E_c/N_o}}} \quad (1)$$

$T(D,N)$ is the transfer function of the CC and d_{free} denotes the *Hamming* free distance. *Ec/No* is the ratio between the energy per coded symbol and the noise spectral density.

It can be seen from Figure 1 that increasing the constraint length of the CC eases the power requirements. However, the number of states in the trellis associated with the CC grows exponentially with the constraint length. Since the decoding complexity directly depends on this parameter, the constraint length cannot be chosen arbitrarily high. The CC with constraint length seven offers a good trade-off between performance and complexity since the target BER can be achieved with an *Eb/No* of 6.8 dB and 64 states in the trellis. In order to further reduce the power requirements by 1 dB, one would need to implement the CC with constraint length 11 at the cost of increasing the decoding complexity by a factor equal to 16.

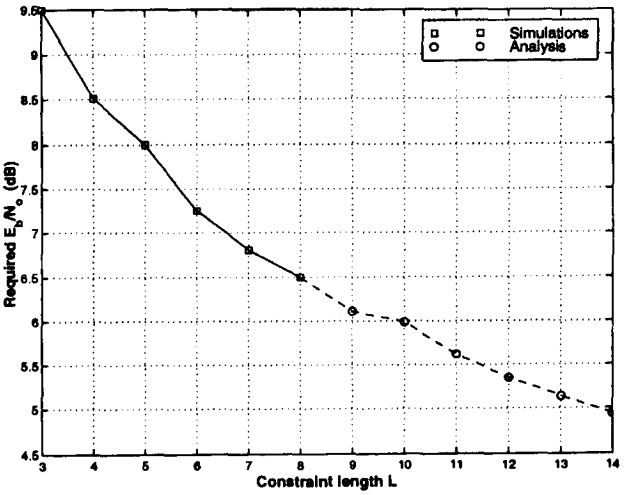


Figure 1: *Eb/No* requirements *vs.* constraint length

The decoding complexity of the uplink scheme will have a major impact on the satellite on-board processing section. Hence, it is of interest to investigate the possibility of reducing this computational burden. A number of sub-optimum algorithms such as the M algorithm [4] for the decoding of trellis codes have been proposed in the past. The M algorithm restricts the trellis search to only a subset of all possible survivors. Figure 2 presents the performance of the M algorithm for the CC(½,7) and different sizes of the survivors' set. The *Eb/No* is equal to 4.5 dB.

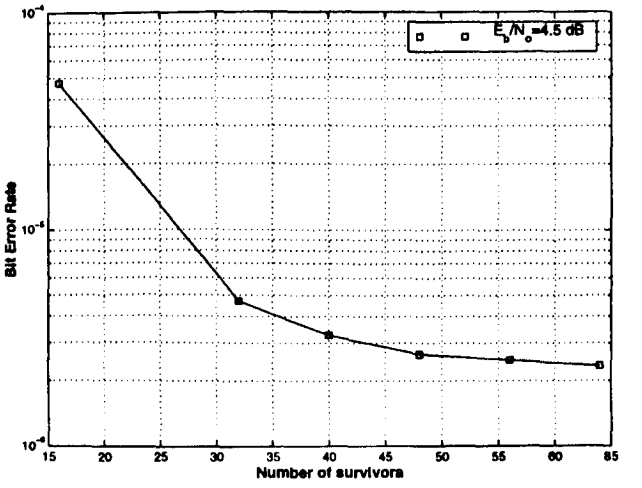


Figure 2: BER *vs.* number of survivors

As expected, the performance of the M algorithm is equal to that of the VA when the number of survivors is equal to the number of states in the trellis. When only 16 states are used

for the decoding procedure, the performance noticeably degrades and the *Eb/No* loss is equal to 0.6 dB. However, if the comparison is made for the same number of survivors, the CC(½,7) with the M algorithm and 16 survivors outperforms the CC(½,5) decoded by the VA by more than 0.5 dB. When 32 survivors are kept in the trellis, the M algorithm leads to the same power requirements than the VA. Hence, using the M algorithm, the decoding complexity can be reduced without any performance degradation.

When very high power efficiency is to be achieved, the inner CC can be concatenated with an outer *Reed-Solomon (RS)* code. The simulation of concatenated CC+*RS* codes is far to complex when very low BER are considered. The approach that has been chosen is to simulate the inner code and then analytically derive the performance of the end-to-end transmission. Using the assumptions presented in [5], the bit error probability at the output of the *RS(n,k,b)* can be derived:

$$P_e^b \le \sum_{i=t+1}^{n} \frac{i+t}{2 \cdot n} \times P(\text{i RS symbols in error in the codeword}) \quad (2)$$

n and *k* denote the length and the dimension of the *RS* code. The symbol error correction capability is denoted as *t*. The *RS* symbols take their values in the *Galois* field $GF(2^b)$. If the errors produced by the VA are independant, the number of erroneous symbols in the *RS* codewords follows a binomial distribution. The BER at the output of *RS* decoder can then easily be derived with the sole knowledge of the Symbol Error Rate (SER) at the output of the VA. However, due to the structure of the CC, bursts of error tend to happen. In order to take into account this effect, it is proposed to model the errors produced by the VA with a two-state *Markov* chain. The bad state corresponds to the bytes in error and the good state represents the bytes that have been correctly decoded. With the *Markov* model, the distribution of the number of erroneous bytes in the *RS* codewords can be calculated using the method proposed in [6].

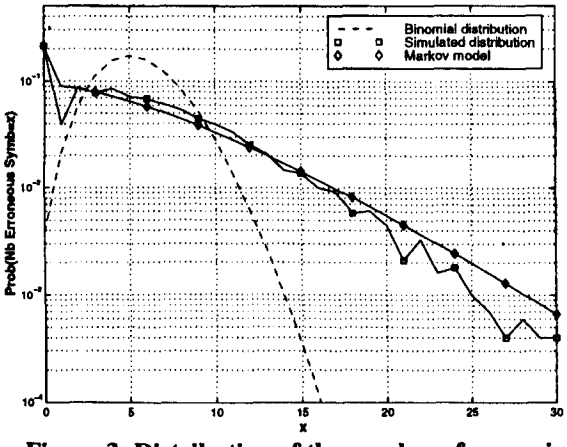


Figure 3: Distribution of the number of errors in the codeword using the *Markov* model

Figure 3 presents the distribution of the number of errors in *RS(236,220,8)* codewords for the inner CC(¾,7) and an *Eb/No* of 3 dB. The distribution obtained from the simulation results is very different from the binomial distribution. As explained earlier, this is due to the correlation in the errors at the output of the VA. On the other hand, the distributions obtained via simulation and using the *Markov* model are extremely close. Hence, the *Markov* model will be used to assess the performance of concatenated coding schemes. It is also interesting to note that using the *Markov* model, the effect of interleaving can easily be taken into account.

IV. CONCATENATED TCM-*RS*

In order to improve the spectral efficiency of the FEC coding schemes, it is possible to use TCM. TCM were first proposed by *Ungerboeck* [7] and allow the achievement of good coding gains without any reduction in the spectral efficiency by introducing memory in the modulation operation while expanding the signal set. The improvement of TCM when compared to CC in spectral efficiency leads to an increase for the power requirements. In order to lower this increase, the concatenation of the inner TCM with an outer *RS* is considered.

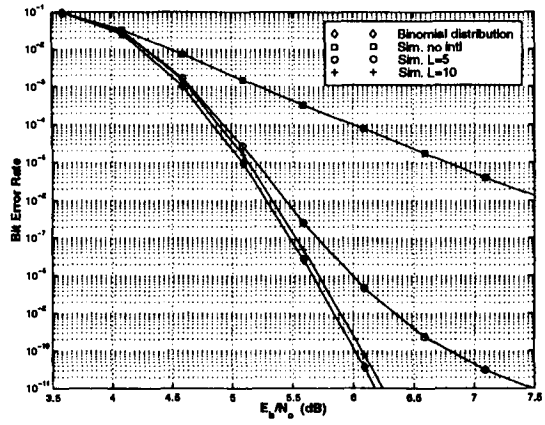


Figure 4: BER of the $^2/_3$-8PSK TCM + *RS(63,55,8)*

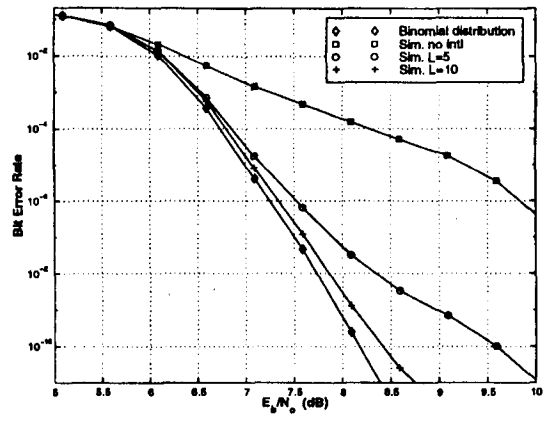


Figure 5: BER of the ¾-16QAM TCM + *RS(63,55,8)*

Figure 4 presents the performance of the *Ungerboeck* $^2/_3$-8PSK TCM with 16 states and outer *RS(63,55,8)*. In Figure 5, an inner ¾-16QAM TCM is considered. The *Markov* model has been used in order to assess the performance for different interleaving length values.It can be seen from Figure 4 and Figure 5 that an interleaving depth of 10 is required in order to achieve a performance close to the optimum one. The performance of TCM having larger number of states has also been assessed. However, it was observed that by increasing the TCM memory, the correlation in the errors produced by the VA became larger. This affects negatively the performance of the outer *RS* decoder. For example, when the inner $^2/_3$-8PSK TCM is used, an increase in the number of states in the trellis from 16 to 64 with an increase in the interleaving depth from 10 to 15 lowers the power requirements by only 0.2 dB. Hence, the use of TCM with 16 states provides a good trade-off between performance on the one hand and decoding complexity and delay on the other hand. *RS* codes with larger error correction capabilities have also been tested. However, it was found that the increase in decoding complexity provided only very limited gains (<0.2 dB).

V. TURBO CODES

TC are a new class of coding schemes that have first been introduced by *Berrou et al.* [8] in 1993. TC have received a lot of attention because they exhibit extremely good performance. For a BER of 10^{-5}, the first TC was only 0.7 dB away from *Shannon's* limit. TC represent very promising schemes, however a number of issues must be tackled before successful implementation for commercial communication systems.

First, the TC proposed by *Berrou* uses very large interleaving blocks (65536 bits). This would be very problematic for satellite applications since this would entail large delays and buffer sizes. In the SWAID project, small interleaving lengths have been considered. Our results have shown that with the appropriate interleaving algorithm, good performance can be achieved even with an interleaving length equal to the ATM-like cell [9]. Figure 6 present the *Eb/No* required to achieve a BER of 10^{-3} for different values of the code memory and interleaving length. The constituent codes are of rate ½ and puncturing is applied in order to achieve an overall code rate of ½. Four decoding iterations have been performed. The performance degradation due to a reduction in the interleaving length from 7040 bits to 1760 bits is lower than 0.2 dB. When the length is further reduced to 440 bits (one ATM-like cell), the additional loss is limited to 0.5 dB. Hence, TC can be successfully applied even when the interleaving length has to be kept very low.

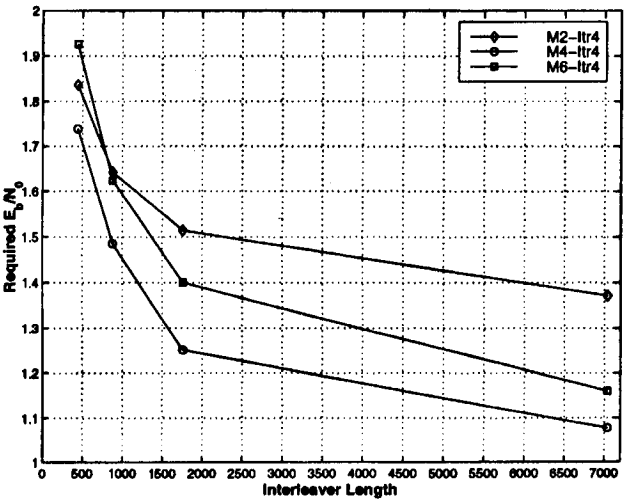


Figure 6: *Eb/No* required to achieve a BER of 10^{-3}

Another issue that needs to be addressed for the implementation of TC is the decoding complexity. The decoding complexity depends on both the Soft Input Soft Output (SISO) algorithm used for the decoding of the constituent codes and the number of required iterations. The modified *Maximum A Posteriori* (MAP) algorithm offers the best performance but is extremely complex. A simplified version, which does not show any performance degradation, has been proposed in [10]. This algorithm can be further simplified by performing the computations in the log domain and is referred to as LogSMAP. The performance of the SubSMAP has also been assessed. The SubSMAP algorithm is based on the LogSMAP algorithm and uses an approximation for the calculation of the logarithm of a sum of exponential terms.

Figure 7 presents the performance of the TC for the different SISO decoding algorithms. The memory of the constituent codes is equal to 2. Better power efficiency could be achieved with constituent codes having a larger memory size. However, it was shown that increasing the memory size beyond 4 did not provide any performance gain. Moreover, the improvement in power efficiency achieved by doubling the memory size from 2 to 4 is very limited in view of the increased decoding complexity. The interleaving length is equal to 440 bits. The dotted line corresponds to the performance of the LogSMAP algorithm for 8 decoding iterations. The plain lines are associated with the SubSMAP algorithm for which different number of decoding iterations have been considered. When 8 decoding iterations are considered for both algorithms, the performance degradation of the SubSMAP is limited to 0.15 dB. Moreover, halving the number of SubSMAP decoding iterations increases the power requirements by only 0.1 dB.

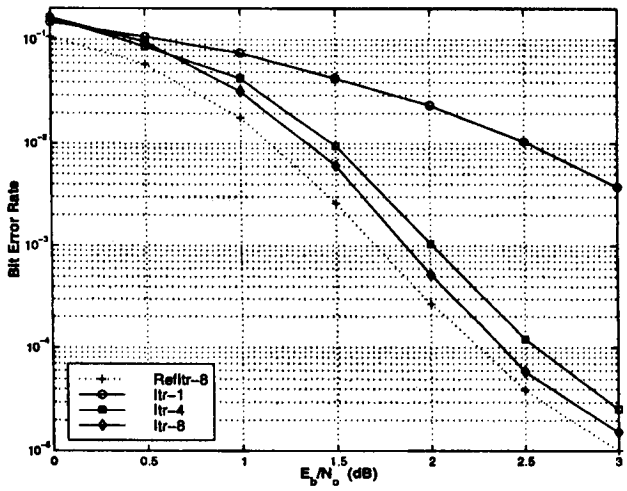


Figure 7: TC performance

TC are very efficient when moderate to low BER values are to be achieved. However, when very low BER are targeted, the flattening in the BER curve, due to the weakness of the constituent codes, means that they do not outperform conventional CC. Taking into account the requirements set, it was decided to concatenate the TC with an outer *RS*. Figure 8 presents the performance of the concatenated TC+*RS(63,55,8)*. Depths of 3, 5 and 10 have been considered for the interleaving between the TC and the *RS* decoders.

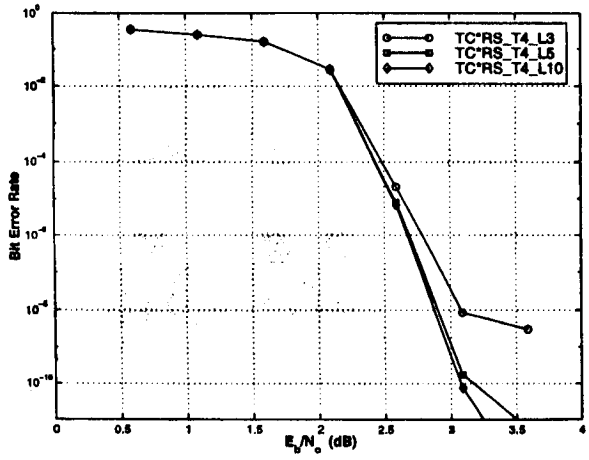


Figure 8: Performance of concatenated TC+*RS*

An interleaving depth larger than 3 is required in order to avoid the BER flattening. On the other hand, the gain provided by increasing the interleaving length from 5 to 10 is lower than 0.1 dB. Hence, when compared to CC, TC require smaller interleaving sizes at the decoder output. This can be justified by the fact that the memory of the constituent codes is very low, and hence the correlation in the errors produced by the decoder is smaller.

VI. SUMMARY AND CONCLUSIONS

Using the results presented in the above sections of this paper, seven candidates have been retained. Table 2 presents the characteristics of these different schemes. The schemes are compared according to their power efficiency, spectral efficiency and decoding complexity. The actual decoding complexity of the different schemes will depend on the implementation and is therefore impossible to exactly quantify. However, in order to have an estimate of the required processing, the decoding complexity is expressed in complexity units per second (cups). Assumptions on the number of complexity units (cu) associated with the different operators have been made:

- Real addition or subtraction: 1 cu
- Real multiplication: 2 cu
- Addition in *GF(256)*: 1 cu
- Multiplication in *GF(256)*: 8 cu
- Division in *GF(256)*: 16 cu

Coding scheme	*Eb/No*	Spec. Eff. (b/symb)	Comp. (Mcups)
CC(½,7)	6.8 dB	0.99	68
2/3-8PSK TCM (16 states) + *RS* (71,55,8)	5.7 dB	1.51	95
2/3-8PSK TCM (16 states) + *RS* (63,55,8)	6.1 dB	1.71	75
¾-16QAM TCM (16 states) + *RS* (63,55,8)	8.4 dB	2.52	135
CC(rate ¾, L=7) + *RS* (237,220,8)	4.4 dB	1.39	80
CC(rate ¾, L=5) + *RS* (237,220,8)	4.8 dB	1.39	35
TC(rate ½, M=2, 4 it.) + *RS* (63,55,8)	3.2 dB	0.87	340

Table 2: Comparison between the different schemes

In order to allow easy visualisation, Figure 9 and Figure 10 plot the results summarised in Table 2.

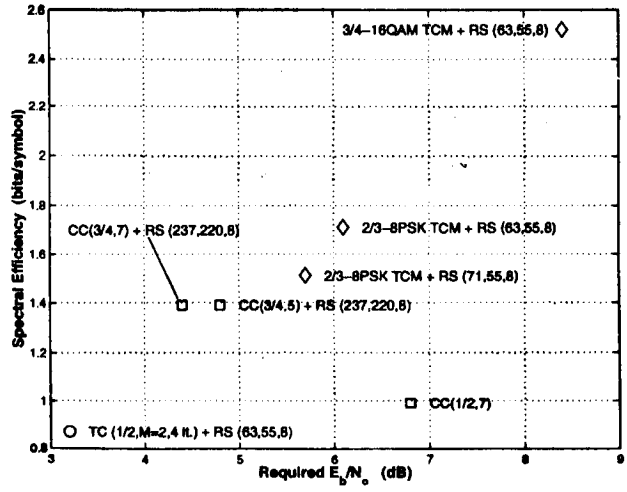


Figure 9: Spectral efficiency of the different schemes

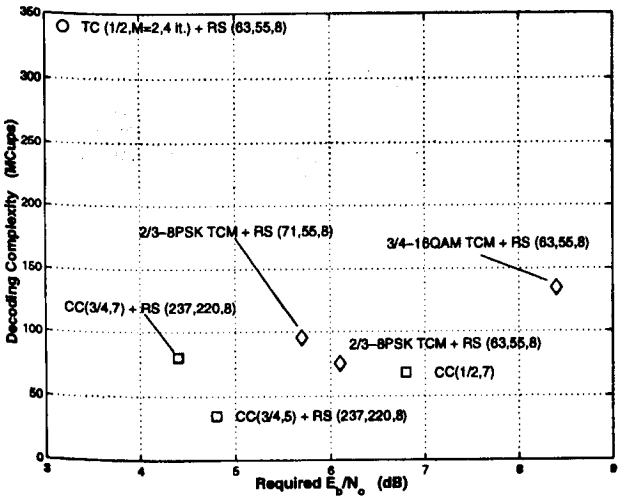


Figure 10: Complexity of the different schemes

The TC provides the best power efficiency. It outperforms all the other schemes by more than 1 dB. However, the spectral efficiency is quite limited and the decoding complexity extremely high. The complexity of decoding the proposed TC with the simplified SubSMAP algorithm is approximately four times that of the CC(½,7).

The concatenation of an inner $^2/_3$-8PSK TCM with 16 states and outer *RS* proves to be more efficient both in terms of signal bandwidth and power requirements than the CC(½,7). With the outer *RS(63,55,8)*, the TCM offers a 70% increase in spectral efficiency and a 0.7 dB gain. Moreover, the decoding complexity of both schemes is very similar. It should however be stressed that the TCM schemes require interleaving at the output of the VA with a length equal to 10.

When very high spectral efficiency values are targeted, the concatenation of the inner ¾-16QAM TCM with the outer *RS(63,55,8)* seems to be the most appropriate. When compared to the concatenated schemes with inner $^2/_3$-8PSK TCM, the gain in spectral efficiency is higher than 50%. However, the power requirements have to be increased by more than 2.3 dB and the decoding complexity is more than 40% higher.

The concatenation of the inner punctured CC of rate ¾ with an outer *RS* code provides a good trade-off between spectral and power efficiency. Spectral efficiency values of 1.4 bit/symbol can be reached for required *Eb/No* lower than 5 dB. Moreover, the number of states in the trellis of the CC can be kept low without any important performance degradation. By so doing, very low decoding complexity can be achieved.

ACKNOWLEDGMENTS

The authors would like to thank the BNSC and MMS for supporting the SWAID project.

REFERENCES

[1] R. De Gaudenzi, C. Elia and R. Viola, "Bandlimited Quasisynchronous CDMA: A Novel Satellite Access Technique for Mobile and Personal Communication Systems", *IEEE J. Select. Areas Commun.*, vol. 10, pp. 328-343, Feb. 1992.

[2] C. Valadon, G. Verelst, P. Taaghol, R. Tafazolli, B.G. Evans, "Code-Division Multiple Access for Provision of Mobile Multimedia Services with a Geostationary Regenerative Payload", *IEEE J. Select. Areas Commun.*, vol. 17, no. 2, pp. 223-237, Feb. 1999.

[3] C. Valadon, I. Mertzanis, G. Huggins, R. Tafazolli, B.G. Evans, "Air Interface Design for the Provision of Multimedia Services to Compact User Terminals: The SWAID Project", *in Proc. 4th Ka Band Utilization Conf.* (Venice, Italy), Nov. 1998, pp. 317-324.

[4] J.B. Anderson and S. Mohan, "Sequential Coding Algorithms: A Survey and Cost Analysis", *IEEE Trans. Commun.*, vol. 32, pp. 169-176, Feb. 1984.

[5] S. Bellini, "On the Interleaver Depth in Concatenated Coding Schemes", *in Proc. of the European Workshop on Mobile/Personal Satcoms (EMPS'96)*, 1996.

[6] J.K. Wolf, "ECC Performance of Interleaved RS Codes with Burst Errors", *IEEE Trans. Magnetics*, vol. 34, no. 1, Jan. 1998.

[7] G. Ungerboeck, "Channel Coding with Multilevel/Phase Signals", *IEEE Trans. Info. Theory*, vol. 28, no. 2, pp. 5-66, Jan. 1982.

[8] C. Berrou, A. Glavieux, and P. Thitimasjshima, "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-Codes", *in Proc. IEEE Int. Conf. Commun.*, pp. 1064-1070, May 1993.

[9] C. Valadon, Y. Rosmansyah, R. Tafazolli, B.G. Evans, "Modulation and Coding Schemes for SWAID", *SWAID-Deliverable 4-Modulation and Coding for SWAID*, 17/11/98.

[10] S.S. Pietrobon and A.S. Barbalescu, "A Simplification of the Modified Bahl Decoding Algorithm for Systematic Convolutional Codes", *in Proc. Int. Symp. Info. Theory and Applications*, Nov. 1994.

Performance of Turbo-Codes with Relative Prime and Golden Interleaving Strategies

S. Crozier, J. Lodge, P. Guinand, and A. Hunt

Communications Research Centre, 3701 Carling Ave., P.O. Box 11490, Station H, Ottawa, Canada
K2H 8S2, ph: 613-998-9262, fax: 613-990-6339, email: stewart.crozier@crc.ca, web: www.crc.ca/fec

Abstract

This paper describes a number of new interleaving strategies based on the golden section. The new interleavers are called golden relative prime interleavers, golden interleavers, and dithered golden interleavers. The latter two approaches involve sorting a real-valued vector derived from the golden section. Random and so-called "spread" interleavers are also considered. Turbo-code performance results are presented and compared for the various interleaving strategies. Of the interleavers considered, the dithered golden interleaver typically provides the best performance, especially for low code rates and large block sizes. The golden relative prime interleaver is shown to work surprisingly well for high puncture rates. These interleavers have excellent spreading properties in general and are thus useful for many applications other than Turbo-codes.

1.0 Introduction

Interleaving is a key component of many digital communication systems involving forward error correction (FEC) coding. This is especially true for channels characterized by fading, multipath, and impulse noise, for example. Interleaving, or permuting, of the transmitted elements, provides time diversity for the FEC scheme being employed. An element is used here to refer to any symbol, sample, digit, or bit that is interleaved. In the past the interleaving strategy was usually only weakly linked to the FEC scheme being employed. Exceptions are concatenated FEC schemes such as concatenated Viterbi and Reed-Solomon decoding. The interleaver is placed between the two FEC encoders to help spread out error-bursts and the depth of interleaving is directly linked to the error correction capability of the inner (Viterbi) decoder. More recently, however, interleavers have become an integral part of the coding and decoding strategy itself. Such is the case for Turbo and Turbo-like codes, where the interleaver is a critical part of the coding scheme. The problem of finding optimal interleavers for such schemes is really a code design problem, and is an on-going area of research.

One problem with classical interleavers is that they are usually designed to provide a specific interleaving depth. This is fine if each burst of errors never exceeds the interleaver depth, but it is wasteful if the interleaver is over-designed (too long) and error-bursts are typically much shorter than the interleaver depth. For example, a simple 10×10 matrix interleaver has an interleaving depth of 10 elements. If a burst of 10 errors occurs, the deinterleaver will optimally spread these 10 errors throughout the block of 100 elements. If the error-burst is 11 elements long, however, then two errors will again be adjacent. If the error-burst is only two elements long then these two errors will only be spaced 10 elements apart after deinterleaving, but they could have been spaced much further apart if it was known that only two errors were present. For example, a 2×50 matrix interleaver would have spaced these two errors 50 elements apart. Of course this interleaver is not good for longer bursts of errors. In practice, most channels usually generate error events of random length, and the average length can be time varying, as well as unknown. This makes it very difficult to design optimum interleaving strategies using the classical approaches. What is sought is an interleaving strategy that is good for any error-burst length.

Section 2 provides some background on Turbo-codes and interleaving methods. Section 3 describes the new interleaving strategies based on the "golden section". Section 4 compares the bit and packet error-rate performance of Turbo-codes with the various interleavers. Section 5 gives the conclusions.

2.0 Background

Turbo-codes [1,2,3] have received considerable attention since their introduction in 1993. This is due to their powerful error correcting capability, reasonable complexity, and flexibility in terms of providing different block sizes and code rates. The canonical Turbo-code encoder consists of two 16-state, rate 1/2 recursive systematic convolutional (RSC) encoders operating in parallel with the information bits interleaved between the two encoders, as shown in Figure 1. Without puncturing, the overall code rate is 1/3. This is because the systematic information bits are only sent once.

Other code rates can be achieved as required by puncturing the parity bits c^1_k and c^2_k. It is the job of the interleaver to break apart low-distance error patterns that belong to one RSC code, in the hope that they will create high-distance error patterns in the other RSC code.

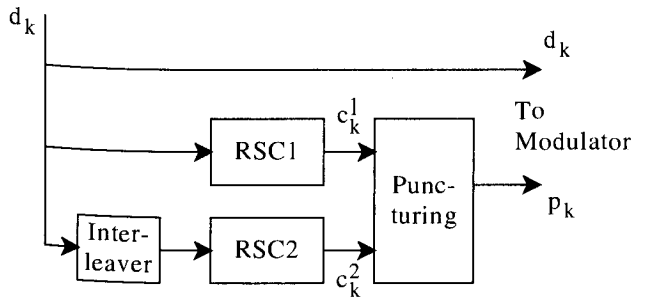


Fig. 1: Turbo-code encoder using two rate 1/2 RSC codes with puncturing.

A number of close-to-optimum and sub-optimum Turbo decoding methods are possible. The simulation results presented here are based on the enhanced maximum-log-*a-posteriori*-probability (max-log-APP) approach, with corrected extrinsic information, as described elsewhere in [4,5,6,7]. It has been found that performance is typically within 0.1 to 0.2 dB of exact, infinite precision log-APP decoding. The amount of degradation is a function of block size, code rate, and signal-to-noise ratio (SNR), with the larger degradations occurring for long blocks, low code rates, and low SNRs. It is convention that one Turbo decoding iteration be defined as two max-log-APP decoding operations.

Interleaving is a key component of any Turbo-code, as shown in Figure 1. Although some form of random or pseudo-random interleaving is usually recommended, it has been found that simple structured interleavers can also offer good performance, especially for short data blocks on the order of a few hundred bits. Examples of common structured block interleavers include relative prime interleavers and LxM matrix (or block) interleavers using L rows and M columns. An LxM matrix interleaver is usually implemented by writing into the rows and reading out of the columns, or vice versa. The rows or columns are sometimes read in or out in a permuted order. This permuted order is often implemented using a relative prime. That is, the row or column index can be generated using modulo arithmetic where the index increment and row or column lengths are relative primes. With L or M equal to 1, this type of interleaver simply becomes a one-dimensional relative prime interleaver. The relative prime interleaver is examined more closely in Section 3.

The original "Turbo" interleaver [1,2] is based on the use of an MxM matrix with a form of (pseudo-random) relative prime indexing, but the design is much more complicated than that described above. This interleaver has been reported to work well, but is not suited to arbitrary block sizes. This interleaver is not considered further in this paper, but the approach definitely merits further investigation.

Two other interleavers that have been investigated are the "random" interleaver, and the so-called "spread" interleaver [8,9,10]. The random interleaver simply performs a random or pseudo-random permutation of the elements without any restrictions. This interleaver is very useful as a benchmark, and has also been used extensively in calculating error-rate bounds [9,10].

The spread interleaver is really a semi-random interleaver. It is based on the random generation of N integers from 0 to N-1, but with the following constraint [8,10]:

> Each randomly selected integer is compared to the S most recently selected integers. If the current selection is within S of at least one of the previous S integers, then it is rejected and a new integer is selected until the previous condition is satisfied.

This process is repeated until all N integers are extracted. The search time increases with S, and there is no guarantee that the process will finish successfully. As a rule of thumb the choice $S < \sqrt{N/2}$ produces a solution in a reasonable amount of time. A number of variations on the spread interleaver are presented in [11,12,13,14]. These variations are not considered further here, but also merit further investigation.

3.0 Golden Section Interleaving

3.1The Golden Section

The golden section arises in many interesting mathematical problems. Figure 2 illustrates the golden section principle in relation to the interleaving problem of interest. Given a line segment of length 1, the problem is to divide it into a long segment of length g, and a shorter segment of length 1-g, such that the ratio of the longer segment to the entire segment is the same as the ratio of the shorter segment to the longer segment.

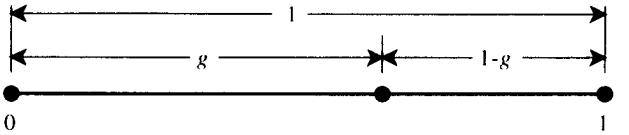


Fig. 2: Illustration of the golden section principle.

That is,

$$\frac{g}{1} = \frac{1-g}{g} \qquad (1)$$

Solving this simple quadratic equation for g gives the golden section value

$$g = \frac{\sqrt{5}-1}{2} \approx 0.618 \qquad (2)$$

Now consider points generated by starting at 0 and adding increments of g, using modulo-1 arithmetic. After the first increment there are two points at 0 and g that are $1-g$ apart, using modulo-1 arithmetic. Modulo distances are used to allow for the option of having the first point start anywhere along the line segment. From (1), the distance of $1-g$ is the same as g^2. After the second increment the first and third points determine the minimum distance and this distance is g^3. Again, this follows from the definition of g in (1). After the third increment the first and fourth points determine the minimum distance and this distance is g^4. The minimum distance after the fifth point is the same. The minimum distance after the sixth point is g^5. This trend continues, with the minimum distance never decreasing by more than a factor of g when it does decrease. This property follows directly from the definition of the golden section in (1). The same distances can also be generated with the complement increment of $(1-g)=g^2\approx 0.382$. Higher powers of g can also be used for the increment value, but the initial minimum distances are reduced to the smaller increment value.

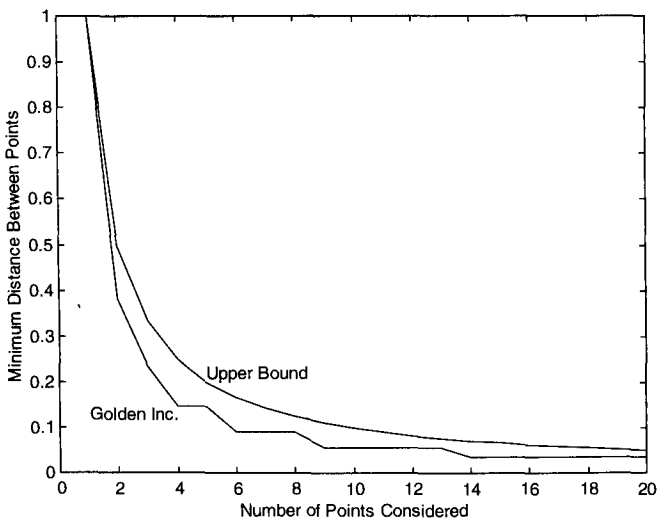


Fig. 3: Minimum distance between points versus number of points with a golden increment.

Figure 3 shows a plot of the minimum distances versus the number of points considered, as points are added using an increment of g with modulo-1 arithmetic. Figure 3 also shows an upper bound for each specific number of points. That is, given n points, and only n points, they could be uniformly spaced with a minimum distance of $1/n$. Of course the golden section increment results are valid for all numbers of points at the same time. The upper bound is not. Even so, the golden section increment results are seen to track the upper bound quite closely. Note that even when the minimum distance drops, most points will still be at the previous minimum distance from their neighbours, with the average distance between points equal to the upper bound.

The distance properties of the golden section increment, illustrated in Figure 3, are desirable for interleavers in general, but in particular are desirable for Turbo-code interleavers. It is now shown how these properties can be used in designing a number of practical interleavers.

3.2 Golden Relative Prime Interleavers

For golden relative prime interleavers, the interleaver indexes are calculated as follows:

$$i(n) = s+np, \text{ modulo } N, \quad n=0\ldots N-1 \qquad (3)$$

where s is an integer starting index, p is an integer index increment, and N is the interleaver length. N and p must be relative primes to ensure that each element is read out once and only once. The starting index s is usually set to 0, but other integer values of s can be selected. The relative prime increment, p, is chosen "close" (as further defined below) to one of the non-integer values of

$$c=N(g^m+j)/r \qquad (4)$$

where g is the golden section value, m is any positive integer greater than zero, r is the index spacing (distance) between nearby elements to be maximally spread, and j is any integer modulo r. As discussed previously, the preferred values for m are typically 1 or 2. In a typical implementation where adjacent elements are to be maximally spread, j is set to 0 and r is set to 1. For Turbo-codes, however, greater values of j and r can be used to obtain the best spreading for elements spaced r apart. For example, r could be set to the repetition period of the feedback polynomial in the RSC encoder, to maximally spread input-weight-2 error events. With this in mind, good choices for j and r are values that result in spreading by approximate golden section spacing for adjacent elements, as well as those spaced by r. For example, j=9 and r=15 are expected to work well for a 16-state Turbo-code with an RSC code repetition period of r=15.

One definition of being a "close" relative prime is to fall within a small window about the exact real value, c, given in (4). The simplest choice is to select the relative prime p closest to c, for predetermined values of N, m, j, and r. The result is a golden relative prime interleaver with quantization error. For large blocks the quantization error is usually not significant for short error-burst lengths, but can grow to be significant after many increments. One way to mitigate the

quantization error problem is to perform a search for the best relative prime increment p in the vicinity of c, by using the minimum distance between interleaved indexes for the maximum number of elements considered, as a measure of the spreading quality of the interleaver. Alternatively, the best relative prime increment, p, in the vicinity of c, is determined by the sum (or weighted sum) of the minimum distances between interleaved indexes for all numbers from two up to the maximum number of elements considered. In this case, the best choice for p is that which maximizes the area under the minimum distance curve.

Figure 4 shows the spreading properties for an interleaver having a size N=1028 (e.g. used in a Turbo-code encoder with 1024 information bits and 4 flush bits per block), m=2, j=0, r=1, and a relative prime increment of p=393. The value of $c=Ng^2$ is approximately 392.7. The value of p=393 is the closest relative prime. As can be seen, this golden relative prime interleaver performs well in tracking the upper bound near the origin, but does not appear to be as good away from the origin where the accumulating quantization error becomes significant. The area under the entire curve is 4620. This spreading measure is used for comparison purposes below.

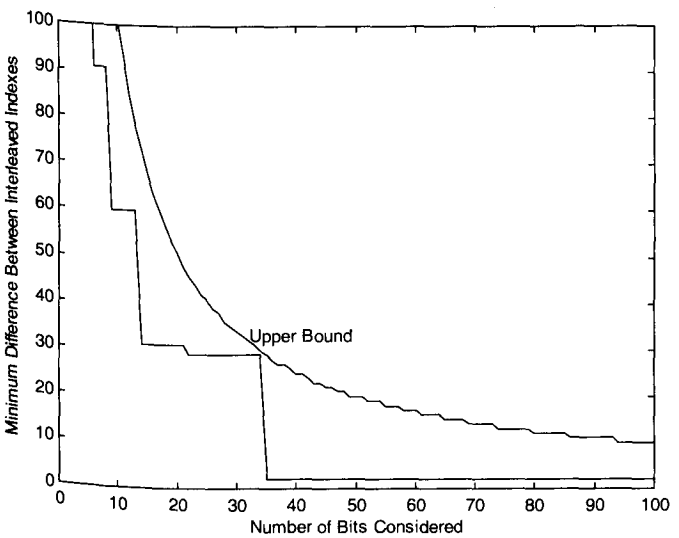


Fig. 4: Minimum difference between interleaved indexes versus number of bits considered with a golden relative prime interleaver. N=1028, p=393, area under curve=4620.

Most general purpose digital signal processors (DSPs) today offer the kind of modulo indexing indicated in (3) to implement circular buffers. It is also trivial to implement in hardware. Thus, golden relative prime interleavers require no additional memory and little or no additional processing compared to that required to store and read an uninterleaved vector.

3.3 Golden Interleavers

Golden interleavers do not use integer relative primes and integer modulo arithmetic, but rather are based on sorting real-valued numbers derived from the golden section. The first step is to compute the golden section value g. The second step is to compute the real increment value c, as defined previously in (4). The third step is to generate real-valued golden vector $\boldsymbol{v}$. The elements of $\boldsymbol{v}$ are calculated as follows:

$$v(n) = s+nc, \text{ modulo } N, \quad n=0\ldots N\text{-}1 \qquad (5)$$

where s is any real starting value. The next step is to sort golden vector $\boldsymbol{v}$ and find the index vector $\boldsymbol{z}$ that defines this sort. That is, find sort vector $\boldsymbol{z}$ such that $a(n)=v(z(n))$, $n=0\ldots N$-1, where $\boldsymbol{a}$=sort($\boldsymbol{v}$). The golden interleaver indexes are then given by $i(z(n))=n$, $n=0\ldots N$-1. In fact, vector $\boldsymbol{z}$ is the inverse interleaver for $\boldsymbol{i}$.

The starting value s is usually set to 0, but other real values of s can be selected. The preferred values for m are typically 1 or 2, as discussed previously. For maximum spreading of adjacent elements, j is set to 0 and r is set to 1. For Turbo-codes, greater values of j and r may be used to obtain the best spreading for elements spaced r apart, as discussed previously.

The golden interleaver does not suffer from accumulating quantization errors, as does the golden relative prime interleaver. In the golden interleaver case, a quantization error only occurs in the final assignment of the indexes. On the other hand, the golden interleaver cannot be implemented using the simple modulo-increment indexing method described above for the golden relative prime interleaver. In contrast, the golden interleaver indexes must be pre-computed and stored in index memory for each block size of interest. If the full indexes are stored, then the index memory can be excessive. For example, an interleaver of length 2^{16} elements would require 16×2^{16} bits of index memory. The required amount of index memory can be significantly reduced by only storing index offsets. For example, the n-th index can be easily calculated as required using $i(n)$=floor[$v(n)$]+$o(n)$, where the floor function extracts the integer part, $v(n)$ is calculated using real modulo N arithmetic as in (5), and by definition $o(n)$ is the required index offset stored in index memory. The number of bits that are required to store each index offset is typically only one or two. Thus, for the example above, the index memory is reduced to 2×2^{16} bits, or about 1/8 that required for full storage of the indexes.

Figure 5 shows the spreading properties for a golden interleaver having size N=1028, m=2, j=0, and r=1. The value of real increment $c= Ng^2$ is approximately 392.7. As can be seen from Figure 5, the golden interleaver performs very well in tracking the theoretical upper bound, and tracks it better than the golden relative prime interleaver curve shown in Figure 4. Note that the area under the curve has increased

from 4620, for the golden relative prime interleaver, to 5250, for the golden interleaver, indicating that the golden interleaver is better at spreading out error-bursts of arbitrary length.

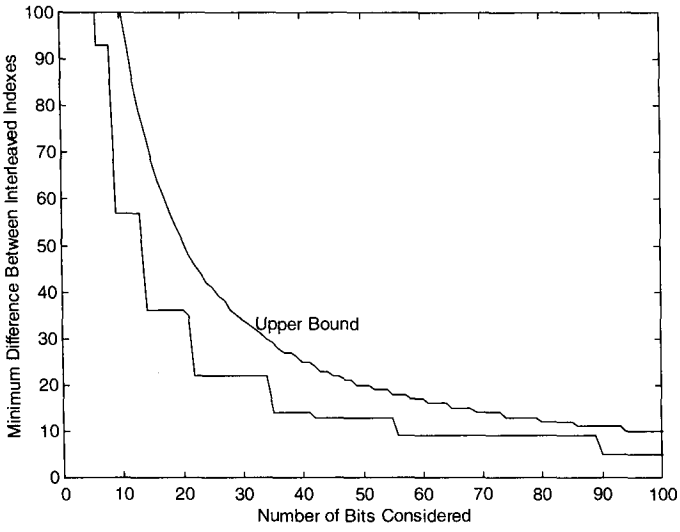


Fig. 5: Minimum difference between interleaved indexes versus number of bits considered with a golden interleaver. N=1028, m=2, j=0, r=1, area under curve=5250.

3.4 Dithered Golden Interleavers

It has been found for Turbo-codes that interleavers with some randomness tend to perform better than completely structured interleavers, especially for large block sizes on the order of 1000 or more bits. However, the spreading properties of the golden interleaver are still very desirable, both to maintain a good minimum distance (a steep error curve at high SNRs) and to ensure rapid convergence by efficiently spreading error-bursts throughout the block. These two features are encompassed in the dithered golden interleaver. The only difference between the golden interleaver and the dithered golden interleaver is the inclusion of a real perturbation (dither) vector $\boldsymbol{d}$, in golden vector $\boldsymbol{v}$. That is,

$$v(n)=s+nc+d(n), \text{ modulo } N, \quad n=0\ldots N-1, \qquad (6)$$

where $d(n)$ is the n-th dither component. The added dither is uniformly distributed between 0 and ND, where D is the normalized width of the dither distribution. The dithered golden vector $\boldsymbol{v}$ is sorted, and interleaver indexes are generated in a similar manner to that for the golden interleaver described above.

It has been found experimentally, for Turbo-codes, that a crude rule of thumb for any block size is to use $D\approx 0.01$. The result is that for small blocks, on the order of 1000 bits or less, the effect of the dither component is small. For large blocks, on the order of 1000 bits or more, the effect of the dither component naturally increases as the block size increases. In practice, the optimum amount of dither for a specific Turbo-code is a function of the block size and the code rate obtained with puncturing.

Similar to the golden interleaver, the dithered golden interleaver requires the use of index memory for storing pre-computed indexes, and therefore cannot be implemented using the simpler method of modulo-increment indexing. As for the golden interleaver, the required amount of index memory can be significantly reduced by only storing index offsets. The amount of memory required now depends on the degree of dither, and whether the dither component is included in the calculation of each approximate index, or whether it is totally accounted for in the stored index offset.

In conclusion, the dithered golden interleaver maintains most of the desirable spreading properties of the golden interleaver, but is also capable of adding randomness to the interleaver to improve Turbo-code performance. Further, the golden interleaver is now just a special case of the dithered golden interleaver with D=0.

4.0 Performance Results

Performance results are presented for a fixed interleaver size of N=1028 (historically selected for a 1024 info-bit block with 4 flush bits). The Turbo-code uses two identical, parallel, 16-state, rate 1/2 RSC codes, with polynomials $(23,35)_8$. The repetition period of the feedback polynomial is r=15. Results are presented for nominal code rates of 1/3 (unpunctured), 1/2 and 4/5. The Turbo decoding method used is the enhanced max-log-APP (*a posterior* probability) approach presented in [4,5,6,7]. This method typically provides performance with 0.1 to 0.2 dB of exact, infinite precision APP processing. The maximum number of decoding iterations was set to 16. A simple early stopping criterion was used, which helped speed up the simulations [6]. A more extensive list of the encoder and decoder specifications is given in [7]. More information on the Turbo decoder can be found at www.crc.ca/fec [17].

The RSC code trellis termination method is critical to the performance of Turbo-codes, especially with good interleavers. A number of generally applicable dual-termination and dual-tail-biting techniques are presented in [15,16]. These termination techniques do not place any restrictions on the interleaver design. The recommended approach for large blocks (>1000 info-bits) is to perform dual-termination. Without termination, both RSC encoders start in the zero-state and both stop in an unknown state. With single-termination, a commonly used approach in the literature, the interleaver includes 4 flush bits, both RSC encoders start in the zero-state, and one RSC encoder is

known to stop in the zero-state. With dual-termination, the interleaver includes 8 flush bits and both RSC encoders are known to start and stop in the zero-state. For large blocks the flush overhead is negligible.

Figure 6 shows the packet error rate (PER) performance for rate 1/2 codes, with the three termination options mentioned above. The same dithered golden interleaver, with m=1, j=0, r=1, and D=0.02, was used in all three cases. This figure clearly shows the importance of proper trellis termination, when a good interleaver is used. The conclusion is quite different for a random interleaver. The bit error rate (BER) results corresponding to Figure 6 are shown in Figure 7. The remaining results were all obtained with dual-termination.

Figure 8 shows the PER results for a code rate of 1/3, and four different interleavers of size N=1028. The interleavers used are the "random" interleaver, the relative "prime" interleaver with p=393 (closest relative prime to Ng^2), the "spread" interleaver with spread parameter S=18, and the dithered "golden" interleaver with m=1, j=0, r=1, and D=0.02. The spread interleaver address generator was obtained from [13]. As expected, the highly structured relative prime interleaver does not perform well at high SNRs due to its inability to break up coupled error events. It does, however, do an excellent job of eliminating the low-distance input-weight-2 events (multiples of 15 to multiples of 15). Note that the random interleaver is not much better for this block length. The spread interleaver offers a significant improvement, but the dithered golden interleaver provides the best performance. Figure 9 shows the corresponding BER results. It is worth noting that the BER results for the random interleaver, at high SNRs, agree quite closely with the theoretical bounds presented in [9,10].

Concerning statistical reliability, 1000 packet errors were counted in the upper portion of each curve. The goal for the lowest point on each curve was to count on the order of 100 packet errors. The least reliable result is for the lowest point on Figures 8 and 9, for which only 20 packet errors were counted. Even so, it is safe to say that the "bend" in the BER curve, for the dithered golden interleaver, is in the vicinity of 10^{-10}.

Figure 10 shows the PER results for a punctured code rate of 4/5, and the same four interleaver types. The random, relative prime, and spread interleavers were exactly the same as before. The best parameters found for the dithered golden interleaver were m=1, j=9, r=15, and D=0.005. The dithered golden interleaver is again the best. What is somewhat surprising is how well the relative prime interleaver performs. This is partly explained by the fact that, for high puncture rates, it becomes more important to eliminate low-distance input-weight-2 events, and the relative prime interleaver is ideally suited to this task. The spread interleaver is not as effective at eliminating such events, and therefore performance is degraded. Figure 11 shows the corresponding BER results. Note that the bend in the BER curve occurs much higher for highly punctured codes. Even so, this high rate code, with a dithered golden interleaver, still provides excellent performance for BERs down to 10^{-7}.

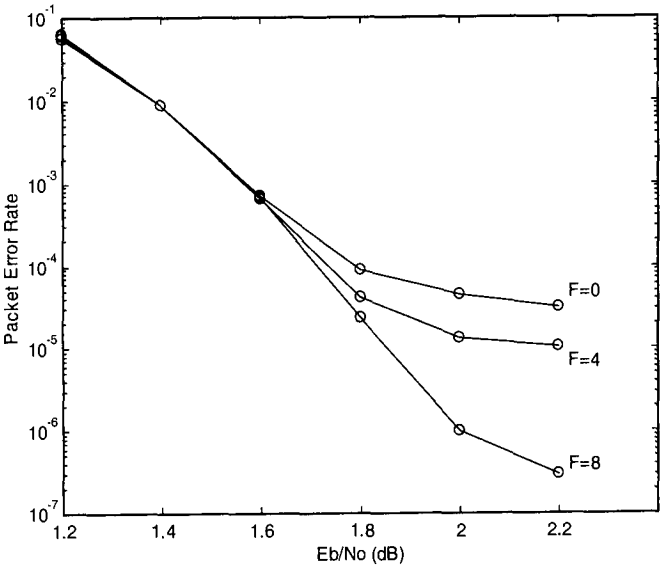


Fig. 6: PER performance for rate 1/2 codes and a dithered golden interleaver with N=1028, m=1, j=0, r=1, and D=0.02. Results are shown for no-termination (number of flush bits F=0), single-termination (F=4), and dual-termination (F=8).

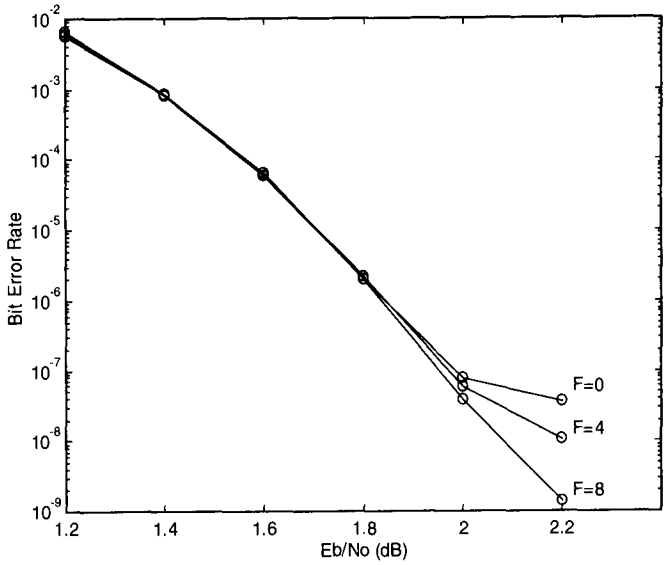


Fig. 7: BER performance for rate 1/2 codes and a dithered golden interleaver with N=1028, m=1, j=0, r=1, and D=0.02. Results are shown for no-termination (number of flush bits F=0), single-termination (F=4), and dual-termination (F=8).

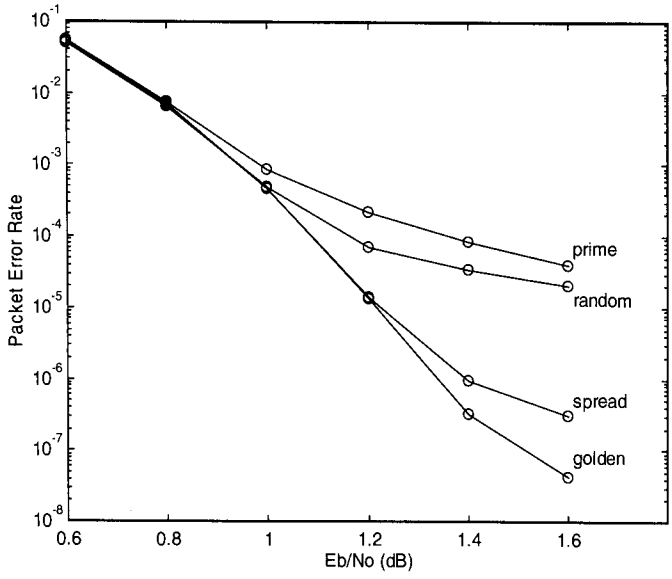


Fig. 8: PER performance for rate 1/3 codes with dual-termination and N=1028. The interleavers used are the "random" interleaver, the relative "prime" interleaver (p=393), the "spread" interleaver (S=18), and the dithered "golden" interleaver (m=1, j=0, r=1, and D=0.02).

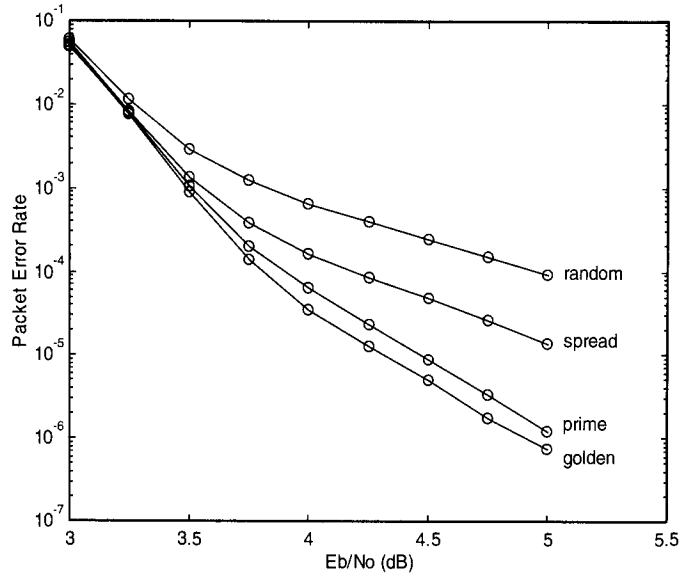


Fig. 10: PER performance for rate 4/5 codes with dual-termination and N=1028. The interleavers used are the "random" interleaver, the relative "prime" interleaver (p=393), the "spread" interleaver (S=18), and the dithered "golden" interleaver (m=1, j=9, r=15, and D=0.005).

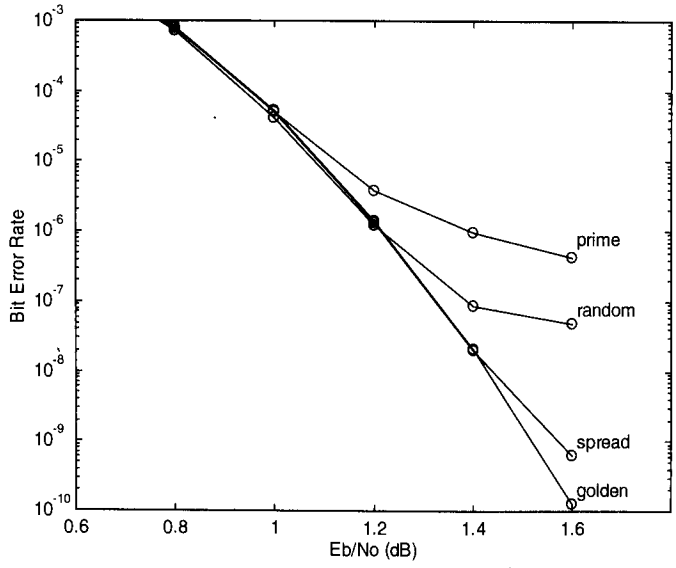


Fig. 9: BER performance for rate 1/3 codes with dual-termination and N=1028. The interleavers used are the "random" interleaver, the relative "prime" interleaver (p=393), the "spread" interleaver (S=18), and the dithered "golden" interleaver (m=1, j=0, r=1, and D=0.02).

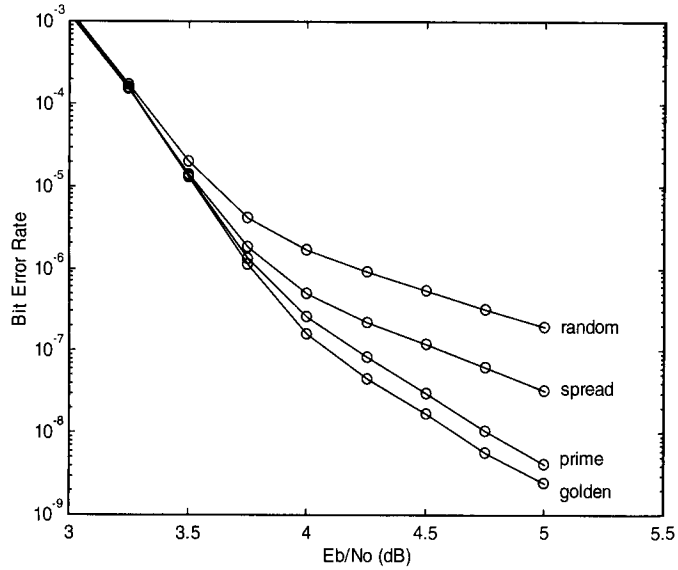


Fig. 11: BER performance for rate 4/5 codes with dual-termination and N=1028. The interleavers used are the "random" interleaver, the relative "prime" interleaver (p=393), the "spread" interleaver (S=18), and the dithered "golden" interleaver (m=1, j=9, r=15, and D=0.005).

5.0 Conclusions

Three new interleavers based on the golden section were presented. They are called the golden relative prime interleaver, the golden interleaver, and the dithered golden interleaver. Random and spread interleavers were also considered. Turbo-code performance results were presented and compared for the various interleavers. The dithered golden interleaver provided the best performance in all cases considered. The golden relative prime interleaver, although highly structured with no random component, worked surprisingly well for high code rates.

Using a dithered golden interleaver of size N=1028, it was shown that a parallel, dual-terminated, 16-state, rate 1/3 Turbo-code can achieve a BER of 10^{-10} at an E_b/N_0 value of 1.6 dB. The bend in the BER curve also occurs at a BER of about 10^{-10}. Puncturing this same code to rate 4/5 moved the bend out and up to a BER of about 10^{-7}. Further improvements should be possible by incorporating more specific knowledge about the punctured component RSC codes into the interleaver design.

The various "golden" interleavers have excellent spreading properties in general and are thus useful for many applications other than Turbo-codes. In addition, there are no restrictions on the block size, and a time-consuming search is not required. Thus, interleavers can be easily generated on an as needed basis for any block length.

References

[1] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-Codes", Proceedings of ICC'93, Geneva, Switzerland, pp. 1064-1070, May, 1993.

[2] C. Berrou, and A. Glavieux, "Near Optimum Error Correcting Coding and Decoding: Turbo-Codes", IEEE Trans. On Comm., Vol. 44, No. 10, October 1996.

[3] B. Talibart and C. Berrou, "Notice Preliminaire du Circuit Turbo-Codeur/Decodeur TURBO4", Version 0.0, June, 1995.

[4] S. Crozier, A. Hunt, K. Gracie, and J. Lodge, "Performance and Complexity Comparison of Block Turbo-Codes, Hyper-Codes, and Tail-Biting Convolutional Codes", 19-th Biennial Symposium on Communications, Kingston, Ontario, Canada, pp.84-88, May 31-June 3, 1998.

[5] K. Gracie, S. Crozier, A. Hunt, and J. Lodge, "Performance of a Low-Complexity Turbo Decoder and its Implementation on a Low-Cost, 16-Bit Fixed-Point DSP", The 10-th International Conference on Wireless Communications (Wireless'98), Calgary, Alberta, Canada, pp.229-238, July 6-8, 1998.

[6] K. Gracie, S. Crozier, and A. Hunt, "Performance of a Low-Complexity Turbo Decoder with a Simple Early Stopping Criterion Implemented on a SHARC Processor", International Mobile Satellite Conference (IMSC'99), Ottawa, Canada, June 16-18, 1999.

[7] A. Hunt, S. Crozier, M. Richards, and K. Gracie, "Performance Degradation as a Function of Overlap Depth when using Sub-Block Processing in the Decoding of Turbo Codes", International Mobile Satellite Conference (IMSC'99), Ottawa, Canada, June 16-18, 1999.

[8] D. Divsalar and F. Pollara, "Multiple Turbo Codes for Deep-Space Communications", JPL, TDA Progress Report 42-121, May 15, 1995.

[9] S. Benedetto and G. Montorsi, "Unveiling Turbo Codes: Some Results on Parallel Concatenated Coding Schemes", IEEE Trans. on Inform. Theory, Vol. 42, No. 2, pp.409-428, March 1996.

[10] S. Benedetto and G. Montorsi, "Tutorial 11: Turbo Codes: Comprehension, Performance Analysis, Design, Iterative Decoding", IEEE International Conference on Communications (ICC'97), Montreal, Quebec, Canada, June 8-12, 1997.

[11] A. Barbulescu and S. Pietrobon, "Interleaver Design for Turbo Codes", Electronics Letters, Vol. 30, No. 25, pp.2107-08, December 8, 1994.

[12] M. Ho, S. Pietrobon, and T. Giles, "Interleavers for Punctured Turbo Codes", IEEE Asia-Pacific Conf. on Commun. (APCC'98) and Singapore Int. Conf. on Commun. Systems, Vol. 2, pp. 520-524, Singapore, November, 1998.

[13] S. Pietrobon, "Interleaver Address Generator", Version 1.01, October 4, 1998, available from Small World Communications. See www.sworld.com.au.

[14] P.-P. Sauvé, "Multibit Decoding of Turbo Codes," Master's Thesis, University of Toronto, Canada, October 1998.

[15] P. Guinand and J. Lodge, "Trellis Termination for Turbo Encoders", Proc. 17th Biennial Symp. On Communications, Queen's University, Kingston, Canada, pp. 389-392, May 30-June 1, 1994.

[16] S. Crozier, P. Guinand, J. Lodge, and A. Hunt, "Construction and Performance of New Tail-Biting Turbo Codes", 6-th International Workshop on Digital Signal Processing Techniques for Space Applications (DSP'98), ESTEC, Noordwijk, The Netherlands, paper 1.3, September 23-25, 1998.

[17] "Ultra-Fast Turbo and Viterbi Decoders for PCs", CD-ROM, available from the Communications Research Centre (CRC), Ottawa, Canada. See www.crc.ca/fec.

Performance Degradation as a Function of Overlap Depth when using Sub-Block Processing in the Decoding of Turbo Codes

Andrew Hunt, Stewart Crozier, Mark Richards, and Ken Gracie
Communications Research Centre (CRC)
Ottawa, Ontario, Canada
E-mail: andrew.hunt@crc.ca
URL: http://www.crc.ca/fec

ABSTRACT

This paper presents results related to the decoding of Turbo codes. In particular, the results relate to the decoding of Turbo codes using sub-blocks, and show the degradation in error-rate performance as a function of the overlap depth used. These results are relevant to the design of high-speed hardware or software Turbo code decoders. Sub-block processing is desirable in such implementations because the technique can provide significant reductions in the memory requirements of the decoder, as well as allowing increased parallelism. There is a computational overhead, however, associated with using sub-block processing, and so it is desirable to keep the overlap depth as short as possible while still achieving the required error-rate performance specifications. Results are presented for a range of overlap depths showing the bit error rate (BER) and packet error rate (PER) performance over an additive white Gaussian noise (AWGN) channel. Results for code rates r=1/3, r=1/2, and r=4/5 are presented, showing the need for longer overlap depths as the code rate is increased.

BACKGROUND

Turbo codes are a broad family of forward-error-correcting (FEC) codes introduced in 1993 by Berrou et al. [1,2]. Sub-block processing is a method commonly used in Turbo code decoding to reduce the memory requirements associated with decoding large blocks. For example, the MAP04B product, an FPGA (field-programmable gate array) design developed by *Small World Communications* that can be used for Turbo code decoding, has a selectable "minimum decoding depth" (i.e. overlap) of either 32 or 64 bits [7]. The use of sub-block processing in this design keeps the memory associated with storing state metric values constant, regardless of the size of the block being processed.

A general discussion on Turbo code decoding methods can be found in [4]. Sub-block processing simply means that the forward state metric calculations, backward state metric calculations, and metric combining operations are performed one sub-block at a time, the size of the sub-blocks typically being a fraction of the size of the total block. Decoding in this fashion means that there are more backward state metric calculations; specifically, for each sub-block, except the last sub-block, an additional "overlap depth" of backward state metric calculations is required. These additional calculations begin one overlap depth into the next sub-block and start with all states weighted equally, and proceed backwards until the boundary of the current sub-block is reached. At this point, the backward state metrics are initialized, the accuracy of the metrics being dependent on the overlap depth being used, and the normal backward state metric calculations and metric combining operations for the current sub-block can begin. This additional processing overhead is the penalty paid in order to realize the memory savings afforded by adopting a sub-block decoding approach.

TURBO CODE AND DECODER SPECIFICATIONS

This paper presents data for three example Turbo codes. The specifications for these example Turbo codes, as well as details on the Turbo code decoding method used, are given below.

Turbo code specifications

- Parallel arrangement
  - Called "parallel concatenation"
  - Like the original paper by Berrou et al. [1,2]
- Two identical constituent codes
  - Recursive systematic convolutional (RSC) codes
  - Binary, rate r=1/2 (each RSC code)
  - Memory-4 (i.e. 16 states)
  - Polynomials (23,35) octal; see [6]
  - Period of feedback polynomial is 15 (i.e. maximal length)
- Termination of both codes [3,5]
  - Both RSC codes end in the all-zeroes state
  - Requires 8 flush bits (because memory-4)
  - No restrictions on interleaver
  - Positioning of flush bits depends on interleaver
- Interleaving
  - Information bits and flush bits are interleaved between the two RSC codes
  - Dithered "golden" interleaving [8]
  - These interleavers are based on the golden section, and have good spreading properties
  - Varying amounts of "dither", depending on the degree of puncturing (see figure captions)
- Block size
  - Interleaver length is 1028 for all cases
  - Number of information bits is 1020
  - Number of coded bits depends on puncturing
- Puncturing to achieve higher code rates

- Overall code rate nominally r=1/3
- Parity bits are punctured to achieve higher code rates (r=1/2, 4/5)

Turbo decoder specifications

- Iterative enhanced-max-log-APP decoding method [4]
  - APP stands for *a posteriori* probability
  - Max-log-APP also known as max-log-MAP (maximum *a posteriori*)
  - Error rate performance typically within 0.1 to 0.2 dB of exact, infinite-precision iterative APP decoding, for rate-1/2 coding
- Iterations
  - Maximum of 16 decoding iterations (i.e. 32 half-iterations)
  - Early-stopping feature used: decoding is halted after two successive half-iterations have same maximum-likelihood sequence estimate (MLSE); see [9]
- Sub-block processing with overlap
  - Overlap depth is the decoding parameter being investigated
  - Block divided into 8 sub-blocks for all cases (therefore, 7 sub-blocks have overlap)
  - Each sub-block about 128 bits long
- More information on decoder at http://www.crc.ca/fec

RESULTS

This paper presents data on the effect of overlap depth on error-rate performance for three example Turbo codes. The simulation details are as follows:

- Overlap depth is the decoding parameter being investigated
- Code rates simulated: 1/3, 1/2, 4/5
- Additive white Gaussian noise (AWGN) channel
- Both bit-error-rate (BER) and packet-error-rate (PER) performance are shown

The error-rate performance degradation caused by sub-block processing with a particular overlap depth will depend on the interleaver being used. With a poor interleaver, a certain overlap depth may not cause any performance degradation, simply because the interleaver itself is dominating the error-rate performance. On the other hand, with a better interleaver, the same overlap depth may indeed result in a performance degradation. That is, the better the interleaver, the longer the overlap depth required to avoid a performance degradation.

The same reasoning applies to the termination of the two constituent RSC encodings. To achieve the best error-rate performance at higher SNR's, it is important to terminate both encodings (or tail-bite). For this reason, all three of the example Turbo codes used in the simulations were double-terminated [3,5].

ANALYSIS

Figures 1, 2, and 3 show that the appropriate amount of overlap depends on the operating point; for lower SNR's, a shorter overlap can be tolerated, whereas at higher SNR's, a longer overlap is required in order not to degrade performance. Also, it is clear that the higher the puncture rate, the longer the overlap depth required. It must be emphasized that the performance data presented are specific to the three example Turbo codes that were simulated, and therefore only very general guidelines can be extracted from these results.

CONCLUSIONS

The appropriate amount of overlap to use when employing sub-block processing in the decoding of Turbo codes depends on many factors. These include the nature of the Turbo code itself (e.g. memory, termination), the degree of puncturing, the SNR at the operating point, the number of iterations, the block size, the interleaver, and the number of sub-blocks. For any potential application, an investigation should be undertaken to determine the overlap depth that provides the optimal trade-off between error-rate performance and decoder processing requirements.

REFERENCES

[1] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-Codes", *Proceedings*, ICC '93 (Geneva, Switzerland, May 1993), pp. 1064-1070.

[2] C. Berrou and A. Glavieux, "Near Optimum Error Correcting Coding and Decoding: Turbo-Codes", *IEEE Transactions on Communications*, Vol. 44, No. 10, October 1996, pp. 1261-1271.

[3] P. Guinand and J. Lodge, "Trellis Termination for Turbo Encoders", *Proceedings*, 17th Biennial Symposium on Communications (Kingston, May 1994), pp. 389-392.

[4] K. Gracie, S. Crozier, A. Hunt, and J. Lodge, "Performance of a Low-Complexity Turbo Decoder and its Implementation on a Low-Cost, 16-Bit Fixed-Point DSP", *Proceedings*, Wireless '98 (Calgary, Canada, July 1998), pp. 229-238.

[5] S. Crozier, P. Guinand, J. Lodge, and A. Hunt, "Construction and Performance of New Tail-Biting Turbo Codes", *Proceedings*, DSP '98 (Noordwijk, The Netherlands, September 1998), paper 1.3.

[6] M. Ho, S. Pietrobon, and T. Giles, "Improving the Constituent Codes of Turbo Encoders", *Proceedings*, Globecom '98 (Sydney, Australia, November 1998), Vol. 6, pp. 3525-3529.

[7] Steven S. Pietrobon, "MAP04B 16 State MAP Decoder Data Sheet", *Small World Communications*, http://www.sworld.com.au.

[8] S. Crozier, J. Lodge, P. Guinand, and A. Hunt, "Performance of Turbo Codes with Relative Prime and Golden Interleaving Strategies", *Proceedings*, IMSC '99 (Ottawa, Canada, June 1999).

[9] K. Gracie, S. Crozier, and A. Hunt, "Performance of a Low-Complexity Turbo Decoder with a Simple Early Stopping Criterion Implemented on a SHARC Processor", *Proceedings*, IMSC '99 (Ottawa, Canada, June 1999).

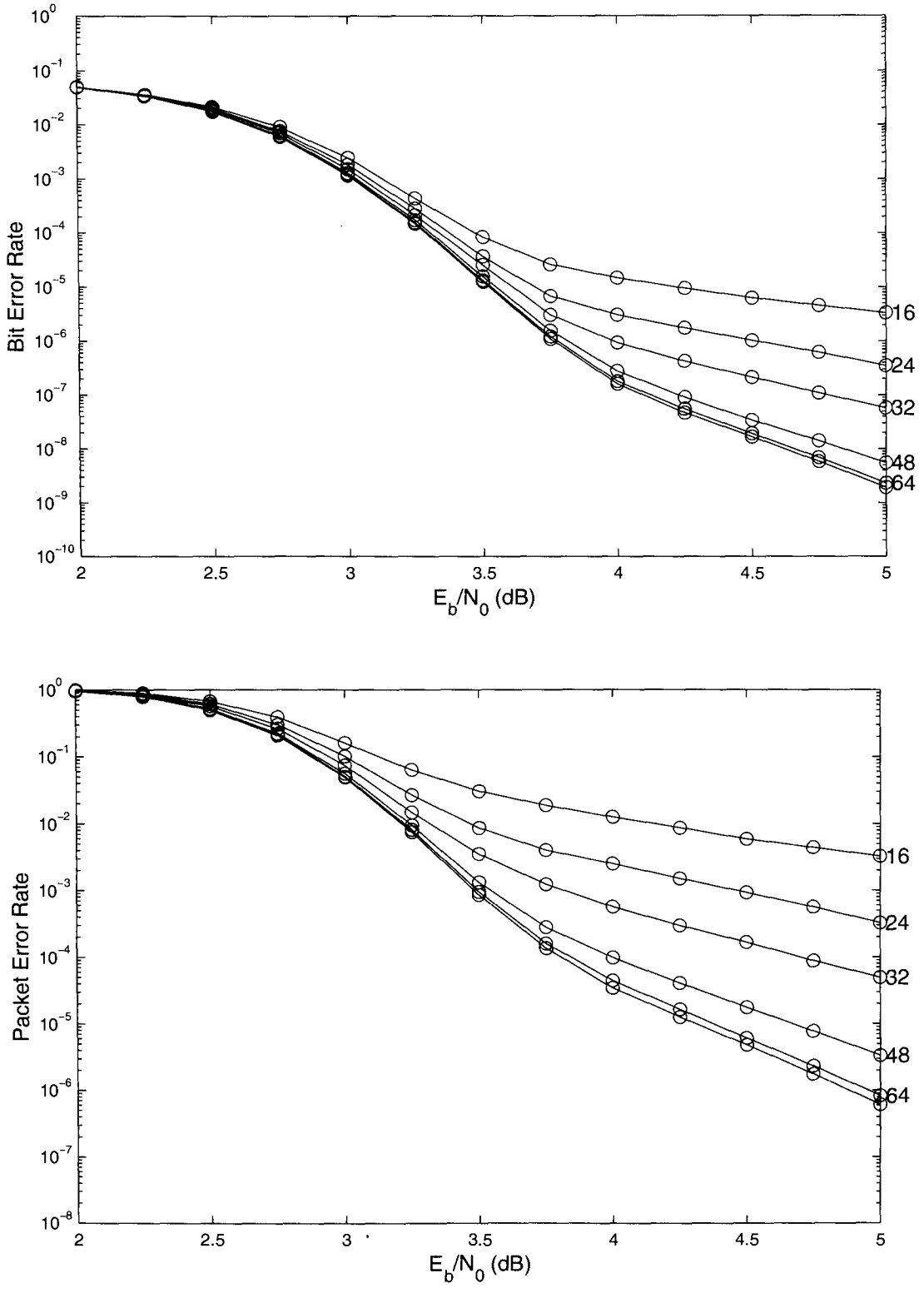
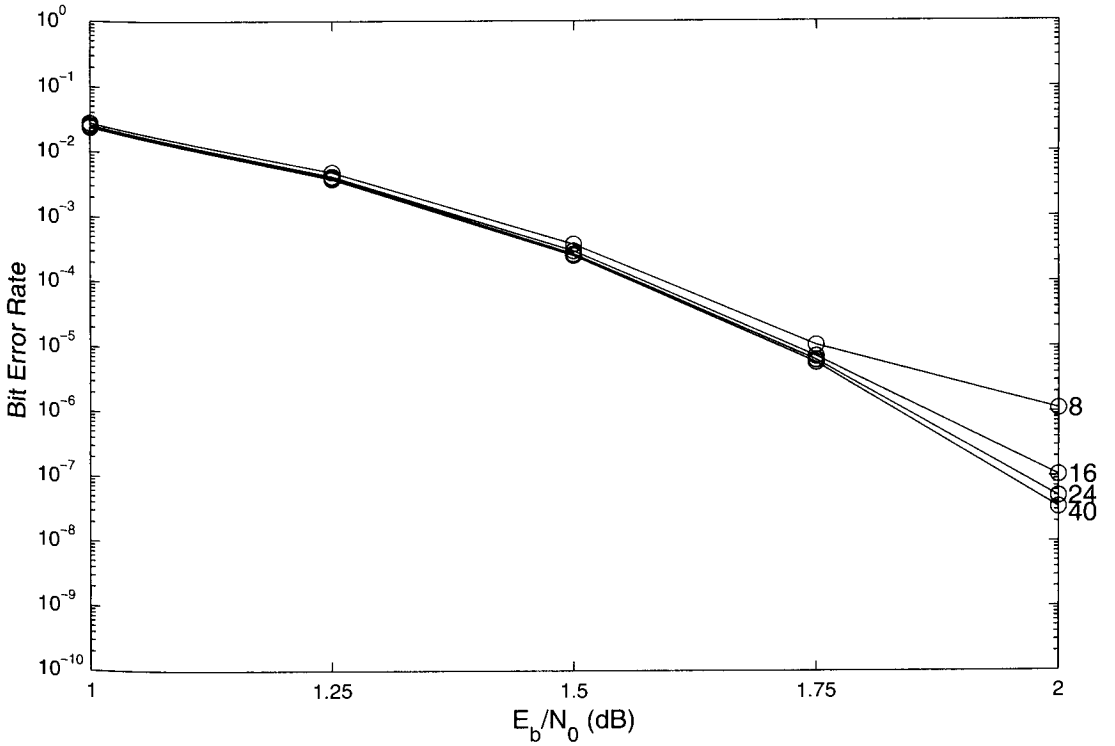


Figure 1. Results for code rate r=4/5 (nominal). Performance for overlap depths of 16, 24, 32, 48, and 64 bits is shown, as well as performance without sub-block processing (the bottom curves). The interleaver used was a type of dithered golden interleaver designed to spread bits separated by multiples of the period, and the dither was 0.5 percent. The block size is (n,k)=(1285,1020).



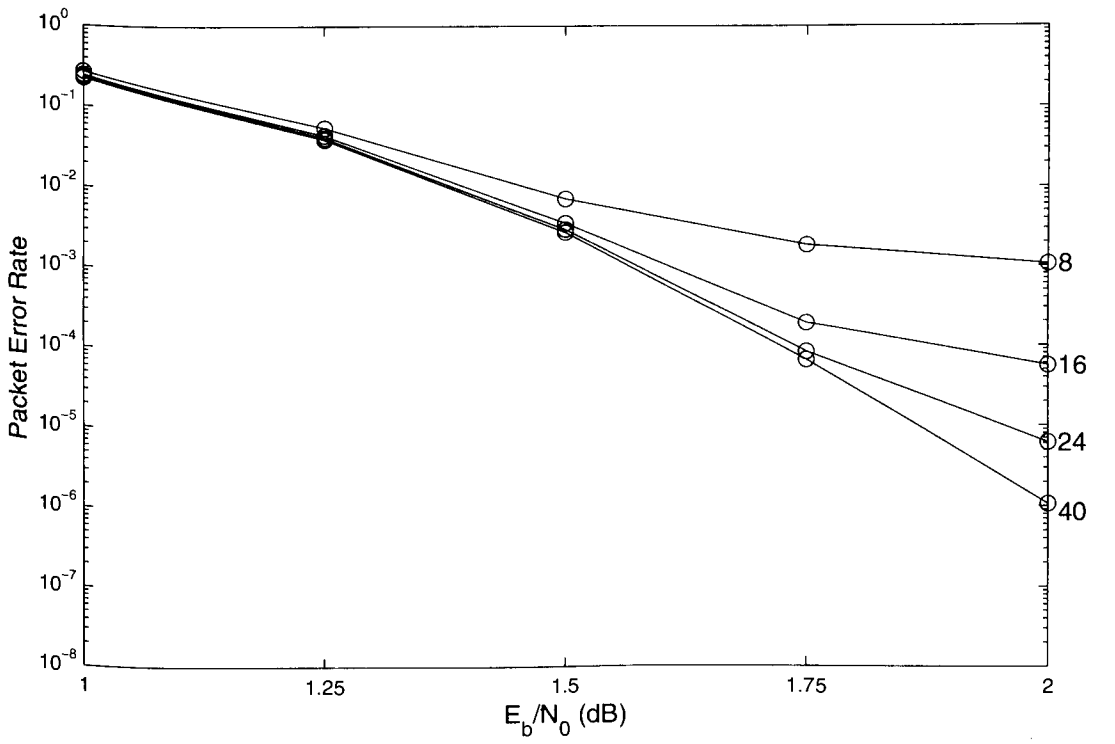


Figure 2. Results for code rate r=1/2 (nominal). Performance for overlap depths of 8, 16, 24, and 40 bits is shown. Performance without sub-block processing is effectively equivalent to that with 40 bits of overlap, for the SNR range shown. The interleaver used was a dithered golden interleaver designed to spread adjacent bits, and the dither was 2 percent. The block size is (n,k)=(2056,1020).

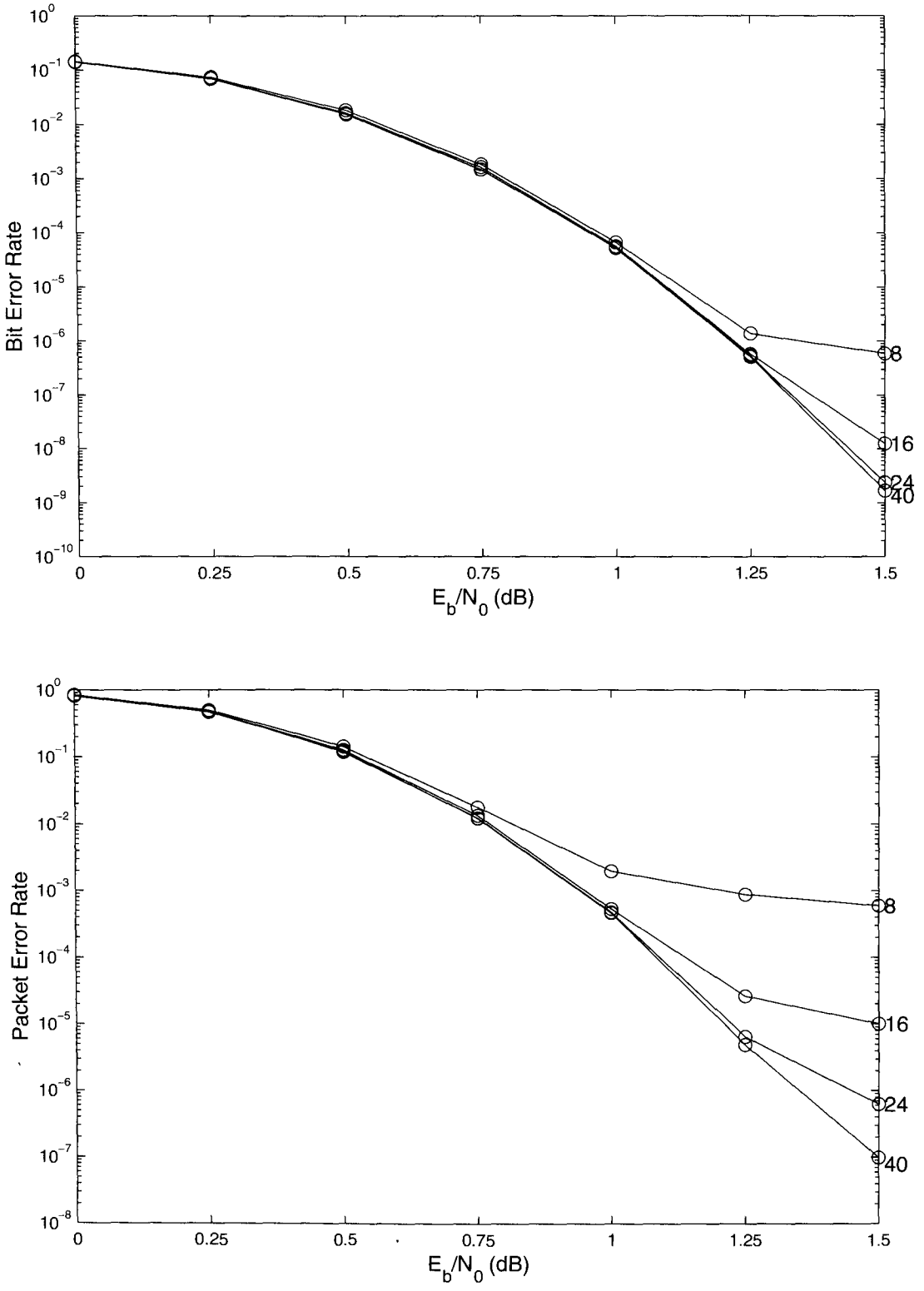


Figure 3. Results for code rate r=1/3 (nominal). Performance for overlap depths of 8, 16, 24, and 40 bits is shown. Performance without sub-block processing is effectively equivalent to that with 40 bits of overlap, for the SNR range shown. The interleaver used was the same as that of the rate-1/2 code. The block size is (n,k)=(3084,1020).

Performance of a Low-Complexity Turbo Decoder with a Simple Early Stopping Criterion Implemented on a SHARC Processor

Ken Gracie, Stewart Crozier, and Andrew Hunt
Communications Research Centre
3701 Carling Avenue, P.O. Box 11490, Station H
Ottawa, Ontario, Canada, K2H 8S2
Phone: 613-990-5846, FAX: 613-990-6339
Email: ken.gracie@crc.ca
Web: www.crc.ca/fec

ABSTRACT

A modified turbo decoder structure using the max-log-*a-posteriori-probability* (APP) algorithm with correction is described, followed by a simple method of halting the decoding process when a packet has converged early. Use of the early stopping feature is shown to substantially increase average throughput as the signal-to-noise ratio (SNR) rises while incurring virtually no loss in either bit error rate (BER) or packet error rate (PER) performance.

INTRODUCTION

Turbo-codes have attracted much attention due to their excellent error rate performance [1,2], but have often been viewed as too computationally intensive for some applications. This paper presents a practical and effective method of stopping the turbo decoding process early, significantly reducing the average amount of processing required.

A low-complexity turbo decoding structure that has been used effectively with this early stopping criterion is presented first. This structure utilizes the max-log-APP algorithm [3,4] in the constituent decoder and then performs a correction operation on the resulting log-likelihood ratios (LLRs). This approach typically yields error rate performance within 0.1 dB to 0.2 dB of the so-called "true MAP" or "true-APP" decoding method, assuming an equal number of decoding iterations.

The early stopping criterion compares successive sets of hard decisions derived from the max-log-APP decoder. When these data sets match, the decoder is considered to have converged and decoding ceases on the current block. It has been found that using this criterion results in substantial throughput gains which increase with SNR while incurring virtually no performance degradation relative to cases where the number of decoding iterations is fixed.

A decoder using these methods has been implemented on the ADSP-21061 SHARC processor from Analog Devices. The performance of this implementation is presented in terms of BER, PER, and throughput for block sizes between 64 and 1024 information bits. Results with and without the early stopping feature are shown for block sizes of 256, 512, and 1024 information bits. Using this early stopping technique on a 40 MIPS version of the SHARC, a block size of 512 information bits achieves an average throughput greater than 70 kbps for BER values below 10^{-4}.

LOW-COMPLEXITY TURBO DECODING

The most common turbo encoder structure is shown in Figure 1 [1,2]. Two identical, constraint length K=5, rate 1/2 *recursive systematic convolutional* (RSC) encoders operate in parallel on different versions of the input block, one non-interleaved (RSC1) and one interleaved (RSC2).

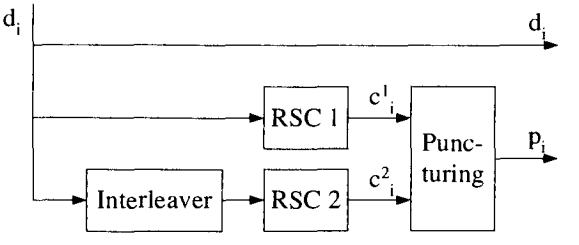
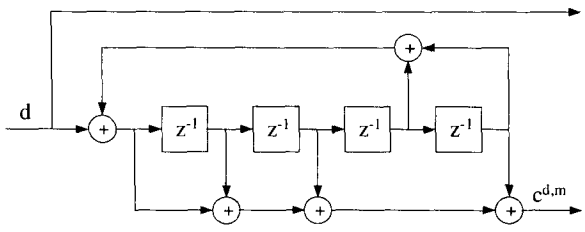


Figure 1 : The turbo encoder



Feedback = $(23)_8 = (10011)_2$
Feedforward = $(35)_8 = (11101)_2$

Figure 2: The constituent RSC encoder. The most significant bit of each polynomial refers to the leftmost tap in the diagram.

The constituent encoder implements the so-called "TURBO4" polynomials [5] and is shown in Figure 2. Both are initialized to the all-zeroes state for each block. Since each encoder is rate 1/2 and the systematic bits only need to be transmitted once, the nominal output code rate of the turbo encoder is 1/3. Other code rates can be realized easily with puncturing.

In addition, a set of K-1 termination or flush bits that return RSC1 to the all-zeroes state is appended to each block of information bits. These same flush bits are interleaved and re-encoded by RSC2, causing it to terminate in an arbitrary state. The inclusion of these flush bits lowers the effective code rate slightly relative to using no flush bits; performance can be improved and this code rate reduction potentially removed by using dual termination or tail-biting [6,7].

As shown in Figure 3, the turbo decoder consists of two stages, one for each set of parity, separated by an interleaver. Each stage is composed of extrinsic subtraction, a constituent max-log-APP decoder, and the extrinsic and LLR correction operations. A number of other algorithms have been proposed for the constituent decoder including the so-called "true-APP" and "log-APP" algorithms as well as the soft output Viterbi algorithm (SOVA) [3,4]. The max-log-APP algorithm is very simple and, with the correction operation, also very effective.

Like other methods, the max-log-APP algorithm calculates approximate LLRs for each input sample as an estimate of which possible information bit was transmitted at each sample time. The LLRs are calculated according to

$$L_i = \max_m[A_i^m + D_i^{0,m} + B_{i+1}^{f(0,m)}] - \max_m[A_i^m + D_i^{1,m} + B_{i+1}^{f(1,m)}] \quad (1)$$

where i is the bit time index, N_s is the number of encoder states, m is the present state ($m = 0, \ldots, N_s\text{-}1$), $f(d,m)$ is the next state given present state m and input bit $d \in \{0,1\}$, A_i^m is the *forward state metric* for state m at time i, B_i^m is the *reverse* or *backward state metric* for state m at time i, and $D_i^{d,m}$ is the branch metric at time i given present state m and input bit $d \in \{0,1\}$. The forward state metrics are calculated starting at the first sample received in the block and moving forward in time to the last sample received. The reverse state metrics are calculated starting at the last sample received and moving backward in time to the first sample received. More formally, the state and branch metrics are given by

$$A_i^m = \max[A_{i-1}^{b(0,m)} + D_{i-1}^{0,b(0,m)}, A_{i-1}^{b(1,m)} + D_{i-1}^{1,b(1,m)}] \quad (2)$$

$$B_i^m = \max[D_i^{0,m} + B_{i+1}^{f(0,m)}, D_i^{1,m} + B_{i+1}^{f(1,m)}] \quad (3)$$

$$D_i^{d,m} = \frac{1}{2}(x_i d' + y_i c'^{d,m}) \quad (4)$$

where $b(d,m)$ is the previous state given present state m and previous input bit $d \in \{0,1\}$, x_i is the i^{th} systematic sample, y_i is the i^{th} parity sample, d is a systematic bit, $c^{d,m}$ is the corresponding coded bit given state m and bit d, $d' = 1-2d$, and $c'^{d,m} = 1-2c^{d,m}$. The state metrics provide a measure of the probability that state m is the correct one at time i, while the branch metrics are a measure of the probability that each possible combination of encoder outputs is the correct one given the channel outputs x_i and y_i. Note that the calculations in (2) and (3) are exactly the Viterbi algorithm without history and will therefore find the same winning path through the decoding trellis given the same inputs and initial conditions. Also, in contrast to the true-APP or log-APP algorithms, no estimate of the channel SNR is required.

The max-log-APP algorithm is sub-optimum due to the approximations involved. However, most of the

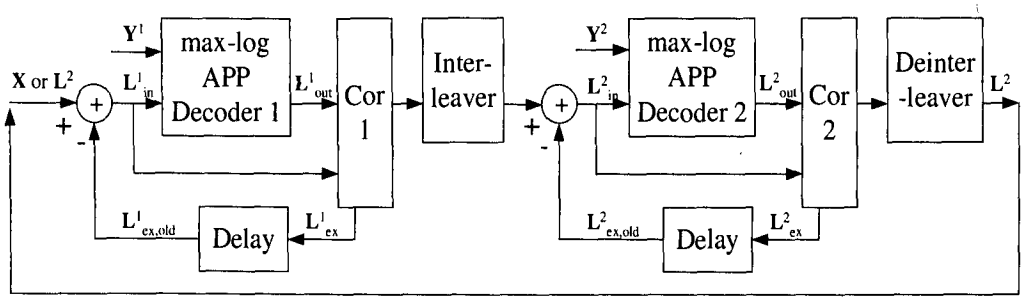


Figure 3: The turbo decoder using max-log-APP decoding and scale factor correction.

performance loss associated with this sub-optimality can be recovered by applying a simple correction factor to the output of the constituent decoder. The so-called *extrinsic information* may be approximated as

$$L_{ex}^{n} = sf \cdot \left(L_{out}^{n} - L_{in}^{n}\right) \qquad (5)$$

where $n \in \{1,2\}$ denotes one of the constituent decoders, L^{n}_{out} represents the set of LLRs produced by the max-log-APP decoder, L^{n}_{in} represents the set of input LLRs, and sf is an appropriate scale factor. Corrected LLRs are then calculated according to

$$\begin{aligned} L_{cor}^{n} &= L_{in}^{n} + L_{ex}^{n} \\ &= L_{in}^{n} + sf \cdot (L_{out}^{n} - L_{in}^{n}) \end{aligned} \qquad (6)$$

These corrected LLRs are the systematic input to the next constituent decoder, while the extrinsic information L^{n}_{ex} is fed back to be subtracted off on the next iteration. For equal numbers of iterations, it has been found that setting sf=5/8=0.625 yields performance within approximately 0.1 dB to 0.2 dB of a true-APP decoder while vastly reducing the amount of computation required.

As is customary in discussions of turbo-codes, this document defines one decoding iteration to be a single execution of both decoding stages, i.e. *max-log-APP Decoder 1* and *max-log-APP Decoder 2* are each invoked once. One *half-iteration* is therefore defined as the execution of a single stage of the decoder, or one max-log-APP decoder.

A SIMPLE EARLY STOPPING CRITERION

An effective strategy for further reducing computational requirements is to stop decoding on each received packet as soon as it has converged, eliminating processing that does not affect the final bit decisions. This requires a reliable and efficient convergence test, such as the following:

> During each max-log-APP decoding operation, generate and store hard bit decisions corresponding to the *pseudo-maximum-likelihood* (pseudo-ML) path through the decoding trellis. Compare these decisions with the pseudo-ML decisions that were generated during the previous half-iteration, i.e., by the previous max-log-APP decoder. When the two sets of decisions match, stop decoding on the current block and return the pseudo-ML bits as output.

The term "pseudo-ML" refers to a path through the decoding trellis for one RSC code that is identical to that produced by the Viterbi algorithm given the same inputs. For independent inputs this will be the true ML path, but this independence assumption is not strictly true after the first half-iteration. The early stopping feature is controlled via two parameters that specify the minimum desired number of iterations (I_{min}) and the maximum desired number of iterations (I_{max}). Both quantities are a multiple of 0.5.

In order for this test to be practical, it must be possible to efficiently isolate and compare the pseudo-ML bits. The bits can be easily set aside if a so-called "retrace" through the trellis along the pseudo-ML path is already used in the calculation of approximate LLRs [8]. No extra processing is required to determine the bits. The bit comparison itself nominally requires as many clock cycles as there are systematic samples in the block, but can be simplified by halting the test as soon as a single discrepancy is observed. This eliminates most of the processing involved, since most of the bits will agree only when the block is nearing convergence, and discrepancies could occur anywhere in the block.

Other convergence tests were investigated, including comparing the signs of the corrected LLRs and the signs of the scaled extrinsic information. However, simulations indicate that comparing pseudo-ML decisions yields better results than either of these, both in terms of error-rate performance and the average number of iterations required. That testing the pseudo-ML sequence is better than testing decisions on the LLRs may be intuitively justified as follows. Testing the pseudo-ML bit sequence is equivalent to testing hard decisions made on LLRs which have been corrected with sf=1 rather than sf=0.625 in (5) and (6). With sf=1, a sign change from one decoding stage to the next is more likely, since the two decoders will more confidently assert disagreements regarding a given bit. That is, using sf=1 weights the information gleaned from each set of parity more heavily, so that if the estimates from the two decoders differ, that difference is more likely to be reflected in the signs of the LLRs.

Several other convergence criteria are proposed in [9] and [10]. The amount of computation involved appears to be greater than that required by the pseudo-ML sequence test, though perhaps only marginally so in some cases. Performance results are given in [10] using two K=3, rate 1/2 RSC encoders, block sizes of 900 and 10,000 systematic bits, and "true-APP" decoding with a maximum of 6 iterations. The HDA criterion proposed in [10] appears to be similar to the approach presented here, though it does not compare pseudo-ML bit sequences. However, very different codecs and parameters make it extremely difficult to compare results.

Finally, it is important to note that reductions in processing due to any early stopping criterion will yield increases in *average* throughput only. That is, some packets will decode more quickly than others, resulting in varying decoding delays. I_{max} could be set such that the maximum decoding delay was matched to the data rate, but this would leave the decoder idle whenever a packet decoded early. Taking advantage of the early stopping feature therefore implies that packets must be buffered at the

decoder input. This increases average throughput and also improves average performance, since a higher value of I_{max} can be used. At the same time, using such a buffer means that it will sometimes fill up and require processing on the current packet to stop before it has converged. This will cause a performance degradation that is a function of buffer length, but this degradation can be minimized by optimizing the value of I_{max} for a given buffer length.

PERFORMANCE RESULTS

A decoder using this early stopping feature has been implemented on a 40 MHz (40 MIPS) version of the SHARC processor from Analog Devices. Figure 4 and Figure 5 show its BER and PER performance, respectively, in additive white Gaussian noise (AWGN) for block sizes between 64 and 1024 information bits, a nominal code rate of 1/2, and 4 and 8 full iterations of the decoder. These results matched those gathered from a PC simulation running on a Pentium processor which used the same decoder structure and interleavers [11].

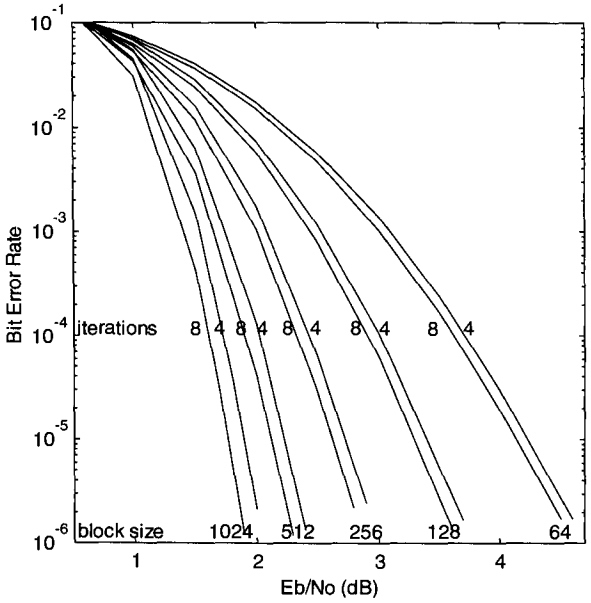


Figure 4: BER performance of the SHARC turbo decoder for 4 and 8 full iterations (nominal code rate=1/2).

Figure 6 and Figure 7 show the BER and PER performance in AWGN, respectively, of the decoder for I_{min}=1.0 and I_{max}=8.0. Results for 8 full iterations from Figure 4 and Figure 5 are also given for the same test points, showing negligible difference in performance. I_{min} is set at 1.0 so that the decoder can stop as early as possible without having to make hard decisions directly on the channel outputs. The values included beside the test points show the measured average throughput in kilo-bits per second (kbps) when early stopping is enabled. At low SNRs, few packets are converging before the maximum number of iterations are performed, resulting in the same average throughput as that observed for 8 full iterations. However, as the SNR increases, a greater proportion of the received packets are converging early, increasing the throughput dramatically (Table 1). Large gains in throughput can therefore be realized with little or no loss in error rate performance, i.e., performance is virtually the same as a decoder with I_{min}= I_{max}. For example, for k=512 and I_{max}=8.0, throughput rose from 22.7 kbps at E_b/N_o=0.0 dB to 75.1 kbps at E_b/N_o=2.0 dB (BER $\cong$ 4×10^{-5}), corresponding to a reduction from 8.0 to an average of approximately 2.2 decoding iterations. The ADSP-2116x "Hammerhead", also from Analog Devices, advertises a dual SHARC core on a single 100 MHz (200 MIPS) part. A straightforward scaling by 5 (200 MIPS/40 MIPS) yields average throughputs between 113 kbps and approximately 370 kbps for the parameters just quoted, which is sufficient for a wide variety of applications. The so-called "TigerSHARC" promises even higher throughputs.

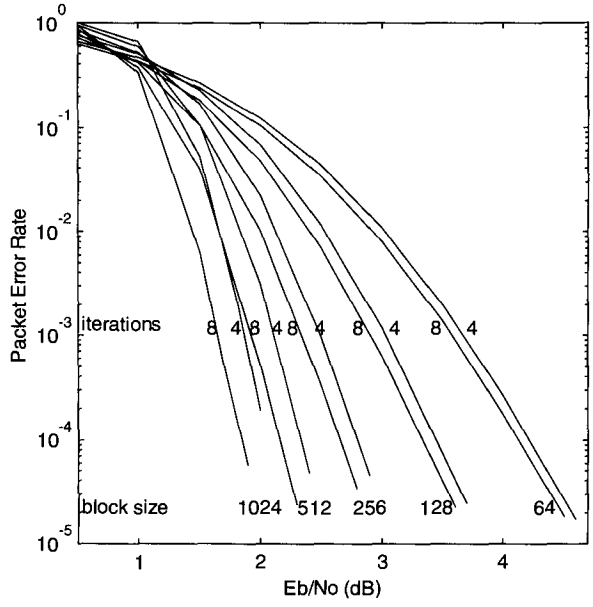


Figure 5: PER performance of the SHARC turbo decoder for 4 and 8 full iterations (nominal code rate=1/2).

An important variable to be aware of when evaluating the performance of turbo-codes is the quality of the interleaver. All of the results shown here used simple *relative prime* interleavers [12]. This type of interleaver has been found to work well for small block lengths such as those presented here. For block sizes of approximately 1000 bits or larger, either a random interleaver or a so-called "dithered golden" interleaver yields superior performance [12], particularly in terms of where the flattening of the error-rate curve occurs. For a block size of 1024 information bits, a degradation in BER performance of less than 0.1 dB was observed above BER=10^{-6} compared to using a dithered golden interleaver. Though not apparent from the figures, it is expected that the flattening of the curve for our relative prime interleaver will occur just below the lowest point that has been shown, i.e., in the vicinity of BER=10^{-6}, which is below many operating points.

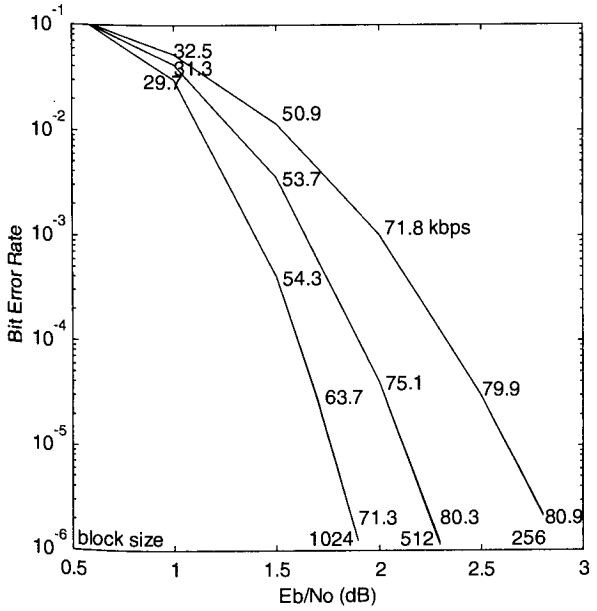


Figure 6: BER performance of the SHARC turbo decoder with early stopping enabled (I_{min}=1.0, I_{max}=8.0, nominal code rate=1/2). Performance with 8 full iterations is also shown. Measured throughput in kbps is shown beside several of the points.

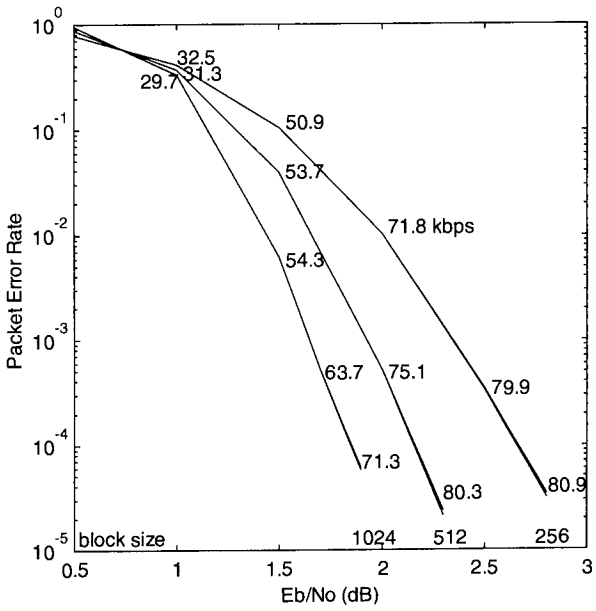


Figure 7: PER performance of the SHARC turbo decoder with early stopping enabled (I_{min}=1.0, I_{max}=8.0, nominal code rate=1/2). Performance with 8 full iterations is also shown. Measured throughput in kbps is shown beside several of the points.

The maximum block size that the decoder can accommodate is determined by the amount of available memory. The ADSP-21061,which offers 0.5 MBits of on-chip data memory and 0.5 MBits of on-chip program memory [13], can handle block sizes of approximately 3000 information bits for code rates of 1/2 or higher. The ADSP-21060, which offers 2 MBits of on-chip data memory and 2 MBits of on-chip program memory, can support block sizes of approximately 18,000 information bits for code rates of 1/2 or higher. No external memory is required in either case, making for a highly flexible single-chip solution. These two processors represent the minimum and maximum amounts of internal memory, respectively, that are available with members of the ADSP-2106x SHARC family. It is significant that these limits could not be achieved without using overlap processing in the decoder [14]. Without overlap processing, the ADSP-21061 would only be able to accommodate a maximum block size of approximately 650 information bits without requiring external memory. The 1024 bit block presented here makes use of overlap processing with an overlap interval of 20 samples, and does so with no noticeable degradation in error rate performance and a reduction in throughput of less than 1%. However, different scenarios, for example higher puncture rates, may require larger overlap intervals [14].

The number of discrepancies that the early stopping criterion will tolerate and still declare convergence is a potential decoder parameter. In increasing this parameter, the hope is that the decoder will be able to declare convergence after fewer half iterations, resulting in increased throughput at the cost of some loss in performance that is hopefully small. Conceptually, performance could be traded off against throughput by varying the decoder's convergence criterion in this way. However, this has not been found to be useful. Allowing even one discrepancy results in a throughput gain of less than 1 kbps at the cost of more than a tenth of a dB in PER performance at PER=10^{-2}. Allowing no discrepancies yields the performance shown in Figure 6 and Figure 7 and is the preferred approach.

Processor	I_{min}	I_{max}	kbps, 512 info. bits, BER=$4*10^{-5}$	Avg. # of Iterations
ADSP-2106x SHARC (40 MIPS)	4.0	4.0	47.1	4.0
	8.0	8.0	22.6	8.0
	1.0	8.0	75.1	2.2
ADSP-2181 (40 MIPS)	4.0	4.0	16.8	4.0
	8.0	8.0	8.5	8.0
Pentium II (400 MHz)	8.0	8.0	224	8.0
	1.0	8.0	773	2.2

Table 1: Measured throughputs and corresponding average numbers of iterations for several implementations of the turbo decoder (nominal code rate=1/2). Results for the ADSP-2181 are taken from [15], while results for the Pentium II were gathered from [11].

Another modification is to require the early stopping criterion to match pseudo-ML bit sequences for more than two half-iterations in a row. That is, rather than having

only two successive sets of bits agree, the decoder is forced to have three or more in agreement before decoding stops. Simulations where three successive sets of bits must agree have shown that this increases the average number of iterations by approximately 0.5 while yielding no gain in performance above the flattening of the curve and gains of a few hundredths of a dB in the flat portion of the curve. Once again, taking the approach presented here is preferable.

CONCLUSIONS

A low-complexity turbo decoding structure and a simple early stopping criterion that compares pseudo-ML bit sequences was presented. The performance of the decoder on the ADSP-2106x SHARC processor was given with and without the early stopping feature. It was found that allowing no discrepancies between pseudo-ML bit sequences and requiring only two sequences in a row to match was an effective way to stop decoding early and achieve performance within a few hundredths of a dB of full iteration performance. Average throughput for a block size of 512 information bits was observed to rise from 22.7 kbps at E_b/N_o=0 dB to 75.1 kbps at E_b/N_o=2.0 dB (BER $\cong$ $4x10^{-5}$) for a maximum of 8 iterations (I_{max}=8.0) on a 40 MIPS device. With the newer processors expected soon from Analog Devices, the decoder could achieve throughputs in excess of 370 kbps.

REFERENCES

[1] C. Berrou and A. Glavieux, "Near Optimum Error Correcting Coding and Decoding: Turbo-Codes", *IEEE Transactions on Communications*, Vol.44, No.10, pp. 1261-1271, October 1996.

[2] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-Codes", *Proceedings of the 1993 International Conference on Communications (ICC '93)*, Geneva, Switzerland, pp.1064-1070, May, 1993.

[3] P. Robertson, P. Hoeher, and E. Villebrun, "Optimal and Sub-Optimal Maximum a Posteriori Algorithms Suitable for Turbo Decoding", *IEEE Communications Theory*, Vol. 8, No. 2, pp. 119-125, March-April 1997.

[4] P. Robertson, E. Villebrun, and P. Hoeher, "A Comparison of Optimal and Sub-Optimal MAP Decoding Algorithms Operating in the Log Domain", *Proceedings of the 1995 International Conference on Communications (ICC'95)*, Seattle, WA, USA, pp. 1009-1013, June 1995.

[5] B. Talibart and C. Berrou, "Notice Preliminaire du Circuit Turbo-Codeur/Decodeur TURBO4", Version 0.0, June, 1995.

[6] P. Guinand and J. Lodge, "Trellis Termination for Turbo Encoders", *Proceedings of the 17th Biennial Symposium on Communications*, Queen's University, Kingston, ON, Canada, pp. 389-392, May 30-June 1, 1994.

[7] S. Crozier, P. Guinand, J. Lodge, and A. Hunt, "Construction and Performance of New Tail-biting Turbo Codes", *Proceedings of the 6th International Workshop on Digitial Signal Processing Techniques for Space Applications (DSP '98)*, ESTEC, Noordwijk, The Netherlands, paper 1.3, September 23-25, 1998.

[8] S. Crozier, K. Gracie, and A. Hunt, "Efficient Turbo Decoding Techniques", *Submitted to the 11th International Conference on Wireless Communications (Wireless '99)*, Calgary, AB, Canada, July 12-14, 1999.

[9] J. Hagenauer, E. Offer, and L. Papke, "Iterative Decoding of Binary Block and Convolutional Codes", *IEEE Transactions on Information Theory*, Vol. 42, No. 2, pp. 429-445, March 1996.

[10] R. Y. Shao, M. Fossorier, and S. Lin, "Two Simple Stopping Criteria for Iterative Decoding", *Proceedings of the 1998 International Symposium on Information Theory*, MIT, Cambridge, MA, USA, p. 279, August 16-21, 1998.

[11] "Ultra-Fast Turbo and Viterbi Decoders for PCs", CD-ROM, available from the Communications Research Centre, Ottawa, Ontario, Canada, 1998 (http://www.crc.ca/fec).

[12] S. Crozier, J. Lodge, P. Guinand and A. Hunt, "Performance of Turbo Codes with Relative Prime and Golden Interleaving Strategies", *International Mobile Satellite Conference '99 (IMSC '99)*, Ottawa, Ontario, Canada, June 16 - 18, 1999.

[13] Analog Devices, "ADSP-2106x SHARC User's Manual", July 1996.

[14] A. Hunt, M. Richards, S. Crozier and K. Gracie, "Performance Degradation as a Function of Overlap Depth when using Sub-Block Processing in the Decoding of Turbo Codes", *International Mobile Satellite Conference '99 (IMSC '99)*, Ottawa, Ontario, Canada, June 16 - 18, 1999.

[15] K. Gracie, S. Crozier, A. Hunt, and J. Lodge, "Performance of a Low-Complexity Turbo Decoder and its Implementation on a Low-Cost, 16-Bit Fixed-Point DSP", *Proceedings of the 10th International Conference on Wireless Communications (Wireless '98)*, Calgary, AB, Canada, pp. 229-238, July 6-8, 1998.

Turbo Code Performance over Aeronautical Channel for High Rate Mobile Satellite Communications

Mohammad S. Akhter[1], Mark Rice[2], and Feng Rice[3]

[1]Institute for Telecommunications Research
University of South Australia
Levels Campus, Mawson Lakes
SA 5095, Australia
mohammad.akhter@unisa.edu.au

[2]DSpace
Innovation House, First Avenue,
Technology Park, Mawson Lakes,
SA 5095 Australia
mark@dspace.com.au

[3]CSSIP, SPRI Bldg.
Mawson Lakes, SA 5095,
Australia
feng@cssip.edu.au

ABSTRACT

This paper presents results from analysis into the performance of Turbo Codes over the aeronautical satellite channel. Currently, only limited services (voice, fax, low rate data) are available for aeronautical travelers. A major constraint is the cost of providing an aircraft qualified antenna with high gain. This limits the data transmission rates that can be supported. In the future, more powerful spot beam satellites with greater EIRP and G/T will provide high rate data services to ground based users and similar services will become feasible for air travelers. In parallel with developments of satellite technology, turbo coding has been proven to be a major advance in power efficient transmission for satellite communications.

A key problem arises for high rate transmissions over the aeronautical channel due to the differential delay between the line of sight path and the reflected diffuse component gives rise to frequency selective fading. The bit error rate performance is found to degrade significantly over the aeronautical channel due to inter-symbol-interference. The effect of equalization is considered as a technique to reduce inter-symbol-interference (ISI). The potential benefits of equalisation are discussed and some results presented which indicate that Turbo Coding greatly reduces sensitivity to frequency selective fading and hence shows that equalisation is not necessarily required. The channel interleaving is known to produce significant improvements by utilizing time diversity to mitigate the effects of fading. Turbo coding has an interleaver built in to its structure which provides good under fading conditions.

1. INTRODUCTION

Mobile satellite services are globally available for low rate applications such as voice, fax and 2.4 kbit/s data. In addition, some services provide in-flight communications of a similar nature. Extension of these services to higher rate applications is constrained by the high cost of flight qualified hardware, particularly antennae with suitable gain. However, the business traveler of the future is likely to wish to be able to access the office network or Integrated Service Distributed Network (ISDN), even at a premium price. Hours spent on aircraft could be spent in video-conference or working with other interactive multimedia applications. Recently, key technology developments are making this scenario more practical.

Primarily, the development of more powerful satellites with spot beams for high EIRP and G/T makes communications with medium gain antennae on board aircraft feasible from the point of view of received signal power. However, achieving high transmission rates is hindered by the effects of multipath reflections. For aircraft, reflections from the ground can give rise to differential delay spread, τ, of the order of tens of microseconds with significant magnitude compared with that of the direct line of sight signal (for example refer to [1] for measurements at L Band).

In the case of ISDN data rates (144 kbit/s) the required transmission rate might be approximately 150 ksym/s assuming rate 1/2 coded QPSK signaling, with a symbol period of 6.7 μs. Clearly, the delay spread could be very significant giving rise to frequency selective fading. The resulting degradation from the effects of ISI depends on the antenna gain, satellite elevation angle and aircraft altitude. It is an important affect to consider when designing the system.

Turbo codes are being considered for mobile satellite applications due to their great power efficiency and adaptability to bandwidth efficient schemes [2]. Turbo codes contain an interleaver and are therefore inherently well suited to operate efficiently in fading channel conditions. In this paper a turbo coded system is described for a range of high rate data aeronautical environments. Simulation results based on the model are presented in Section 3 to examine bit error performance of the turbo codes. Also, the sensitivity of a typical receiver is considered with and without an equaliser for comparison.

Section 4 presents the conclusions of this work, with discussion of future directions.

2. TURBO CODED SYSTEM

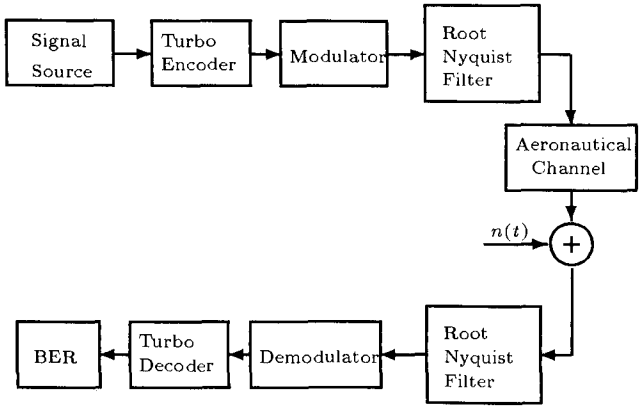


Figure 1: Block diagram of Turbo Coded system.

For the purpose of this investigation, a simplified baseband model is considered as shown in Figure 1. This is used to investigate the performance of turbo coding over aeronautical channel environments. The signal source generates a random binary stream and form in to blocks. The length of these blocks is equal to the interleaver size. The blocks are then rate 1/2 turbo encoded and modulated. QPSK and 16QAM modulation sc hemes were considered in this work. The modulated signal is then filtered with a root-Nyquist filter with roll-off factor of 0.25 and passed to the channel with two different impairm ents namely aeronautical fading and additive white Gaussian noise (AW GN) with mean zero and variance$\sigma^2 = N_0/2$.

The faded signal is passed through a root-Nyquist receive filter assuming ideal synchronisation and no equalisation with known diffuse Rayleigh faded component α at the receiver. The demodulator assumes ideal frequency, phase, amplitude and timing synchronisation. The turbo decoder outputs soft decision and the BER block calculates the bit error rate base on the decoded output. Details of the turbo codec and aeronautical channel model follow.

2.1 Turbo Codec

The concept of turbo-code was first proposed by Berrou *et al.* [3] as a way of dramatically reducing the errors in a forward error correction system. Turbo codes results from the concatenation of two Recursive Systematic Convolutional (RSC) codes. The turbo coded system discussed here uses a rate 1/2 turbo-code encoder which consists of two parallel concatenated RSC encoders with constraint length $K = 5$, memory $M = K - 1 = 4$ and the same encoder generators $G_1 = 37_8$ and $G_2 = 21_8$ as shown in Figure 2. Note that the first encoder receive the information bit vector $\underline{b} = [b_1, b_2, \cdots, b_N] \in \{0, 1\}$ and the second encoder receive randomly interleaved version of the same information bit v ector $\underline{b}$. Note that the N independent information bits in the information bit v ector has equal probability. The encoder initial state S_0 for both the encoders set to 0. For an input bit b_k at bit interval k the turbo-code encoder outputs $s_k = b_k$ and the punctured redundant information bit $c_k \in \{0, 1\}$ which consists of alternate samples of $c_{1,k}$ and $c_{2,k}$. The encoder outputs d_k and c_k are then modulated and transmitted over the channel.

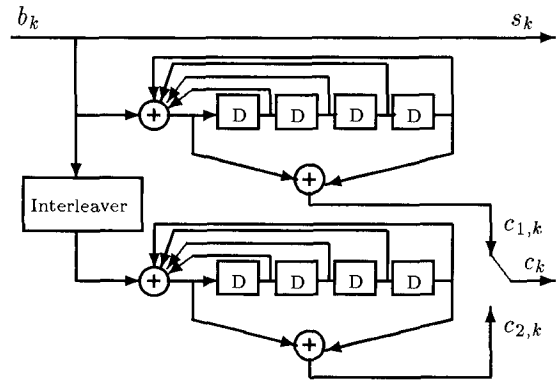


Figure 2: Rate 1/2 Turbo-code Encoder with two parallel concatenated Recursive Systematic Convolutional codes at bit interval k.

The turbo decoder at the receiver uses 4 iterations of the MAP algorithm b y Bahl *et al.* [4] to decode the received demodulated signal. The decoder can receive the normalised bit log-likelihood ratios (LLR) or soft decisions for both I and Q channels and variance estimates. The demodulator block makes ideal variance estimates for the turbo decoder.

2.2 Aeronautical Channel Model

The aeronautical channel implement s a two-path fading channel consisting of a line of sight component and a diffuse Rayleigh faded component based on reflected waves from the ground. The block diagram of the channel is shown in Figure 3.

Three important parameters in the channel are the fading bandwidth, f_d, carrier to multi-path ratio, C/M, and the multi-path delay, τ where, f_d for second order But-

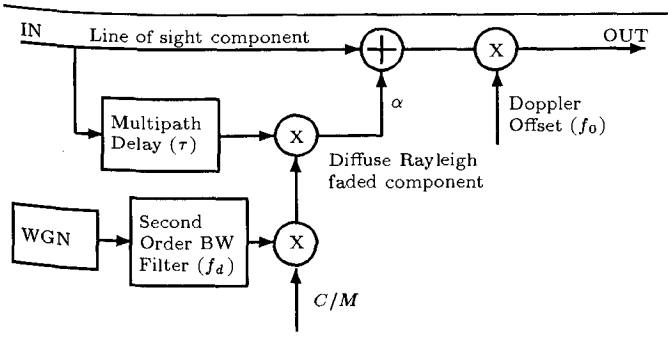


Figure 3: Block diagram of the Aeronautical Channel Model.

terworth (BW) filter is calculated as follo ws:

$$f_d = \frac{f_r}{0.624} \quad (1)$$

where, f_r is the fade rate. These main three parameters can be used in the aeronautical channel to create two different channel effects known as - GOOD and BAD as shown in Table 1. The channel also includes the effect of Doppler offset f_0. Note that the multi-path delay, τ, results in frequency selective fading when it is greater than 0 μs. From Table 1, τ of 12.5 μs and 10 μs are equivalent to approximately 1 and 1.5 symbols for a symbol rate of 100 kHz. Therefore, ISI will degrade the performance significantly especially in the BAD channel condition. An equaliser may therefore be required to improve the performance especially in the BAD channel condition.

Table 1: Aeronautical Channel Parameter's Values.

Channel Effects	τ (μs)	f_d (Hz)	C/M (dB)
GOOD	12.5	50	12
BAD	10	50	7

The error performance can be improved under fading conditions by using a channel interleaver. In this work the effect of a pseudo random interleaver was studied for three different frame lengths of 50 ms, 100 ms and 250 ms.

2.3 Performance Issues

The performance of turbo codes is often effected by delay and complexity associated with it. The delay results from the use of an interleaver embedded in the turbo code. Both the encoding and decoding process have to wait until the complete content of the interleaver gets filled. In practice decoding can start as soon as data arrives but cannot complete a single iteration until the whole block is available. Delay also occurs in the actual decoding process. Assuming the decoder processing can operate at the bit rate, the total processing and buffering delay is taken to be $\frac{3\times N}{\text{Bit Rate}}$ as a typical value.

Complexity of the turbo decoder depends on the number of iterations, the number of states in the constituent codes and the algorithm used. Therefore, performance can be traded-off against system performance. For small interleavers (< 1000 bits), four iterations at the decoder can give most of the gain whereas for larger interleavers more iterations are required. Delay and performance issues are discussed in [2, 5] in detail.

Performance of turbo code can also be limited due to bad interleaver design or truncation in the decoding algorithm. If the sequences are highly correlated, the bit error rate (BER) performance decreases to a certain level from where there is no further improvement in the decoding process. This effect is well known as the "error floor" effect of the turbo code BER curve.

3. SIMULATION RESULTS

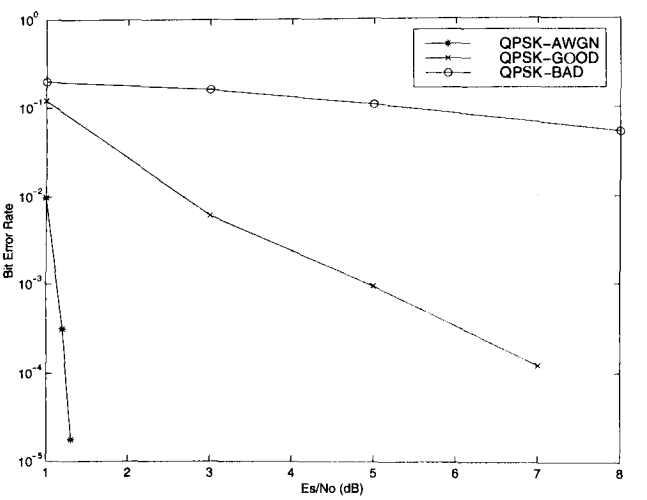


Figure 4: QPSK codec performance in terms of BER versus E_s/N_0 in GOOD and BAD channel conditions assuming ideal synchronisation and no equalisation.

The model developed in Section 2 was simulated to generate BER performance figures as a function of E_s/N_0. Two sets of results are shown; the first is the coded error performance and the second is the uncoded losses due to phase synchronisation under fading conditions. This paper includes only QPSK codec performance. Performance of different frame sizes and 16QAM modulation will be presented in the conference.

3.1 Turbo Coding Performance

Figure 4 shows the QPSK codec performance in AW GN, GOOD and BAD channel conditions assuming ideal synchronisation and no equalisation. With 4 iterations QPSK codec achieves a BER of 10^{-4} at an E_s/N_0 value of 1.23 dB. Note that the E_s/N_0 value used as the abscissa is the ratio of energy per symbol to noise density ratio for a line of sight signal, ie., no fading. It can be seen that over GOOD channel performance loss is approximately 5.77 dB, however, performance degrades significantly over BAD channel. Other interleaver sizes were considered in the study but were not available to be included in this paper.

3.2 Synchronisation and Equalisation Performance

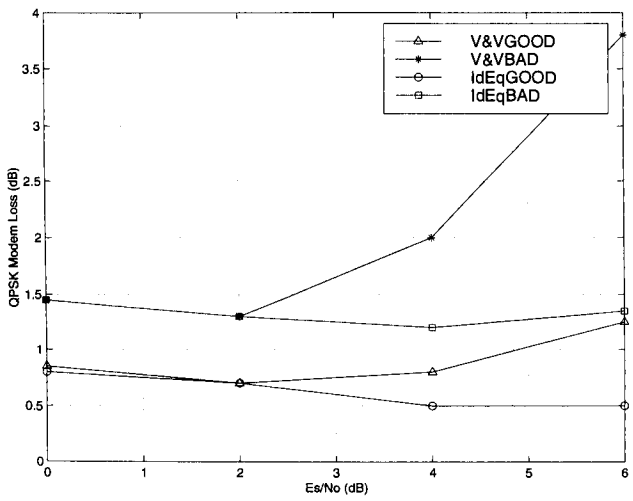


Figure 5: Uncoded QPSK modem loss with an ideal equaliser and V&V phase estimation in GOOD and BAD channel conditions.

Figure 5 shows the performance of rate 1/2 turbo coded system with two different approaches to channel estimation and correction in two different channel conditions. Note that the symbol rate, R_s, is set to 150 kHz. V&VGOOD and V&VBAD denote the modem loss with Viterbi and Viterbi (V&V) QPSK phase estimator [6] in GOOD and BAD channel conditions respectively whereas IdEqGOOD and IdEqBAD denote the modem loss with an ideal equaliser in GOOD and BAD channel conditions respectively. An ideal equaliser means perfect knowledge of channel phase, amplitude and delay, which is used to set the taps of an adaptive infinite impulse response filter. The QPSK modem loss shown in Figure 5 is calculated by measuring the BER of the uncoded QPSK signal and calculating the power difference required to meet the same error rate under ideal modem and AW GN assumptions. The most interesting observation is that at low E_s/N_0, there is no significant benefit in using an equaliser over standard V&V phase correction. This is important because this the operating region for the turbo codes considered to achieve BER of 10^{-6} or lower which might be expected for data services.

It can be seen that ideal equaliser performs better at higher E_s/N_0 values, ie., approximately 0.25 dB better over GOOD channel and 0.2 dB better over BAD channel at an E_s/N_0 value of 6 dB as compared to the E_s/N_0 value of 0 dB. The effectiveness of the equaliser depends on the operating point.

4. CONCLUSION

Turbo coding is the most efficient known error control coding scheme under certain conditions and so was considered in this work. The codes have been simulated to show how excellent performance can be obtained with moderate delay impact for high rate data.

An important result is that there is little, if any, benefit from using an equaliser at the operating point for the turbo code. In practice, the difficulty in training an equaliser at low signal to noise ratio (SNR) could make it impractical anyway. At higher SNR, the equaliser may prove beneficial. This may be true for higher order modulation schemes or high rate codes. Further work is required to determine this.

Although results were not available to include in this paper, the work has shown that significant improvements can be obtained for increased turbo code interleaver sizes at the cost of increased processing delay. This is largely due to the improved diversity gains from a large interleaver; the static performance only improves marginally for larger interleavers.

In summary, the application of turbo codes to aeronautical satellite transmission can produce a highly power efficient and practical scheme capable of meeting low BER targets.

References

[1] J. Lodge, "Modulation for Aeronautical Mobile Channels," tech. rep., FANS WG/B Report, Nov. 1986.

[2] S. A. Barbulescu, W. Farrell, P. Gray, and M. Rice, "Bandwidth Efficient Turbo Coding for High Speed Mobile Satellite Communications," in *Proc. of the International Symposium on Turbo Codes & Related Topics*, (Brest, France), pp. 119–126, ESNT de Bretagne, Sept. 1997.

[3] C. Berrou, A. Glavieux, and P. Thitmajshima, "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-Codes," in *IEEE Int. Conf. on Communications*, (Geneva, Switzerland), pp. 1064–1070, May 1993.

[4] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, pp. 284–287, March 1974.

[5] M. Rice, P. Gray, and S. A. Barbulescu, "Coding and Modulation Techniques for High Speed Data Services by Satellite," in *Proc. of 1997 Int. conf. on Information, Communications and Signal Processing*, (Singapore), Sept. 1997.

[6] A. J. Viterbi and A. M. Viterbi, "Nonlinear Estimation of PSK Modulated Carrier Phase with Applicaiton to Burst Digital Transmission," *IEEE Trans. Inform. Theory*, vol. IT, pp. 543–551, July 1983.

Hyper-codes for TCP/IP over Mobile Satellites

R.W. Kerr and M. Moher
Communication Research Centre
3701 Carling Ave., P.O. Box 11490 Station H, Ottawa, ON, K2H 8S2, Canada
Email: ron.kerr@crc.ca

ABSTRACT

In this paper, we compare and contrast the design considerations when using Turbo-like codes versus traditional coding techniques for transmitting Internet traffic using TCP/IP over a mobile satellite link. This paper presents the throughput efficiency for various file sizes over a geostationary link using a commonly available TCP/IP protocol stack. We briefly discuss some proposed changes to TCP/IP and their effect on throughput for our application.

INTRODUCTION

The traditional coding techniques considered here for comparison are convolutional codes. The Turbo-like codes considered are Hyper-codes [1,2]. They are a family of block codes and are made up by multiple simple parity equations. The iterative decoder for these codes takes advantage of the low complexity associated with soft-in, soft-out decoders for simple parity equations. They also have the advantage that high rate versions of these codes have very good performance.

The performance of convolutional codes with soft-decision Viterbi decoding is used in this evaluation. The fundamental difference between the performance of these two coding approaches is the dependence of the packet error rate on packet length. For convolutional codes the packet error rate exhibits approximately linear growth with code length, this is not the case for Hyper-codes. In several cases, it is possible to find Hyper-codes that have lower PER for longer block lengths. This factor has significant effects on system design and throughput.

In the following sections, we discuss the trade-offs associated with the use of convolutional and Hyper-codes for TCP/IP over a mobile satellite link.

TCP/IP is a connection-oriented reliable communication protocol. TCP/IP is very effective on terrestrial wired networks. There are several extensions to the protocol under development, which will improve the throughput over satellite and wireless links. However, at this time they are not widely distributed. For this reason, we will assume that the TCP/IP protocol used is the one that is in common use. We will use a Go-Back-N protocol for ARQ and assume for the purposes of analysis that each packet requires the typical TCP/IP 40 byte header.

Other system parameters used in the analysis are as follows: a return channel transmission rate is 32 kbps, and one-way propagation delay over the satellite of 250 ms. We will use message or file sizes for transmission of 64 bytes, 1 kbyte, 8 kbyte and 64 kbyte to test the throughput efficiency for various packet sizes.

CODE COMPARISONS

In the first part of our paper, we provide simulation results of the PER performance of the Hyper-codes and the convolutional codes on the additive white Gaussian channel. In the following, we will refer to the Hyper-codes with the prefix 'hc' followed by the number of information bits that are contained in each codeword .

The hc512 Hyper-code is a rate 0.632 code and has a block size of 512 information bits (64 bytes). Figure 1 shows the PER performance of the hc512 code and the punctured rate 2/3 K=9 convolutional code using tail-biting soft-decision Viterbi decoding with a block size of 512. The rate 2/3 K=9 convolutional code was obtained from the K=9 rate 1/2 convolutional code with taps (561,753) in octal notation with a puncture mask of (11,10) [4]. The hc512 code uses iterative decoding and the packet error rate (PER) results are shown in Figure 1 are for a maximum of 8 iterations. Early termination was used in the iterative decoder. Early termination allows the decoder to stop decoding if the decoder had converged to a valid codeword. The hc512 code at a PER of 10^{-3} performs 1.3 dB better than the punctured rate 2/3 convolutional code.

In Figure 2, the PER performances of the hc8000 Hyper-code and the punctured rate 4/5 K=9 convolutional code are shown. The hc8000 has a block size of 8000 bits and a code rate of 0.825. The convolutional code used the rate 1/2 convolutional code with taps (561,753) in octal notation and a puncture mask of (1101,1010) [4]. The decoder used tail-biting soft-decision Viterbi decoding and the block size was set to 8000 bits. The Hyper-code decoder used a maximum of 8 iterations and early termination was used. At a PER of 10^{-3} the hc8000 outperforms the punctured rate 4/5 K=9 convolutional code by 2.2 dB.

In Figures 1 and 2, we see performance of the convolutional code degrade by 1.4 dB at the PER of 10^{-3},

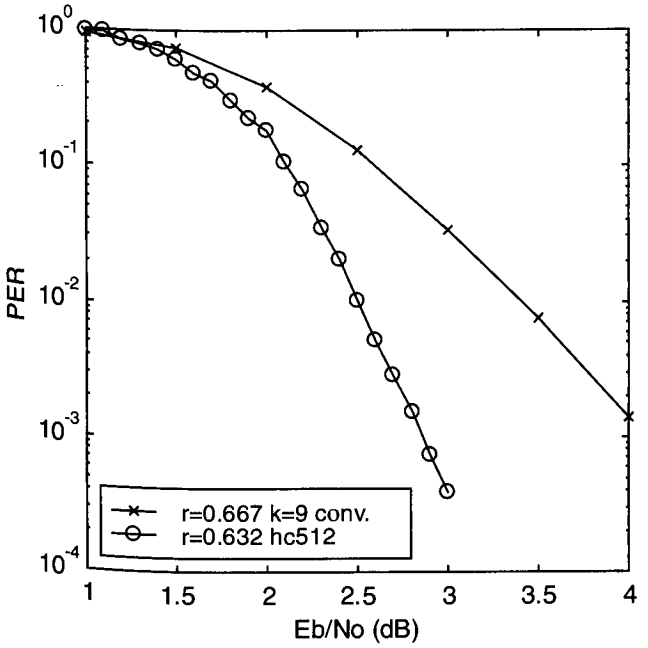


Figure 1 PER performance on the AWGN channel with BPSK modulation for a punctured rate 2/3 k=9 tail-biting convolutional code with packet size of 512 information bits and the rate 0.632 hc512 Hyper-code.

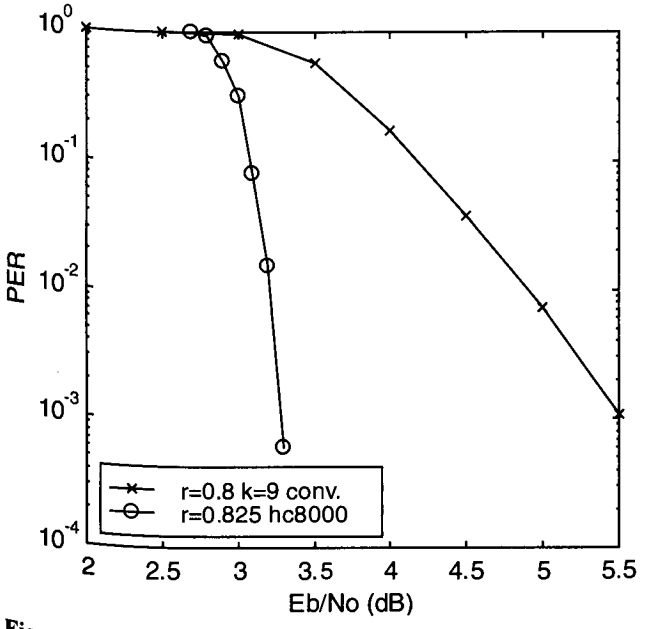


Figure 2 PER performance on the AWGN channel with BPSK modulation for the punctured rate 4/5 k=9 tail-biting convolutional code with packet size of 8000 bits and the rate 0.825 hc8000 Hyper-code.

when the block length is increased from 512 to 8000 bits and the code rate is increased from 2/3 to 4/5. The Hyper-code performance only degraded by 0.4 dB under the same conditions. From the family of Hyper-codes, it is possible to increase the block length and code rate of many of the Hyper-codes with relatively little impact on the BER and PER [1] . Also, fixing the code rate and increasing the packet length increases the power of the code, so it possible to achieve better PER results with increased packet length [2]. This is opposite to the results that can be obtained with convolutional codes.

The tail-biting K= 9 rate 1/2 convolutional code with a block length of 1000 bits, a PER of 10^{-3} can be achieved using QPSK modulation at an E_b/N_0 of 3.6 dB [2]. For Hyper-codes the same block length requires approximately 2 dB [2]. This is a gain of approximately 1.6 dB over the K=9 rate 1/2 convolutional code for the same PER. For code rates ranging from 0.63 to 0.79, the required E_b/N_0 for Hyper-codes to achieve a PER of 10^{-3} is approximately 3.0 dB using QPSK modulation. Thus, Hyper-codes can operate at a higher information rate with a lower E_b/N_0 than required for a K=9 rate 1/2 convolutional code.

THROUGHPUT COMPARISONS

If we consider using a conventional convolutional code then we must carry out the trade-off between packet size and the packet error rate. For a convolutional code, error events are typically short and independent. It follows that increasing the packet length also increases the packet error rate. In this section, we examine the throughput of the channel when a tail-biting convolutional code is used. A tail-biting convolutional code has no overhead associated with sending bits to 'flush' the encoder at the end of packet.

A packet error occurs when a packet has at least one bit error. The PER is at various packet lengths is estimated by

$$p = 1 - (1 - p_b / c)^N \qquad (1)$$

where p_b is the bit error rate, c is a constant dependent on the constraint length and code rate of the code, and N is the packet length. The values for c was determined from the simulations of the punctured k=9 convolutional codes. In Figure 3, the estimated packet error rate versus packet length for the punctured rate 2/3 k=9 convolutional code is plotted with bit error rates of 10^{-5}, 10^{-6} and 10^{-7} is plotted.

A Go-Back-N ARQ strategy is considered here. In this method, a transmitter is required to repeat any packet transmitted since the last acknowledged packet if a negative acknowledgement is received or a timeout occurs. The number of packets retransmitted depends on the transmission rate, propagation delay of the channel and the packet size. The expected number of packet transmissions (N_t) per packet for the Go-Back-N algorithm is given by [3],

$$\mathrm{E}[N_t] = \frac{1 + 2ap}{1 - p}, \qquad (2)$$

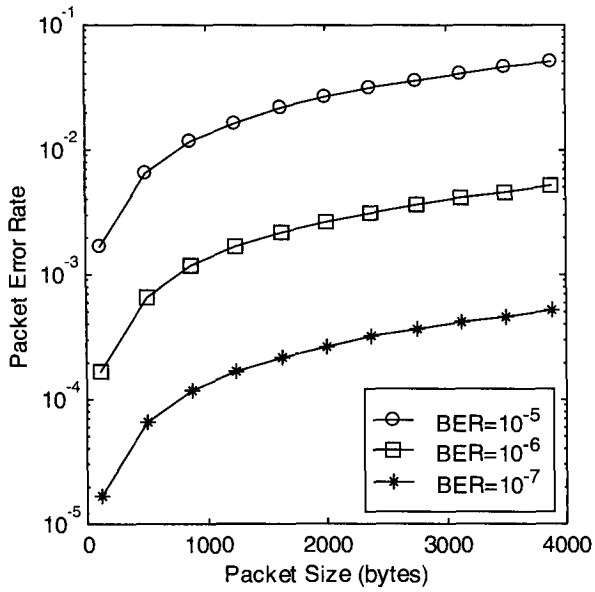


Figure 3 Estimated Packet Error rate versus Packet Length for punctured rate 2/3 k=9 convolutional code

where a= (propagation delay * bit rate)/ (packet size), p is the probability of packet error defined in Eq. 1 and the window size for flow control is greater than the $2a+1$.

An assumption was made in the derivation of Eq. 2 that the retransmitted packets and the acknowledgements are error-free. In a real system, the effect of such errors should have little impact on the performance of the system.

The number of packets of size N bits required to transmit a file of size M bits when each packet has an overhead of H bit is given by

$$N_p = \left\lceil \frac{M}{N-H} \right\rceil, \qquad (3)$$

where $\lceil \bullet \rceil$ is the ceiling function.

The expected number of transmissions for a file is then obtained by combining Eq. 2 and 3 to give

$$N_r = \frac{(1+2ap)N_p}{1-p}. \qquad (4)$$

To provide a fair comparison of the packet sizes, we define the throughput as the number of information bits divided by the total number of transmitted bits (including overhead), that is,

$$T = \frac{M}{N_r N}, \qquad (5)$$

where T is throughput, N_r is the expected number of packet transmissions per file, N is the size of the packet and M is the file size.

The throughput results for the punctured rate 2/3 k=9 convolutional code is plotted in Figure 4 for various file sizes with the packet sizes ranging from 100 to 4100. The

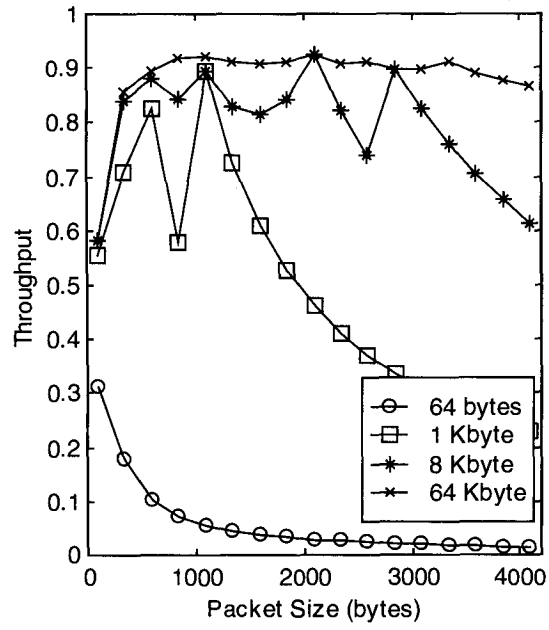


Figure 4 Expected Throughput (info bits/total transmitted bits) for various file sizes (64 bytes to 64 kbytes) and packet sizes. Punctured rate 2/3 k=9 convolutional code with a BER = 10^{-5}.

BER is set to 10^{-5} and the header is 40 bytes. The PER was calculated using Eq. 1. From the figure, the throughput is low for the 64 byte message. A 300 byte packet obtains a 0.73, 0.83, and 0.85 throughput for the 1, 8 and 64 kilobyte files, respectively. The throughput for the 64 kbyte file gradually drops as the packet size increases, supporting the claim that smaller packet sizes have better throughput than larger packet sizes for convolutional codes.

Many of the variations seen in throughput plots are due to the fact that the example file sizes do not divide exactly into the given packet size. As a result the last packet transmitted may contain little information. This lowers the throughput for the specific file size considerably. Also, the throughput lowers when the packet size is greater than the file size

For some Hyper-codes it is possible to increase the code rate and length of the packet at a fixed E_b/N_0 with very little impact on the PER [1]. In order to evaluate the throughput of the system using Hyper-codes, we set the packet error rate to a constant regardless of packet size. The expected throughput results for PERs of 10^{-2} and 10^{-3} are shown in Figures 5 and 6, respectively. The system has the same assumptions used in the previous section. (i.e., 40 byte header per packet, 32 kbps transmission rate, Go-Back-N ARQ.)

From Figures 5 and 6, we see that packets around 1000 bytes are efficient for the 1, 8, and 64 kilobyte files. The throughput for these file sizes is 0.9 or above for when the PER is 10^{-2} and 10^{-3}.

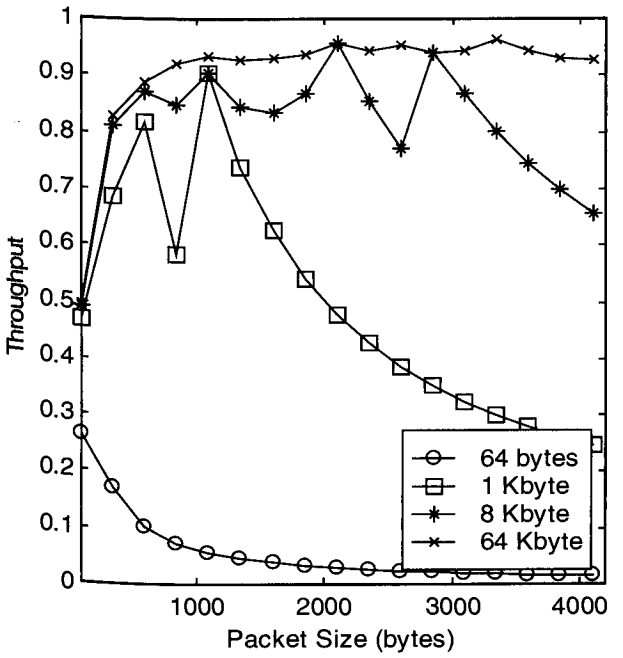


Figure 5 Expected throughput (info bits/ transmitted bits) for various file sizes. Hyper-codes with a PER = 10^{-2}.

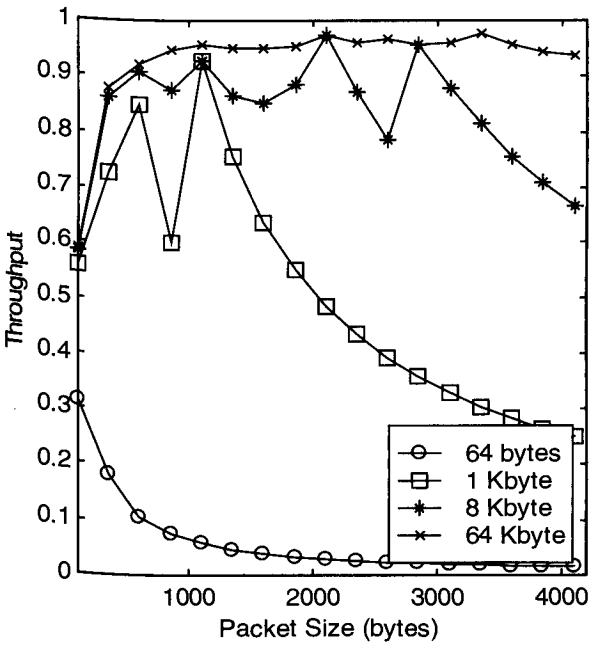


Figure 6 Expected throughput (info bits/ transmitted bits) for various file sizes. Hyper-codes with PER = 10^{-3}.

The best choice of packet size will depend on the distribution of the actual file sizes. In some cases, multiple packet sizes may be preferable. The throughput plots for Hyper-codes and convolutional codes are similar. As shown in Figures 1 and 2, Hyper-codes can achieve a better PER than a similar rate convolutional code for a given E_b/N_0. A difference that will affect throughput, is the sensitivity of the codes to the channel BER. The Hyper-codes can be sensitive to the channel BER where a small degradation in the channel can cause a large degradation in the PER (e.g., Fig. 2). The convolutional codes tend to degrade more gracefully when subjected to increasing channel BER.

TCP/IP EXTENSIONS

Selective Repeat

A proposed modification to the TCP protocol is to include selective acknowledgements (SACK) rather than the current acknowledgement scheme [5]. With SACKs only the packets received in error are retransmitted. This increases the complexity of the receiver, as it is required to store correctly received packets (after the packet in error) and perform reordering when the retransmitted packet arrives correctly. In the Go-Back-N protocol, the receiver does not need to buffer correctly received packets after an error, as the transmitter was required to retransmit all packets after a packet error occurred.

With a Selective Repeat strategy, the expected number of packet transmissions per packet is [3]:

$$\mathrm{E}[N_t] = \frac{1}{1-p}. \qquad (6)$$

The throughput improvement expected is then obtained as

$$T_g = 1 + 2ap.$$

For a PER of 10^{-2}, this entails an improvement in throughput of 31%, 16%, and 2%, for 64, 1000, and 8000 byte packets over Go-Back-N. For a PER of 10^{-3}, the improvement is limited to 3%, 1% and 0.2% for the 64, 1000 and 8000 byte packets. For a PER of 10^{-2}, the change in ARQ strategy is useful, but the impact on the throughput at a PER of 10^{-3} is not significant. A selective-repeat strategy will be useful, if the channel degrades past the expected operating point.

TCP/IP Header Compression

In [6], a method for compressing IPv6 headers was proposed that takes advantage of the fact that most fields in the TCP/IP headers do not change between consecutive packets from the same data stream. When possible, the full header is replaced with a compressed header, which can contain a connection identifier, and any fields that change regularly between packets. This adds some complexity into the protocol as the receiver must detect an inconsistent compression state and there must be some techniques to repair or update the compression state. However, it is possible to reduce the IPv6 header from 60 bytes to an average of 5 bytes [6]. We will make an assumption that we can achieve the same compression

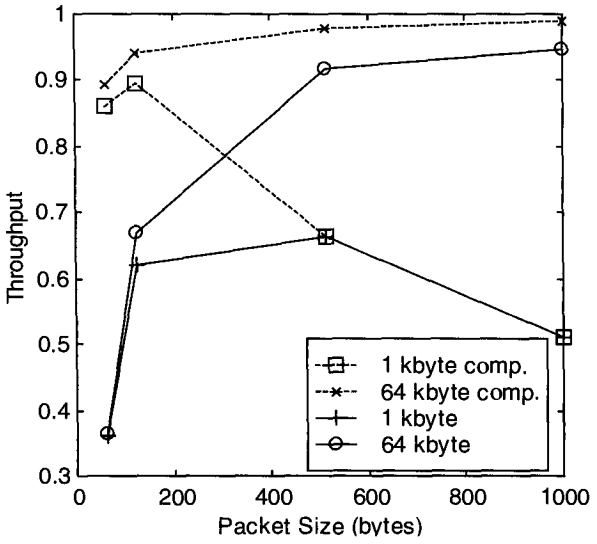


Figure 7 Throughput for 1 and 64 kbyte files with and without header compression. PER = 10^{-3}.

with an IPv4 header of 40 bytes. To determine if a compressed header will increase throughput for our application, let us make the following assumptions: a full header is 40 bytes, and a compressed header is on average of 5 bytes. In Figure 7, the throughput for transmitting a 1 and 64 kbyte file is shown with and without header compression. When compression is used there is one full header transmitted and the remaining packets contain a compressed header. The number of packets to transmit the file is reduced, as there is more data per packet due to the reduced headers.

The impact on the throughput is significant for small packets where the overhead of 40 bytes is a significant portion of the packet. For example, for the 1 kbyte file transmitted with 64 byte packets, the throughput improved from 0.35 to 0.85. For the longer packets and messages where the header is not a significant portion of the packet there was only a small improvement in throughput. For example, for the 64 kbyte file using 1000 byte packets, header compression improved the performance from 0.95 to 0.97.

The throughput is increased using header compression. Here, it was assumed that there were mechanisms in place to detect and correct for corrupted headers. The actual algorithm would require further testing at the operating PER. It does appear from the initial investigation that a header compression algorithm, which could tolerate error-rates common to the satellite channel, looks promising to reduce the overhead associated with TCP/IP.

CONCLUSIONS

We have shown that for our application Hyper-codes have a better PER performance than punctured K=9 convolutional of similar rates. The key difference is that, in general, Hyper-codes with longer block sizes perform better than the Hyper-codes with short blocks in terms of BER and PER.

The throughput performance for Hyper-codes and convolutional codes when using a Go-Back-N strategy and the 40 bytes headers used in TCP/IPv4 was tested . The punctured rate 2/3 convolutional code with a constant BER of 10^{-5} performed worse than the Hyper-codes at longer packet sizes. The better PER performance of the Hyper-codes over the convolutional improves the channel throughput by not requiring as many retransmissions.

Proposed techniques to improve the performance of TCP/IP over satellite were briefly investigated for our application. For our return link data rate of 32 kbps we found that selective repeat ARQ, although beneficial at a PER of 10^{-2}, did not, for the packet sizes of interest give significant improvement over the Go-Back-N ARQ for a PER of 10^{-3}. Selected-repeat ARQ will be beneficial to the when the channel degrades past the Hyper-codes operating point.

TCP/IP header compression [6] was shown to have an improved throughput for some of the message sizes considered. The improvement is due to shortening the header and using the extra bits for data, thus reducing the number of packets required for transmission. The use of header compression requires modification of the protocol in order to ensure that the compressor and decompressor at either end of the link are synchronized.

REFERENCES

[1] A. Hunt, S. Crozier, and D. Falconer, "Hyper-codes: High-performance low-complexity error-correcting codes," Proceedings of the 19th Biennial Symposium on Communications, May 31 to June 3, 1998, Kingston, On. pp. 263-270.

[2] S. Crozier, A. Hunt, K. Gracie, and J. Lodge, "Performance and Complexity Comparison of Block Turbo-Codes, Hyper-codes and Tail Biting Convolutional Codes," Proceedings of the 19th Biennial Symposium on Communications, May 31 to June 3, 1998, Kingston, On., pp. 84-88.

[3] W. Stallings, Data and Computer Communications, Macmillan Publishing, New York, 1985.

[4] Y. Yasuda, K. Kashiki and Y. Hirata, "High-Rate Punctured Convolutional Codes for Soft Decision Viterbi Decoding", IEEE Trans. on Commun., Vol. COM-32, NO.3, p.315-319, March, 1984.

[5] M. Mathis, J. Mahdavi, S. Floyd, and A. Romanov, "TCP Selective Acknowledgement Options," Oct. 1996, RFC 2018.

[6] M. Degermark, M. Engan, B. Nordgren, and S. Pink, "Low-loss TCP/IP Header Compression for Wireless Networks," Proceedings of Mobicom'96, Rye, New York, Nov. 10-12, 1996.

High-Speed DSP Implementations of Viterbi Decoders

Pierre-Paul Sauvé, Stewart Crozier and Andrew Hunt
Communications Research Centre
3701 Carling Ave., P.O. Box 11490, Station H
Ottawa, Ontario, Canada, K2H 8S2
Email: pierre-paul.sauve@crc.ca

ABSTRACT

This paper describes a number of techniques useful in implementing very efficient Viterbi decoders on general-purpose digital signal processors (DSPs). Also described is the application of these techniques to the design of a fast decoder for the ADSP-2106x (SHARC) DSP with an efficiency approaching 2 processor cycles per state for constraint lengths above K=7. Options include rates 1/2 and 1/3, as well as both flushed and tail-biting blocks.

BACKGROUND

This section contains a brief summary of convolutional encoding, and Viterbi decoding essentials.

Convolutional Encoding

Figure 1 shows the industry standard rate 1/2, constraint length K=7, binary convolutional encoder. It is made from a shift register with K-1=6 memory elements. For every bit input, $d(i)$, two coded bits, $c_1(i)$ and $c_2(i)$, are generated. The encoder is a state machine whose state is $[d(i\text{-}(K\text{-}2)),\ldots,d(i\text{-}1),d(i)]$ at time i.

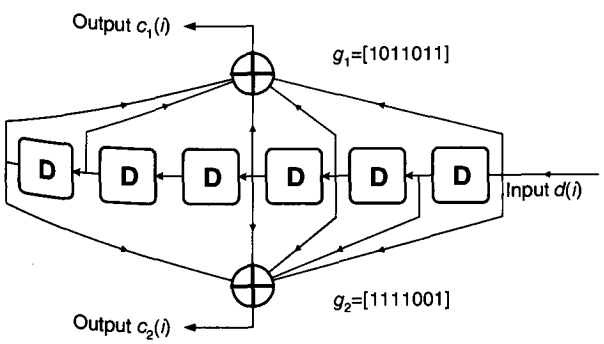


Figure 1: Block diagram of a constraint length K=7, rate 1/2, binary convolutional code.

The encoder is characterized by the generator polynomials g_1 and g_2, which specify how delayed versions of the input are added up (modulo-2) to produce each output. Each binary digit of the generator polynomial corresponds to a tap on the shift register, where the least significant bit represents the tap for the input delayed by K-1 time-steps.

The code bits $c_1(i)$ and $c_2(i)$, taking values in {0,1}, are converted to channel symbols s_1(i) and $s_2(i)$, taking values in {+1,-1} for transmission purposes. For simulation purposes, these channel symbols are transmitted over an additive white gaussian noise (AWGN) channel using binary antipodal signalling. The outputs of the matched filter at the receiver are $r_1(i)$ and $r_2(i)$.

Viterbi Decoding

The Viterbi Algorithm (VA) performs maximum likelihood sequence estimation (MLSE) by finding the data sequence **d'** that minimizes the Euclidean distance between the corresponding candidate channel symbols **s** and the received noisy samples, **r**. For binary antipodal signaling, which is assumed henceforth, this is equivalent to maximizing the correlation between **s** and **r**.

The VA must keep track of *state metrics,* which are the correlation between the received noisy samples and a particular set of channel symbols corresponding to a surviving path, and *state histories,* which contain the data sequence corresponding to a surviving path. Associated with each state n, at time i, is a state metric, $M_n(i)$, and a state history vector, $H_n(i)$. For each valid transition from state m to state n, there is an associated branch metric, $B_{mn}(i)$ which, for a rate 1/2 code, corresponds to the correlation between $\{s_1(i), s_2(i)\}$ and $\{r_1(i), r_2(i)\}$ for the transition. The branch metrics, state metrics and state history vectors are updated as follows:

$$B_{mn}(i) = s_{1mn} \cdot r_1(i) + s_{2mn} \cdot r_2(i) \quad (1)$$

$$M_n(i) = \max_m [M_m(i-1) + B_{mn}(i)] \quad (2)$$

$$H_n(i) = [H_m(i-1)\ ,\ \mathrm{LSB}(n)], \quad m = \text{best old state} \quad (3)$$

We will generally refer to Equation (2) as the add-compare-select (ACS) operation, and the application of (1)-(3) for all states for one time step as a *Viterbi pass*. Integer arithmetic is assumed in the following discussion.

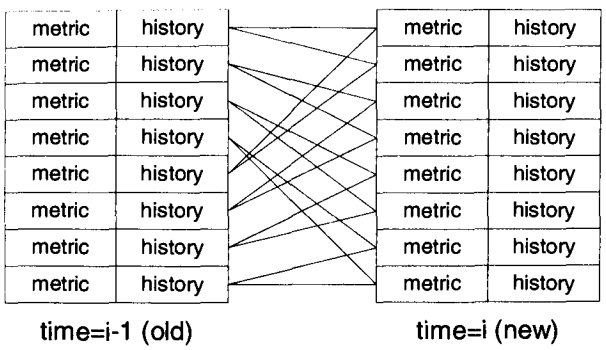


Figure 2: Viterbi pass for a rate 1/n, 8 state code, K=4. All state metrics and histories must be updated at every pass.

Typical Data Structure for Decoding

Figure 2 represents the application of one Viterbi pass. Typical software implementations of the VA require two buffers, each able to hold a full set of state metrics and histories: one for the previous, or "old" data, and one for the updated, or "new" data. Once the new data is obtained, the buffer that contains it is used again, but as the old buffer for a new Viterbi pass.

Normalization of State Metrics and Input Dynamic Range

During the course of decoding, state metrics are subjected to addition of branch metrics and selection. Since state metrics are stored in words with a finite number of bits, they can overflow after a number of Viterbi decoding passes, aliasing large positive numbers to large negative ones (assuming twos-complement arithmetic). This may cause the selection of the best paths to fail. The simplest method of avoiding overflow for short blocks is either to restrict the range of the input samples **r**, or to increase the number of bits available for growth of the metrics. One may also *normalize* the state metrics intermittently to keep their values in check, typically by subtracting the maximum state metric from all state metrics.

To determine the safe range of input samples and the required normalization frequency, we must look at the behaviour of the state metrics with respect to time. There are two components to the evolution of the state metrics: their spread, and their rate of increase. The spread (maximum minus minimum value) of the metrics is restricted to be at most $(K-1)\upsilon$ [1], where υ is the maximum spread of the branch metrics. The maximum rate at which the maximum state metric can increase is $\upsilon/2$ per Viterbi pass. A normalization where the maximum state metric is moved to zero requires the spread to be less than half the total range of the state metrics. Otherwise, normalization would result in an underflow. It is also necessary to ensure that the maximum growth rate of the state metrics does not cause the maximum state metric to overflow within the block length or normalization interval. These requirements can be restated as follows:

$$\log_2((K-1)\upsilon) \le W-1, \text{ and} \qquad (4)$$

$$\log_2(n\upsilon/2) \le W-1, \qquad (5)$$

where W is the number of bits available for the storage of state metrics and n is the normalization interval, in Viterbi passes. Note that in (5), a normalization interval of $n \le 2(K-1)$ is no more constraining than (4).

History Truncation

For true MLSE decoding, the starting and ending states must be known, and the decoder must keep state histories that are as long as the information sequence. With long information sequences, decoding delays and state history storage requirements become impractical.

One may truncate the state histories by storing only a finite number of bits, h, the *history depth*. This depth is chosen so that if one finds the best state metric at time j, the bit corresponding to that path's history at time j-h has a high probability of belonging to the most likely sequence. In practice, a history depth of $h=5K$ [1],[2] is sufficient to ensure performance close to true MLSE decoding for a rate 1/2 code.

Block Termination Methods

There are two common ways of terminating convolutional codes transmitted in blocks: flushing, and tail-biting.

When flushing, the encoder begins and ends in a known state. The beginning state is easily set before the transmission. To force the encoder to end in a known state, K-1 *flush bits* must be appended to the data sequence $d(i)$. Due to the transmission of coded bits corresponding to the flush bits, a (nominal) rate 1/2 encoder, when flushed, has a rate of $r=k/2(k+K-1)$, where k is the number of information bits in a block. The disadvantages of flushing are the reduction in code rate, and the energy lost to flush bits.

Tail-biting is a technique used to eliminate the overhead of flushing the encoder. For tail-biting, the encoder starts and ends in the same state, making the trellis circular. The starting/ending state of the encoder is determined from the data. The conventional way of decoding tail-biting blocks is to decode in a circle, passing at least 2 extra history depths of data through the decoder. (CRC has developed an efficient tail-biting decoder that only requires one extra history depth of processing.) Depending on the application and the block length, the increased decoding complexity may be worth the higher code rate and the energy saved in avoiding the transmission of flush bits.

EFFICIENT DECODING TECHNIQUES

This section discusses a number of techniques and tradeoffs that can be used to minimize the number of computations required for Viterbi decoding.

The Two-State Butterfly

A straightforward implementation of the add-compare-select (ACS) procedure could update the new state metrics and histories one by one. This has the disadvantage of requiring old state metrics and histories to be read from memory two times, when they need to be read only once. New states and histories can be grouped in pairs so that they depend only on one pair of old states and histories. This property is inherent in the way the states change in a convolutional code. This "two-state butterfly" implies a decomposition of the ACS processing into a series of efficient two-state update operations, requiring only one read per state.

Storing Branch Metrics in Registers

The motivation for storing branch metrics in registers is, like the use of the two-state butterfly, to reduce the number of read and write operations. In a binary rate 1/2 code, there are four possible combinations of $\{c_1(i), c_2(i)\}$, and thus four possible branch metrics at every time step. Every state transition requires a particular branch metric to be added to the old state metric. One approach involves pre-calculating the four possible metrics and storing them in the order in which they will be needed. This method requires

supplementary reads and writes, to add to the computations already required for the ACS portion of the algorithm.

A more efficient approach involves storing branch metrics in processor registers rather than in memory. In this approach, the loop around the ACS butterfly calculations must be unrolled because each butterfly requires different branch metric registers. Only two registers are needed to store the branch metrics, because two of the branch metrics are the negative of the other two. Unstored metrics can be obtained by replacing addition by subtraction in (2).

In summary, the core can be expanded, avoiding the reading and writing that memory-stored branch metrics would require. The cost is a fair amount of program memory and two registers dedicated to branch metrics.

Fast-Start

For a flushed block that starts and ends in a known state, the first and last K-1 Viterbi passes require only a fraction of the computation of a regular Viterbi pass. Calculating state metrics only for valid state transitions can save a lot of processing. In particular, the first K-1 passes (starting at a known state) require only additions and no comparisons, since only a single branch enters each state. In processors with many registers, one may even avoid the overhead of storing the state metrics in memory at every pass, as a few state transitions may be calculated entirely within the processor registers.

Efficient State Metric Normalization

Below are some efficient state metric normalization techniques that may be used alone or in combination with the periodic normalization that has already been discussed in a previous section.

Modulo Metric Approach

It is possible to avoid normalization in certain processors by letting the state metrics overflow naturally [3]. The metrics can still be compared correctly by allowing overflows to occur in the difference operations used for comparisons. This method works, as long as the maximum spread of the metrics remains within half the range of the state metric registers. The constraint on the spread is sufficient to ensure that the difference between state metrics modulo the state metric register range is the same as the non-modulo difference, and that the sign bit from this comparison is valid. Unfortunately, this method can only be used with processors for which conditional processing can be based on the sign bit of the ALU.

Branch Metric Approach

The subtraction of a value from all state metrics can be done very effectively when it is combined with the calculation of the branch metrics. Subtracting a constant from all the branch metrics before they are added to the state metrics during every Viterbi pass has the same effect as subtracting it directly from each state metric. This method is not applicable when only half of the branch metrics are stored in registers, because these may be added or subtracted from state metrics.

Normalization from an Arbitrary State Metric

Finding the maximum state metric requires the processing of all state metrics, and is therefore quite expensive. Instead, an arbitrary state metric can be used (say the first one) for normalization rather than the maximum one, eliminating a search through all metrics. The disadvantage is a doubling of the range of the state metrics, because of the possibility of normalizing with the minimum metric rather than the maximum.

Efficient History Management

In a straightforward implementation of the Viterbi algorithm, history management (3) may require as much as or more processing than the ACS operations (2). By using the techniques presented in this section, history processing requirements can be reduced to a fraction of that required by the ACS operation.

Full History Buffer Approach

The full history buffer approach requires a history buffer to be updated and copied at each Viterbi pass. Because practical history depths do not usually fit in a single processor word, many read/write operations are required for each state. This approach is therefore quite computationally expensive, and becomes increasingly so for long history depths.

Split History Buffer Approach

The management of state histories can be made more efficient. Instead of processing the full (multi-word-wide) state history at every pass, the buffer and its associated processing are split into two levels.

At the first level, processing is required at every Viterbi pass, and is identical to the full-buffer approach, except that the history buffers are a single word wide. These buffers are temporary, and hold at most L bits of history. The processing at this level can be made to be very efficient, compared to the full history approach.

The second level operations of *extraction,* and *retrace* occur every L Viterbi passes, and concern groups of L history bits. Extraction involves copying the accumulated state histories from the single-word temporary history buffers to a longer-term storage array. The single-word history buffer must also be reset after extraction. The retrace allows the history of a particular path to be reconstructed from this array.

The extracted history array is arranged in 2^{K-1} rows, by state, and in columns L bits wide by time. This arrangement allows paths to be reconstructed efficiently. This is achieved by storing the history columns so that the most significant K-1 bits of an element can be used as an index to its corresponding predecessor in the previous column. The amount of processing required to construct the proper columns depends on how the paths are stored and updated in the single word temporary buffer, as discussed below. Figure 3 shows the data structures required for the split history buffer approach and illustrates their use during a retrace.

When history is truncated, a retrace must take place every L bits, starting from the best state metric. Since the best state

metric can only be found by going through all 2^{K-1} metrics, it is a relatively expensive operation. This cost depends on the width of the stored history columns. To save the processing related to finding the best state metric, it is possible to retrace from an arbitrary state, but history depth must be increased by two or three constraint lengths to maintain the same performance as a retrace from the maximum metric.

Updating the Temporary History Buffer

First consider the case where each state history word is updated at each Viterbi pass as in (3), by appending the newest bit shifted into the state. Notice that the current state will always be mirrored in the K-1 least significant bits of the path history words. Not only are these last K-1 bits redundant (the state is known from the position of the word in memory), but they need to be removed before the words can be extracted to be copied to a history column (and put back before continuing). Clearly, this is not an ideal approach. A much better approach is to update the path history word as follows:

$$H_n(i) = [H_m(i-1), \mathrm{MSB}(m)], \; m \text{ is the best old state.} \qquad (6)$$

This approach appends bits "as they shift out of the state shift register." The redundancy is removed, and extraction of the words into history columns only requires copying, as the K-1 most significant bits are indices to the rest of the path in the previous retrace history column. This approach may not always be convenient to implement directly, but it leads to further improvements.

Preset History

Up to this point, the word level path history updating requires reading, shifting or masking bits into words, and writing at every Viterbi pass. The shifting or masking operations can be eliminated. We have observed that after K-1 passes, the history words contain the number of the state from which they have originated and we have used this property to produce the retrace data structure. The next step is to notice that if, upon starting, the path history words are *preset* with the state to which they correspond, the only operation required for the next K-1 passes to produce data suitable for extraction, is copying!

If we further require the extraction to occur exactly every L=K-1 passes, the history extraction requires no shifting or masking, and also simplifies to copying. The cost of presetting the history column every K-1 passes can also be eliminated by using a dedicated and pre-initialized column of state numbers for the first pass of each set of L=K-1 Viterbi passes.

The processing efficiency gained by choosing L=K-1 has to be weighed against the possible loss of storage efficiency for the extracted history bits. For example, storing the history columns in 16-bit words would be a fair compromise for most commonly used constraint lengths. Some applications may require multiple extractions to be packed in a processor word to improve storage efficiency.

Embedding History in State Metrics

The amount of processing required for history management has been reduced to the copying of L=K-1 bit history chunks, and a low-complexity retrace operation. The copying operation can also be eliminated.

The ACS processing is already carrying out decisions and moves. For processors having a wide word, the preset history bits can be embedded in the least significant bits of the state metrics, where they will not significantly interfere with the state metric calculations. Care must be taken to ensure that branch metrics added to the state metrics are shifted up so as not to corrupt the embedded history.

This last improvement reduces the history processing to periodic extraction and retrace, which amounts to only a small fraction of ACS processing.

IMPLEMENTATION

This section discusses how some of the techniques described above were applied to the design of a Viterbi decoder for the SHARC DSP. The implementation was optimized for decoding speed, not program size.

ADSP-2106x SHARC Characteristics

The ADSP-2106x Super Harvard Architecture Computer (SHARC) is a floating point DSP with a 32-bit wide data bus, sixteen all-purpose registers and single-cycle instructions. The multiply-accumulate unit and the ALU may be used simultaneously, all the while reading or writing one operand to *each* of program and data memory. Note that the Viterbi decoder implementation discussed here uses only the fixed-point processing capabilities of the SHARC.

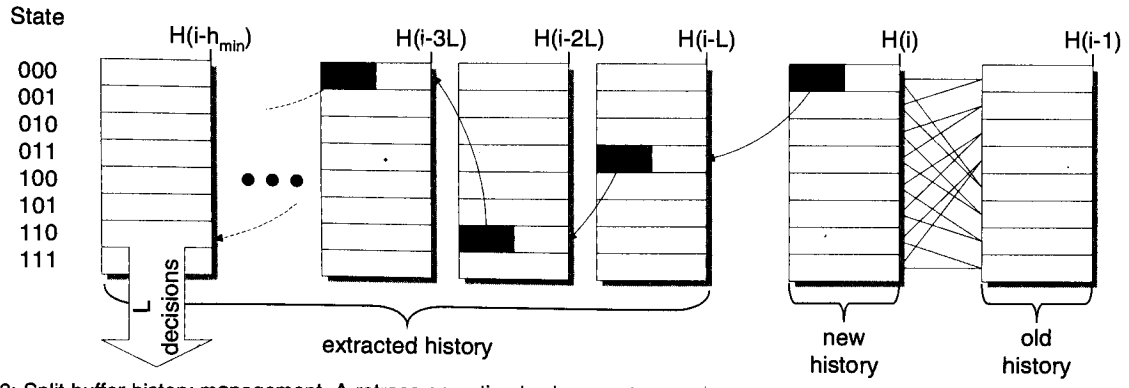


Figure 3: Split buffer history management. A retrace operation is shown, where the most significant K-1 bits of every column is used as an index into the previous one. The history columns are L bits wide, and decisions are taken L bits at a time.

The Four-Cycle, Two-State Core

By requiring that the code's generator polynomials have taps at both ends, one has only to use one branch metric (and its negative) per two-state butterfly. This is particularly useful, as it allows the use of the SHARC's dual-add-subtract and max operations to update two state metrics at a time in the ACS core. The ACS core is unrolled to allow storage of branch metrics in processor registers.

Because of the possibility of reading or writing data memory while doing either a dual-add-subtract or a max with the SHARC, the core ACS operation of Viterbi decoding can be effected in four cycles per two-state butterfly, or two cycles per state. The goal of the implementation is to reduce all other processing to be small, compared to the ACS processing, at two cycles per state.

Embedding History in State Metrics

The efficiency of the fast ACS operation can only be maintained by reading and writing the histories and state metrics simultaneously. The 32-bit wide registers of the SHARC provide plenty of space to embed history information within the state metrics. The most significant bits store the state metrics, and the least significant bits store the history information. The branch metrics must be aligned to the appropriate bits of the state metrics before adding.

The Special Core

A second "special" unrolled core is called every K-1 bits, and performs state metric normalization, history extraction as well as the ACS processing in three cycles per state. All these operations can be performed more efficiently in one routine since the data required for all these operations is identical. The multi-operation instructions allow all of these functions to be pipelined together in an efficient manner. In fact, since history extraction has to be done (at a cost of one cycle per state), the special core effectively performs periodic state metric normalization "for free."

The special core was also made to preset the history data "for free" by using a state metric/history buffer containing permanent preset history. This saves another cycle per state. If the memory budget allows it, the metric buffer with the preset history may be statically allocated, thus saving the 2^{K-1} cycles/block required for its initialization at every call to the decoder.

Only three cycles per state are required to do the usual ACS processing, state metric normalization, history extraction and presets. Since the special core is only used once every K-1 passes, the use of the two cores implies that, not counting overhead, the complexity of the decoder is $2+1/(K-1)$ processor cycles per state.

For rate 1/2, the dynamic range of the inputs is limited by the maximum allowable spread of the state metrics (4), and is reduced by one bit because we are normalizing with an arbitrary state metric. The allowable dynamic range of the inputs is 10 bits for $K \leq 9$ and 9 bits for all other implemented constraint lengths.

History Retrace

The history retrace operation in this implementation is simple because the stored history columns contain exactly K-1 bits of state history, allowing the use of the state history data directly as a pointer into the previous history column. In the actual program, the retrace processing requires 2 cycles per column, not including overhead.

This particular implementation retraces from an arbitrary state and uses three more columns of history to ensure good performance.

Memory Requirements

The use of program memory is divided between the main decoding routine and the unrolled cores. The size of the unrolled cores increases exponentially with constraint length, and the dual unrolled core approach quickly becomes impractical at constraint lengths above, say, K=11. Program memory requirements are about $250+5S$ 48-bit words, which includes the main program and two unrolled cores. Partially unrolled cores can be used to reduce the amount of program memory required, without significantly lowering the decoding speed. Data memory requirements are about $CS/2 + 3S$ 32-bit words, where S is the number of states and C is the number of K-1 bit wide history columns. Although fairly large, this memory is freed after every call.

PERFORMANCE

This section presents some experimental results and the throughputs of the SHARC Viterbi decoder.

Experimental results for a rate 1/2, K=7 code at a bit error rate of 10^{-3} have shown that C=8 and C=12 columns of history are required to give performance 0.16 dB and 0.01 dB from true MLSE decoding, respectively. C columns of history correspond to a minimum history depth of $h_{\min}=C(K-1)+1$.

The number of processor cycles required to decode a block depends on the constraint length, the history depth and the block length. Table 1 gives processing cycles per state and throughputs for a 256 information bit code with 12 columns of history. The cycle counts include all setup and "sanity checks." Note that throughputs include flush bits.

It is interesting to compare Viterbi processing requirements for the present SHARC decoder to those for the Motorola DSP56300/56600 series of DSPs, which include an instruction dedicated to Viterbi decoding [4]. Using the example in [4], a K=6, 168 information bit, flushed, rate 1/2 code, the Motorola DSPs use about 6.40 cycles/state, whereas the SHARC uses only 2.61 cycles/state. For larger constraint lengths, the comparison would be even more favorable for the SHARC.

K	States	Rate	cycles/state	kbps at 40MIPS
5	16	1/2	3.60	695
		1/3	3.77	663
6	32	1/2	2.78	450
		1/3	2.87	436
7	64	1/2	2.42	258
		1/3	2.47	254
9	256	1/2	2.18	71.7
		1/3	2.19	71.3
11	1024	1/2	2.12	18.5
		1/3	2.12	18.4

Table 1: Processing requirements for SHARC Viterbi Codec for a block of 256 information bits, flushed, minimum history depth of $h_{min}=12(K-1)-1$ bits. Throughputs include flush bits.

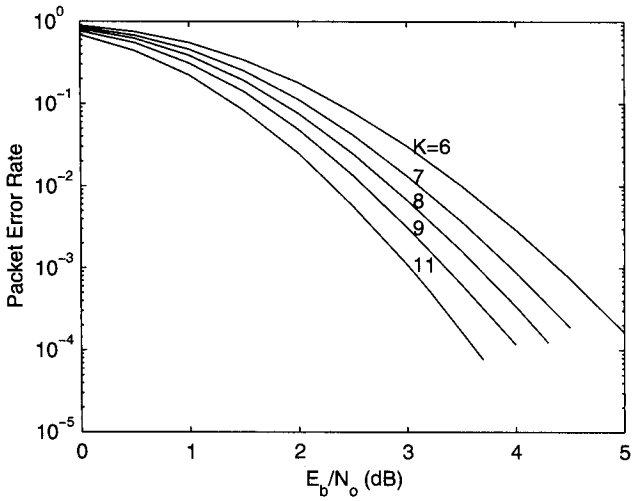


Figure 4: Packet error rate for different constraint lengths of rate 1/2 binary, flushed convolutional codes, 128 information bits per packet. Minimum history depth is $h_{min}=12(K-1)+1$ bits.

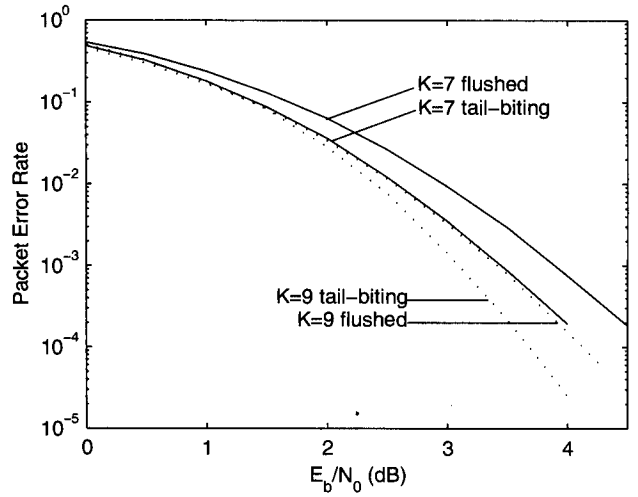


Figure 5: Packet error rate comparison for flushed and tail-biting K=7 and 9, rate 1/2 binary convolutional codes, 48 information bits per packet. Full history was used.

Figure 4 shows the Packet Error Rate (PER) of a rate 1/2 decoder for a 128 information bit flushed block, for various constraint lengths. Twelve columns of history were used. Figure 5 compares performance of tail-biting and flushed block termination for a relatively short block with 48 information bits. Full history was used. Note that for the range of results shown, the K=9 flushed and K=7 tail-biting codes have about the same performance, whereas the K=7 tail-biting decoder has about half the complexity of the K=9 flushed decoder for our implementations.

CONCLUSIONS

A very fast Viterbi decoder has been implemented for the SHARC processor, with a complexity approaching $2+1/(K-1)$ cycles per state. This very low complexity was achieved partly because of the SHARC's architecture, which allows for a very efficient 2 cycle per state ACS core, and partly by the use of techniques that kept all other processing to a minimum.

Very efficient history management enabled the most significant reduction in complexity. The presetting of history information and the embedding of history in the state metrics allow the history to be updated practically "for free." The remaining history retrace processing has an almost negligible cost, and renders the decoding speed quite insensitive to history depth. Eliminating the reading and writing of pre-computed branch metrics further reduced processing. This was achieved by storing branch metrics in registers and unrolling the ACS core.

FUTURE WORK

For high constraint lengths (e.g. $K \geq 11$) the amount of program memory used by the dual unrolled cores may be a problem. One solution is to partially unroll the ACS loop. The first 64 states can be put in a larger loop, which calculates changes in the signs of the branch metrics at each iteration. This approach has been used successfully in existing software to allow the core to fit in a PC's cache[5].

REFERENCES

[1] A. Viterbi and J. Omura, *Principles of Digital Communications and Coding*, McGraw-Hill, 1979.

[2] J. Proakis, *Digital Communications*, McGraw-Hill, 1983.

[3] H. Lou, "Implementing the Viterbi Algorithm," IEEE Signal Processing Magazine, pp. 42-52, September 1995.

[4] D. Taipale, "Implementing Viterbi Decoders Using the VSL Instruction on DSP Families DSP56300 and DSP56600," Motorola Inc., 1998.

[5] "Ultra-Fast Turbo and Viterbi Decoders for PCs," CD-ROM, available from the Communications Research Centre, Ottawa, Ontario, Canada, 1998.
See http://www.crc.ca/fec.

MAC Protocol Issues for Multimedia Satellite Systems

Janez Bostič
German Aerospace Center (DLR)
Postfach 11 16, D-82234 Weßling, Germany
Email: Janez.Bostic@dlr.de

ABSTRACT

The paper is dealing with the multiple access control (MAC) protocol issues for multimedia satellite systems. Such systems are going to support a lot of various access variants such as interconnection of networks, gateway and group terminal access and user terminal direct access. The latter one is probably the most important since it could support the terminal portability globally. The most complex problem is how to guarantee the quality of service (QoS) and still efficiently statistically multiplex the traffic from various terminals over radio medium. For the time division multiple access (TDMA) based techniques the entity which is responsible for guaranteeing the QoS and efficient statistical multiplexing is called scheduler and is a part of the MAC protocol. In this context, the basic approaches will be presented and a simple scenario with real-time variable bit rate (VBR) sources and a satellite as access point will be simulated.

INTRODUCTION

Multimedia satellite systems are enjoying more and more popularity in the last few years. That is mainly because they are able to support multimedia communications globally, together with the freedom of the terminal portability. The satellite communications systems are also essential in establishing the global information infrastructure (GII).

Existing satellite systems consist primarily of geo-stationary earth orbit (GEO) satellites. But future networks will consist of low earth orbit (LEO) and medium earth orbit (MEO) satellites, together with heterogeneous constellations of all three types.

Multimedia communications can be defined as a joint transfer of voice, video, audio, image, graphics, text applications or any combination of these media [1]. The asynchronous transfer mode (ATM) technology already supports efficient transmission with the guarantee of quality of service (QoS) parameters of such applications in the fixed networks. In addition, the standardization process of wireless ATM is going on. Therefore the introduction of ATM technology into new multimedia satellite systems can also be expected.

This paper begins with a short description of satellite multiservice satellite architectures based on ATM technology and arguing the problem of medium access control (MAC) protocol for uplink channel (from portable or fixed terminal to satellite) which is one of the crucial issues for transmitting multimedia applications over such systems. In the second section the MAC techniques for the uplink channel access are discussed. The differences between the scheduling in fixed networks and scheduling in radio networks are also explained. The simulated scenario is based on multi-frequency time division multiple access (MF-TDMA) and a simple scheduling strategy. Since the presented simulator is a good basis for investigations on scheduling issues for multiservice terminal access, the paper concludes with the discussion of research topics for further work in the area of MAC and scheduling for the uplink access.

SATELLITE MULTISERVICE SYSTEM ARHITECTURES

The TIA TSB-91 document [2] provides architectures and guidelines for satellite ATM networks. Two groups of ATM network architectures are defined:

a) ATM network architectures for bent pipe (transparent) satellites, and
b) ATM network architectures for satellites with on-board ATM switches.

These architectures are foreseen for all LEO, MEO and GEO constellations.

For the multimedia services the preferable architecture is the second one (b), because a lot of functions for traffic and bandwidth management can be implemented on-board the satellite. Such strategy is very advantageous for a MAC protocol, since the uplink and downlink propagation delays are major limiting factors on the performance of the MAC protocol.

The terminals can access the satellite (i) directly, (ii) through a concentrator or (iii) through a gateway station in both architectures. The latest two possibilities are preferred for interconnection. They have the advantage that already the end point statistically multiplexes the traffic from various terminals. The output is therefore smoother and the bandwidth required over the air interface could be fixed or changes very slowly.

The big difference among (i), (ii) and (iii) lays in the signaling interfaces. The user-network interface (UNI) is used for the direct satellite ATM access and the network-network interface (NNI) for inter-satellite or satellite to ground station links. The network interfaces provide the signaling procedures for dynamically establishing, maintaining and clearing of ATM connections.

The uplink and downlink channels can be realized in frequency division duplex (FDD) or time division duplex

(TDD) mode. Although some wireless ATM systems have proposed TDD mode MAC protocols, for satellite systems the FDD mode based MAC protocols are preferable. This is mainly because of typical propagation delays for satellite systems. Therefore, the next sections deal only with the uplink channel access, since the dowlink channels can be implemented by TDM mode or asynchronous TDM as proposed for the Teledesic system [3].

MAC TECHNIQUES FOR UPLINK ACCESS

The MAC protocol for the satellite broadband multi-service systems must achieve the following goals:

- QoS guarantee, which means that the MAC protocol will guarantee the connection parameters negotiated at the connection setup for the time of the connection. The QoS parameters such as cell transfer delay (CTD) and cell delay variation (CDV) are very sensitive for real-time services whereas the cell loss ratio (CLR) is very sensitive for data services.
- Support of different kinds of traffic categories such as constant bit rate (CBR), real-time variable bit rate (rt-VBR), non-real-time variable bit rate (nrt-VBR), unspecified bit rate (UBR) and available bit rate (ABR),
- Fairness, i.e. the MAC protocol must serve the terminals of the same priority class with equal probability,
- Efficiency, i.e. the MAC protocol must minimize the network bandwidth usage while guaranteeing QoS,
- Small signaling overhead of the MAC protocol functions, i.e. the information flow between the access point and terminal should be as small as possible.

Let us first give a short look into common multiple access techniques and their suitability for transmission of multimedia applications.

Multiple Access Techniques

There are many techniques for the access to a shared medium such as the radio medium. They can be classified into: fixed assignment, demand assignment, random access, combination of random access and reservation, and adaptive protocols [4].

With the fixed assignment of the bandwidth the users are assigned a priori a constant capacity, and therefore such techniques are not suitable for services with changing bit rates.

The random access techniques are also not appropriate for multimedia services, since they cannot guarantee the QoS parameters, mainly because of collisions. This can be seen also in Figure 1, which shows the comparison of the cell loss probability for the propagation delay 10.6 ms between the fixed TDMA and advanced packet reservation multiple access (A-PRMA) techniques for multiplexing of VBR sources [5]. The VBR sources were modeled as described later, and have standard deviations of 28.5 kbps, 57 kbps, and 101 kbps. The curves for TDMA are calculated with the same standard deviations. It can be concluded that if the standard deviation raises, the cell loss probability also raises.

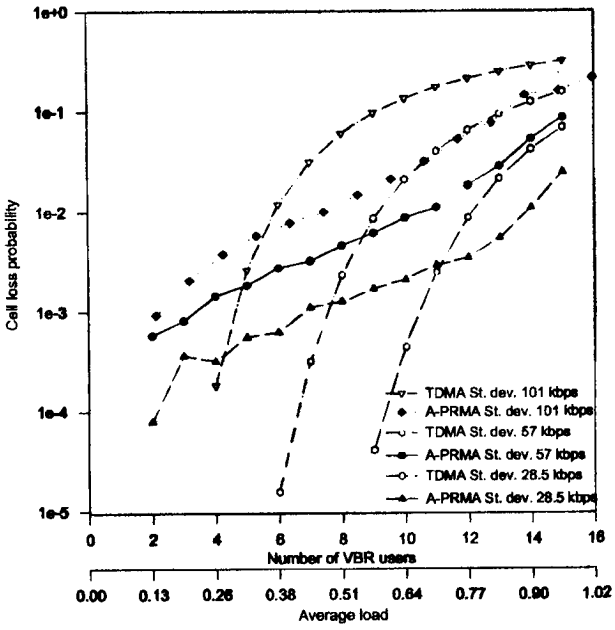


Figure 1: Comparison of fixed TDMA and A-PRMA for VBR traffic

In both cases the delay is not taken into account (the cells are not queued for TDMA, and for the A-PRMA they are not re-transmitted if collision occurs), and because of that the cell loss ratio is very high. The A-PRMA outperforms the fixed TDMA only in a narrow interval of traffic load as can be seen. The average load means the terminals load which should be transmitted.

The demand assigned multiple access techniques can be based on fixed or variable rate demand assignment. In the first case a constant bandwidth is assigned for every new connection. On the other hand, the variable rate demand assignment allocates the capacity for connections dynamically during the duration of the connection based on the changing traffic rates.

The demand assignment techniques can also be combined with the free assignment techniques, which allocate the rest of the capacity to active connections according to the specified criteria.

The combination of random access and reservation is the most preferred solution since the access signaling channel can be shared among various users for the initial access and reservation of the capacity for signaling channels or even data channels if they are used for signaling too. The random access method can also be used for the access of signaling channels during ongoing connections, if it is not done implicitly or there are no dedicated signaling channels.

The adaptive protocols are those which can for example change the number of reservation or contention dedicated channels according to specified parameters.

Since the QoS is the most important criterion for multimedia services, the reservation based techniques

together with MF-TDMA are the most suitable. The scheduler is the entity which guarantees QoS in fixed networks and therefore in the next section some basic features of it are described.

Scheduling of Packets in fixed Networks

In fixed networks the scheduling is associated only with the outgoing links as shown in Figure 2. The scheduling function, which schedules the cells according the pre-negotiated QoS parameters to outgoing links, is realized within the ATM switch. In this context, it is assumed that the access links from the sources and the input buffers are dimensioned in such a way that they do not impose any constraints on the traffic, i.e. cell rate.

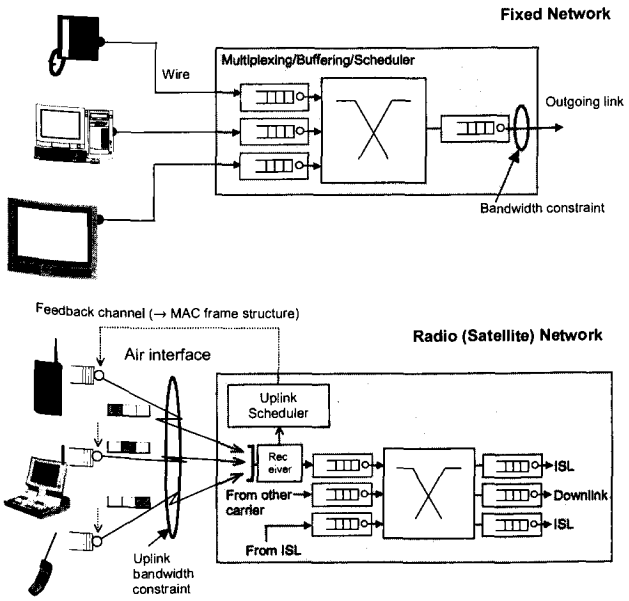


Figure 2: Scheduling in fixed and satellite networks

The scheduling can also be seen as the mechanism that determines which queue is given the opportunity to transmit a cell. In general, the queues can be organized as per-group queuing or per-virtual channel/virtual path (per-VC/VP) queuing [6]. With per-group queuing a number of connections share the same queue in a first-in-first-out (FIFO) arrangement. In this queuing structure, the connections can be categorized according the service category, service class or conformance definition into groups. On the other hand, with the per-VC/VP queuing the cells of each VC or VP are queued independently.
A very simple queue structure scheduling technique is the priority based scheduling, which assigns a priority to each queue and serves them in order of priority. The support of QoS is only limited. Better techniques are those based on fair share scheduling, where for each queue is guaranteed that it gets a share of link bandwidth according to a defined weight. The weight is the criterion for connections with equal priority but different traffic and QoS parameters. In this way, they guarantee a certain minimum rate allocated among the queues and are further classified as rate allocation (work conserving) and rate-controlled (non-work conserving) schedulers. With a work conserving scheduling a server is never idle when there is a packet to send. On the other hand, with a non-work conserving scheduling the server may be idle even when there are packets waiting to be sent [7].
The entities that control and guarantee that the traffic parameters are in accordance with the pre-negotiated at the connection setup are called usage parameter control (UPC) and network parameter control (NPC) and are not part of scheduling [8]. They have similar functions but at different interfaces: UPC is done at UNI, whereas NPC is done at the NNI. Their functions include monitoring cell streams, checking the conformity between the actual cell stream and the nominal cell stream (the traffic descriptor values) and taking necessary actions when unconformity is detected. The actions that can be taken are discard of cells immediately or tagging of them for discard at a congestion point when the network is congested. The cell loss priority (CLP) bit is used for tagging.

Scheduling of Packets in Radio (Satellite) Networks

In radio networks, the constraint is on the total bandwidth available to all users before the access point, i.e. satellite or base station, respectively. Because of that, for the exploitation of the statistical multiplexing in a TDMA system, a scheduler is needed to organize the frame structure in the satellite uplink channels. The information on the slot assignments, as decided by scheduler, is broadcast to the user terminals via a feedback channel as shown in Figure 2.
The connection QoS parameters are negotiated during the connection setup. The entity which decides if a new connection can be accepted or not runs a connection admission control (CAC) algorithm. This entity can be placed on the on-board processing satellite or in the ground network control station. However, the parameters of the accepted connection must be transmitted to the scheduler which uses them for the process of slots allocation.

As already mentioned the big difference between the scheduler in fixed networks and the scheduler in radio networks is that the latter one does not schedule the traffic based on arrived cells in queues but on arrived requests for the bandwidth capacity (number of time slots in case of TDMA-based MAC).
The first criterion for the scheduling decision is the priority of the service category. As shown in Table 1 the highest priority is assigned to CBR service category which is then followed by other service categories.

Priority number	Service category
5	CBR
4	Rt-VBR
3	Nrt-VBR
2	ABR
1	UBR

Table 1: ATM traffic categories priorities

The UPC for the uplink could be performed in the terminal or within the scheduler entity on on-board the satellite. The downlink does not need UPC because this traffic has already gone through the uplink UPC and the traffic is conformed to traffic descriptors. The UPC functions implemented in the terminal can only use discard principles since the traffic over the air interface has also to be conformed to traffic descriptors. On the other hand, the scheduler implemented UPC functions could also use the tagging principle if there is enough capacity. In this case the scheduler could allocate additional slots if there are non-conforming requests and when the cells would reach the access point, the CLP bit of cells in non-conforming slots could be set to lower priority as in fixed networks.

MF-TDMA based Multiple Access Technique

The MF-TDMA access scheme has been proposed for different GEO and LEO systems. MF-TDMA frame is normally divided into two areas: the first one is intended for synchronization and signaling information transmission, whereas in the second one the data is transmitted. The requests for bandwidth allocations can be transmitted via out-of-band request slots, which are part of the first area. However, the in-band (implicit) requests can be sent together with data packets.

The first task of the scheduler is to calculate the number of slots which will be allocated to the specific connection. This has been presented for hierarchical round robin (HRR) scheduling in [9].

Then, the allocated number of slots has to be assigned to actual slots in MF-TDMA frame. This process is much more complex and needs also the time component, such as virtual arrival times.

For this first approach, we decided first to implement a very simple strategy which is based on priority and first-come-first-serve (FCFS) strategies. In this way, when the new connection is setup, the scheduler writes the connection parameters into a linked list. This list is ordered according to priorities of the traffic classes of the connections. The scheduler begins on the top of the list and allocates the slots to services with equal priority in a round robin manner as shown in Figure 3.

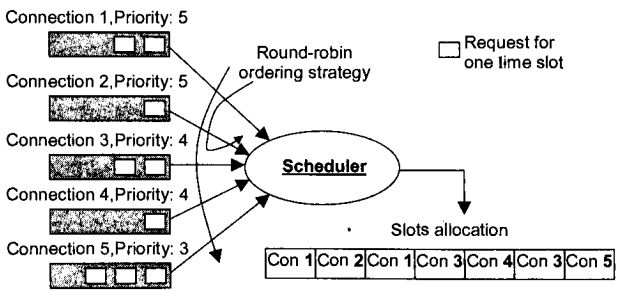


Figure 3: Scheduler operation

First, it serves all CBR connections, then rt-VBR ones and if capacity is available also other connections, until all time slots are allocated. In such a way every connection of the same priority gets a minimum capacity.

SIMULATION SETUP

The simulation has been realized by the SDL Design Tool (SDT), which is very appropriate for real-time, interactive and distributed systems. The basic simulation scenario is shown in

Figure 4. The scenario consists of a terminal segment and an access point (satellite) segment.

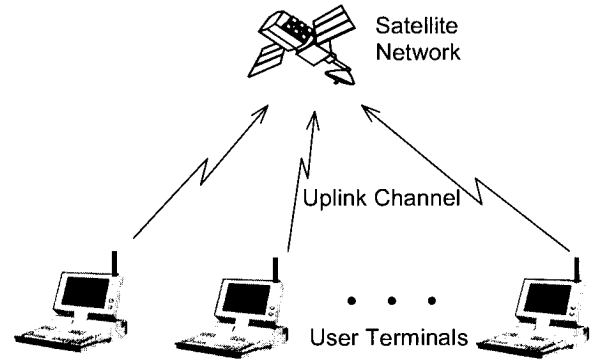


Figure 4: Simulation scenario

Since the slots reservation information is broadcast on a frame basis (the scheduler also operates on a frame basis), the propagation delay can be quantized into the segments of frame length. The minimum delay that can be achieved is one frame, but only if the requests for capacity are sent explicitly at the beginning of the frame. The Figure 5 shows the delay effect on uplink and downlink transmission. The SDT system is programmed in a way that the scheduler runs when the uplink frame arrives.

The terminals operate on call and cell level. Now they can generate of CBR or VBR traffic sequences. The VBR model represents the video-telephone scenes sequence [10] and is based on an autoregressive model as shows the following equation:

$$Y(n) = 0.871 \times Y(n-1) + 0.213 \times w(n) \times M$$

where $Y(n)$ is the cell rate in the n^{th} video frame (a frame is generated every $1/40^{th}$ of a second), $w(n)$ is a Gaussian random variable with a mean of 0.572 and variance 1.0, and M is the mean cell rate per frame of the video source.

The access point segment consists of call control and scheduler part. The call control is responsible for functions on the call level, such as CAC. The scheduler operates as described in the previous section. All connections are ordered according to priority in a linked list. They are not weighted since only one traffic source model was simulated. When the scheduler finishes the operations the satellite broadcast the information to all active terminals. They know then exactly in which time slot they can send the information and then there is no contention.

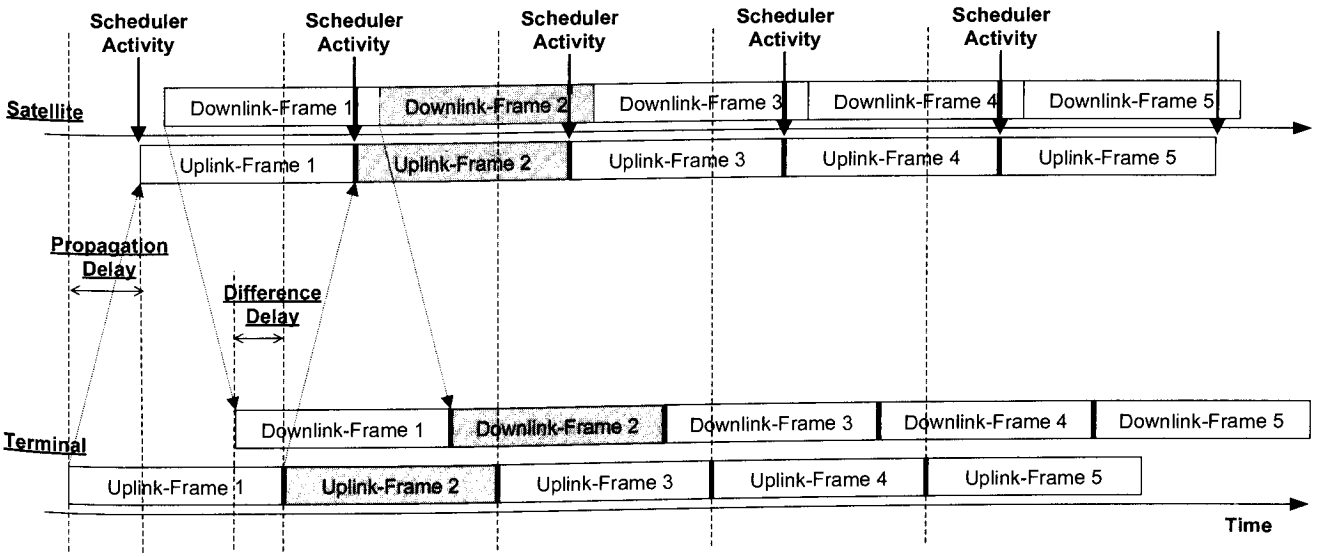


Figure 5: Delay effect on uplink and downlink transmission

RESULTS

The following parameter values have been set:

- Frame length: 26.5 ms,
- Carrier capacity: 2 Mbps,
- Number of time slots per frame: 125,
- Slot capacity (granularity): 16 kbps.

It has also to be mentioned that only one carrier was simulated since the hoping between different carriers have not been foreseen. The simulations have been performed for different number of traffic source. CLR was null, since the cells wait in the terminal queue until they are transmitted. Since the scheduler operates on a frame basis and that there is frame length granularity, the same results are obtained for the delays which are within granularity segment, only the difference delay changes as can be seen in Figure 5. The results for mean cell delay for the delay between one and two frames length shows Figure 6. As expected the delay increases with the number of traffic sources.

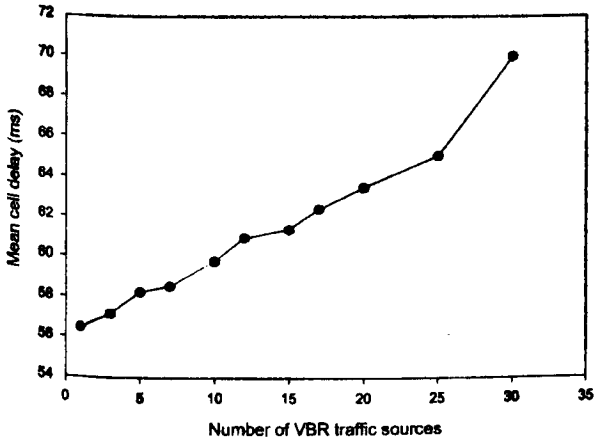


Figure 6: Mean delay of cells for various number of VBR traffic sources

The mean delay changes slowly and the difference between the minimum delay and mean delay remains within one frame length. Such situations appear normally in non-geostationary satellite systems, where still the propagation delay variation appears, which adds problems to the synchronization. In this study it is assumed that these problems are solvable.

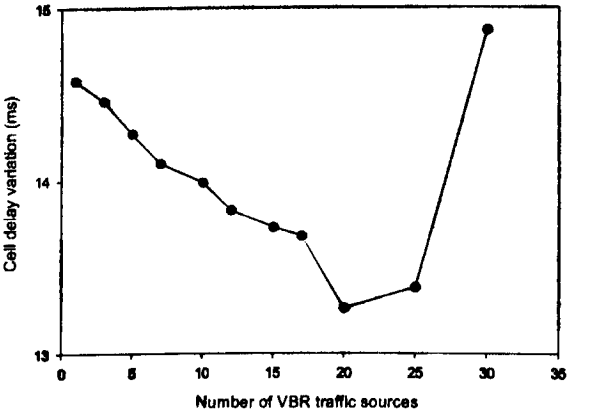


Figure 7 shows the CDV for different number of VBR sources. The minimum occurs not if only one traffic source is active. That can be explained with the fact that all sources have the same traffic descriptors and that the scheduler does not take into account the time difference between two consecutive generated cells. The scheduler thus implicitly better allocates time slots if more traffic sources are active. In the future therefore, the sources with different traffic parameters and weights should be simulated.

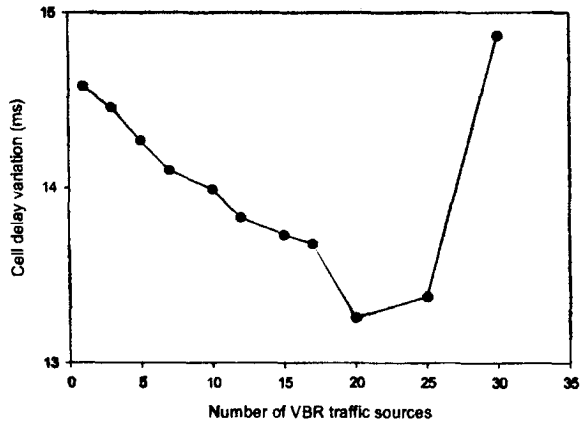


Figure 7: Cell delay variation different number of sources

FURTHER WORK

The results of the first simulations presented in this paper are a good basis for further investigations. There is always a tradeoff between CLR and cell delay. The performed simulations have assumed that the CLR is zero, which is for real-time applications not applicable. The delay should be controlled already in the terminal buffer and cells with exceeding-delay should be discarded already there.

In our simulations the traffic on the uplink was generated conforming to the traffic descriptors, because the UPC functions are already performed in terminal. But as it was already mentioned, the UPC could be performed also on-board the satellite. In this way, the network would control the access of terminals as in fixed networks. For this the scheduler has also to know the basic traffic descriptors such as mean cell rate, peak cell rate and cell rate variation of the service category. According to these parameters the virtual leaky bucket process can generate "tokens" for each virtual connection token pool. The process is shown in Figure 8. For the VBR connection the tokens are generated at the mean bit rate. The scheduler decides then if the connection will get the requested slots in the new frame based on the requests and the tokens in the token pool. When a new slot is allocated the token pool is decremented, and if no tokens are available, the connection will get new slots only if there are any slots free after serving other connections.

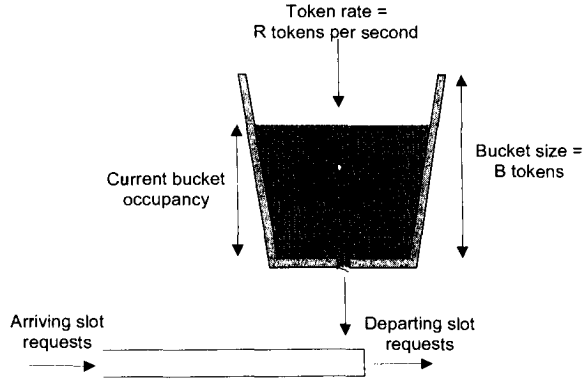


Figure 8: Virtual leaky bucket process

There are still the problems regarding the cell delay variations. If the CTD and CDV must also be guaranteed the virtual leaky bucket is not enough. The time between the two consecutive arriving cells has to be the same or nearly it was in the source. Therefore, the scheduler has to be delay-oriented. Because the uplink queues are not located in satellite, where the scheduler is, the arrival time of the uplink ATM cells should be estimated. If the new slots are implicitly requested the scheduler knows when the last cell has arrived and is also aware of the number of cells waiting for transmission in the terminal buffer. From this information it can estimate the arrival times for the cells (the expiry time of the next cell) and allocate the time slots according to the following formula:

$$T_{Deadline}(i) = T_{Last\_Uplink\_Cell} + (i\text{-}1)/R_{Peak\_Up} + \Delta, \quad i = 1,2,...,N$$

where Δ is the maximum time which can elapse between the arrival of an ATM cell at the satellite and the time in which this cell will be transmitted in the downlink, R_{Peak_Up} is the peak uplink bit rate, $T_{Last\_Uplink\_Cell}$ is the arrival time of the last ATM cell at the satellite, and N is the number of ATM cells waiting in terminal buffer for uplink transmission. This process will probably still not totally eliminate the CDV, but decreasing can be expected.

CONCLUSIONS

In this paper some issues on MAC protocol for multiservice multimedia over satellite were presented. At the beginning a short overview of possible architectures and appropriate satellite MAC protocol techniques was introduced. Then, the differences between the scheduling in fixed networks and uplink scheduling in radio (satellite) neworks was described. To assess the ideas, a simple simulation scenario was implemented and simulated. This part has concluded with some results and research topics for future work.

The future work will concentrate on improving the scheduling technique that it could support all aspects of QoS.

In addition, other traffic sources, such as internet applications should be included. For this kind of traffic the scheduler strategy simulated in this paper may be appropriate since these traffic streams have almost no time constraints.

REFERENCES

[1] TR 101 374-1 V1.2.1, Satellite Earth Stations and Systems (SES), Broadband satellite multimedia; Part1: Survey on standardization objectives, October, 1998.

[2] TIA/EIA Telecommunications Systems Bulletin 91 (TSB-91), Satellite ATM: Architectures and Guidelines, May 1998.

[3] http://www.teledesic.com.

[4] H. Peyravi, Medium Access Control Protocols Performance in Satellite Communications, IEEE Commun. Mag., vol. 37, pp. 62-71, March 1999.

[5] Janez Bostič, et al., Multiple Access Protocols for ATM over Low Earth Orbit Satellites, Proceedings 1st Joint COST 252/259 Workshop, Bradford, UK, April 1998.
[6] N. Giroux and S. Ganti, Quality of Service in ATM Networks, Prentice Hall, ch. 5, pp. 85-119, 1999.
[7] H. Zhang, Service Disciplines for Guaranteed Performance Service in Packet-Switching Networks, Proc.IEE, vol. 83, pp. 1374-1396, October 1995.
[8] H. Saito, Teletraffic Technologies in ATM Networks, Artech House, ch. 3, pp. 71-98, 1994.
[9] A. Hung, M.-J. Montpetit and G. Kesidis, ATM via Satellite: A Framework and Implementation, ACM-Baltzer Wireless Networks, vol. 4, pp. 141-153, no. 2, 1998.
[10] B. Maglaris, et al., Performance Models of Statistical Multiplexing in Video Communications, IEEE Trans. On Commun., vol. 36, pp. 834-844, July 1988.

Observation, Characterization and Modeling of World Wide Web (WWW) Traffic

Carlo Matarasso
DLR-German Aerospace Center
Institute for Communications Technology
Digital Networks Department
Oberpfaffenhofen
Postfach 1116
D-82230 Wessling, Germany
phone: +49 8153 28-2848
fax: +49 8153 28-2844
email: Carlo.Matarasso@dlr.de

ABSTRACT

This paper shows our results in the observation, characterization and modeling of World Wide Web (WWW) traffic. First of all we calculated the main parameters that characterize the behavior of WWW clients. Based on previous studies, we then defined an analytical model for WWW users and we calculated its parameters. By multiplexing together a high number of the above defined single traffic sources, we generated time series that simulated the aggregation of many WWW clients surfing the net simultaneously. With this procedure we were able to produce time series with self-similar properties, similar to the ones exhibited by Internet traffic. Our most important outcome showed that the grade of self-similarity increases together with the number of clients multiplexed.

1. INTRODUCTION

The first public release of WWW was issued in 1992 by the European Laboratory for Particle Physics, as a mean for scientific documents distribution. It had immediately an unexpected success in many different areas, like commercial and educational, becoming in a few years the major cause of Internet traffic [1].
The characteristics of WWW traffic have been matter of intensive studies in the recent years, a fundamental observation is the failure of the Poisson model for the arrival process of the Internet traffic [2]. Traces of aggregated WWW traffic showed bursts of data on a wide range of time scales, and no typical burst size could be found in time intervals up to 1 hour long. These properties cannot be replicated by modeling the packets inter-arrival times with a Poisson process, that would lead to a characteristic burst-length, which would tend to be smoothed when averaging on long enough time intervals [3]. On time spans longer than one hour the wide-area traffic is not stationary anymore, showing cyclical properties: some of them repeating themselves on a daily rate, some on a weekly one, some are even seasonal. [3] made the fundamental discovery that WWW has self-similar properties on time intervals where it can be considered stationary.
We based our studies of WWW single clients traces recorded by the Boston University (BU) Computer Science Department spanning on a period beginning on the 24th of November 1994 and terminating on the 8th of May 1995 [4].
In section 1 we will make a extensive survey on the status of the characterization of WWW traffic.
In section 2 we calculate some parameters that characterize single WWW clients.
This numerical results will lead to the determination of an analytical model for WWW clients that will be described in section 3.
Section 4 will then contain the results of our efforts to generate self-similar time-series, that could be representative of Internet traffic.

2. OVERVIEW ON PREVIOUS STUDIES ON WWW TRAFFIC

The studies of WWW traffic began around 1994, when web browsing was becoming the application that generated the highest amount IP packets in the Internet. Already in 1995 WWW had the highest share in Internet traffic, Table 1 contains the precise picture of the 1995 situation [5].

Table1: Internet traffic shares in 1995 [5].

WWW	21%
FTP-data	14%
NNTP	8%
Telnet	8%
SMTP	6%
IP	6%
Domain	5%
Other	32%

In the following years the situation further evolved towards an increase of WWW share, already in 1997 the bytes generated by web clients and servers together comprised 75% of the overall TCP traffic, which was the cause of 95% of the Internet bytes [5]. The research on WWW was basically pushed by the need of the service providers to more efficiently manage web packets through the networks, in order to offer an higher quality of service.
The BU Computer Science Department carried out an extensive characterization of the web traffic and the behavior of its clients. Their studies were all based on records of WWW sessions generated by its students on a 6 months span between November 1994 and May 1995 [4]. Their main results were [3]

- The packets inter-arrival times cannot be described with a Poisson or Markovian model, because these two processes would lead to a characteristic burst size that tends to be smoothed when averaging over long enough time intervals, while wide area internet traffic shows high burstiness on every time scale.
- Highly bursty traffic can be statistically modeled by a self-similar process. Self-similar processes exhibit long-range dependence: two separated values are non negligibly correlated on any time distance.
- Self-similar traffic can be generated by means of multiplexing a large number on ON/OFF sources, whose ON/OFF duration distributions are heavy-tailed. The ON times represent intervals during which the client is actually receiving data, they can be recognized by the presence of a currently open TCP connection. The OFF times represent the machine idle intervals, due both to user *thinking times* (inactive OFF times) and data parsing (active OFF times). During the OFF times no TCP connection is open.
- The BU studies produced many other important results, but the ones highlighted in the above points are the important ones for this paper.

If self-similar traffic (seen as a time series) is generated by the multiplexing of many ON/OFF processes, with ON/OFF times described by a Pareto distribution, then, if the number of processes is high enough, Willinger et. al. [13] demonstrate that the Hurst parameter of the series is

$$H = \frac{(3 - \min(\alpha_{ON}, \alpha_{OFF}))}{2} \quad (1)$$

Deng [7] observed that the active OFF times may be better described with a Weibull distribution, while the *thinking times* with a Pareto distribution.
Thompson et. al. [5] observed that, although web clients and servers produce almost the same share of TCP packets (respectively 30% and 38%), the clients are the cause of just 6-8% of the overall bytes. This numbers show the typical client-server peculiarity of WWW, that normally consists in a short request from the client to the server, followed by a burst of data in the opposite direction. In this article we will consider only the packets that flow from the server to the client, because they cause the overload of the Internet.

In conclusion we want to point out that Web traffic is not stationary on intervals longer than 1 hour, as a matter of fact [8] observes that the session arrival process exhibits a clear diurnal cycle, with peak during afternoon.

3. WWW CLIENT TRAFFIC CHARACTERISATION

Our calculations are based on HTTP logs recorded in the BU across 1994 and 1995 [4]. This logs were generated by 37 SparcStation 2 workstations located inside the campus facilities. During those years Mosaic was the most popular browser, although in the recent years it has been almost completely replaced by other browsers, we think that the traffic characteristics are mostly affected by network causes and files availability, while improvements of browsers cannot cause radical changes in the packets patterns. And even if the networks underwent to major changes and the number of files on the web increased dramatically, we think that those traces are still representative of the behavior of WWW users.
Each session was recorded in a text file, the file name contains an user identification codename, the machine name and the session starting time. The file body is constituted by lines, every line corresponds to a single URL request. Each line contains the following data: the machine name, the timestamp when the request was made, the URL address, the amount of downloaded bytes and the transfer duration. Cache hits were also recorded, and they can be easily recognized and eventually filtered if needed, because they are characterized by absence of transferred data and duration equal to 0 seconds. This is because when a request can be satisfied by the machine internal cache, it does not need to be sent to the server, and thus there is no data flowing in the net, but the request is nevertheless produced by the browser, and therefore recorded in the session log.

Session Duration Times

The first parameter that we observed is the session duration time. A session begins with the launching of the browser and ends with its closure. While determining its probabilistic properties we have to remember that this parameter can be heavily affected by the machine location: in our case the workstations were located in two rooms inside the BU, and could be accessed by several students. We can easily assume that every student was assigned with a fixed time slot, expired it, he had to leave the machine to the next in the reservation list. If we want to consider the more general case, where the users own the machine they are using, we observe that, while they have no external time constraints, they will probably limit themselves, because in most cases the billing policy will be proportional to the connection duration. These environmental diversities may affect the characteristics of the users behavior, and the numerical results that we obtain will need to be tuned when considering customers in other environments. Figure 1 shows in a logarithmic scale the probability density function (pdf) of sessions shorter than 1 hour. This comprehends about 95% of all logs. With the

exception of a 339 minutes long session, we observed that the other were all shorter than two hours. These evidences make us believe that the students were allocated one hour long time slots, and only in rare cases they could stay longer. We evaluated the average session duration time equal to 14 minutes and 2.502 seconds, a value longer than 74.17% of the samples. The variance is $2.7662*10^6$, this extremely high value shows how spread is the range of occurrences with non negligible probabiliy [9,10].

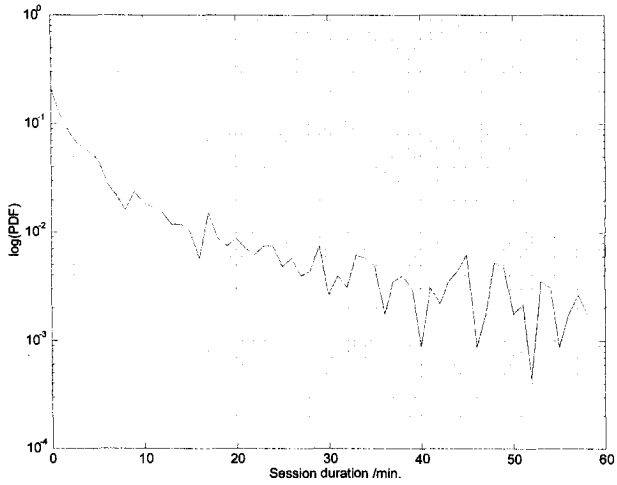


Fig. 1: Pdf of the session duration times shorter than 1 hour. The probability values are given in a logarithmic form.

ON Times

As already said in the introduction every WWW session can be subdivided into two subsets of intervals that alternatively follow each other

- The ON times, or activity intervals: during these periods the machine is actually receiving data from a web server. They can be recognized by the presence of a open TCP connection.
- The OFF times, or inactivity periods, in which no server to client data transmission is currently happening. They will be more accurately described in the next sub-section.

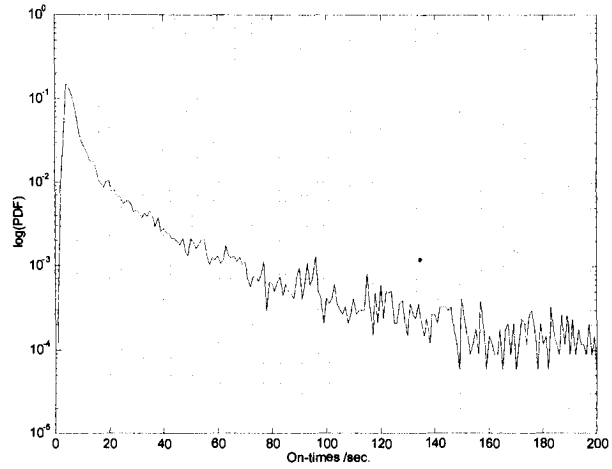


Fig. 2: Logarithmic pdf of the on times between 0 and 200 seconds.

Figure 2 shows the pdf of the ON times shorter that 200 seconds. The mean ON duration time is 3.1095 seconds, comprehending 86% of all records, while 95% is below the 10 seconds mark. The values above 10 seconds are very rare but they significantly contribute to the *tailness* of the distribution, in fact the variance is 547.0771 [9,10].

OFF times

During these periods the machine is idle, either waiting for a new user's request or being busy processing the just received data. In the former case the OFF times are usually called *thinking times* (or inactive OFF times) and they are human generated, the second ones are called active OFF times and they are machine generated. These two sets can be statistically separated, because they span in deeply different time ranges. While it is very unlikely that a human will observe the screen for less than 30 seconds, the amount of time needed by a workstation to process data will almost never pass the second. Of course nothing precise separates the two groups, therefore a separation will be always based on probabilistic basis. In figure 3 we can observe the range of times where the machine generated OFF times are dominant. We can notice two peaks (around 0.1 and 0.4 seconds) that may be produced by two typical processing times needed by the workstations.
Figure 4 shows an extended range of values, there we can observe that the pdf decreases constantly until the 3 seconds mark, than the tangent is not more constant until around the 20 seconds mark, where the plot becomes one more time linear (in a logarithmic scale). The interval between 3 and 20 seconds is characterized by the absence of a dominant process, while after 20 sec. the *thinking times* overwhelm the machine generated times.

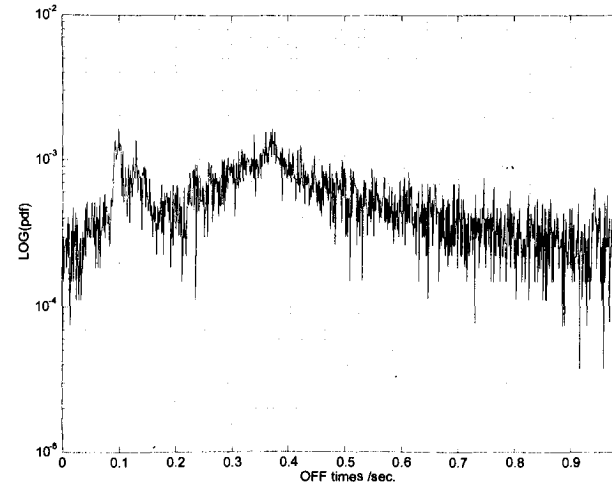


Fig.3: Logarithmic pdf of the OFF times shorter than 1 second.

The mean OFF times duration is 7.5481 seconds (comprehending both types), that is longer than 87% of all occurrences. We noticed that this value is more than two times the ON times one, because it is strongly affected by the human slowness, while the ON times duration depends only on network elements. Finally the variance is 4960.6, showing that the OFF times are a lot more spread than the ON ones [9].

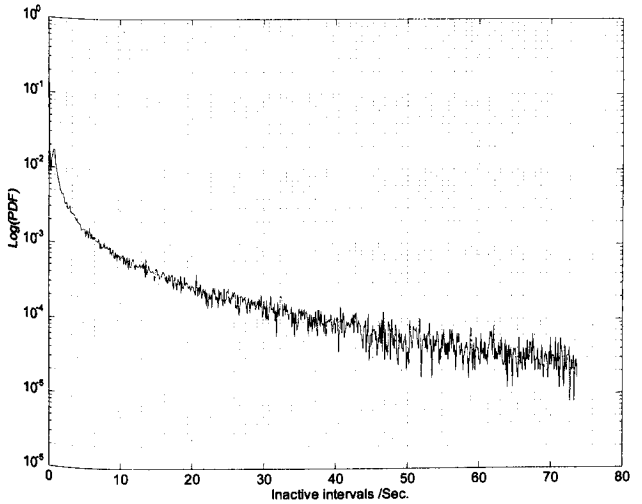


Fig. 4: Pdf of the OFF times ranging to 75 seconds.

4. WWW TRAFFIC MODELING

Completed the analysis of the clients behavior, we now have enough information to proceed to the synthesis, two are the possible approaches to this issue

- Using an **analytical model**: it is always preferable when the conditions for its implementation are fulfilled. Such a model tries to describe an entity as accurately as possible by means of mathematical functions. In order to determine a valid analytical model, the processes that characterize the observed phenomenon have to be correctly understood. Unfortunately many times the analytical model becomes too complex, and incomprehensible to the ones that did not participate to its definition, and thus of no use for anyone. In order to simplify the model scientists have to make assumptions on factors that, they suppose, do not have big impact on the behavior under study. Unfortunately most of the times it is not easy to isolate the important aspects from the negligible ones, and the simplifications lead to a model too far away from the original, thus wrong. A big advantage of a well done analytical model is that it can be applied on a wide range of phenomena, that have the same origin of the one the model was thought for.
- The second way to model an entity is by means of an **empirical model**. It tries not to understand the causes that produce a certain behavior, it simply tries to imitate their effects, treating the object as a black box. This model is clearly simpler than the former, but it lacks of transportability, because it is often valid only for the entity that he was made for.

Several previous studies observed that Internet traffic has self-similar properties [1,2,3,4,6,8], having a significant traffic variance on a wide time range. This characteristic cannot be described by modeling the arrival process with a Poisson or a Markovian model, because they would produce traffic with a characteristic burst size that would tend to smooth on long enough time scales. We will always consider the traffic as a succession of samples, and thus we have to think to self-similarity applied to time-series. If time-series is self-similar, then it exhibits long range dependency: the auto-correlation between two values is not negligible for every time distance [3]. This peculiarity of Internet traffic has shown to have a dramatic influence on tail of the queue waiting times [11], mainly due to the traffic burstiness caused by the long-range dependency.

Many possible models have been proposed to generate Internet conformant traffic loads. The most popular are

1. Multiplexing an high number of ON/OFF processes with heavy-tailed ON/OFF duration distributions [1,3,4,6,8]. This model has the advantage of being very simple to implement, but it needs an extremely high number of sources in order to approximate with acceptable accuracy a self-similar process. This means it requires a lot of CPU time and resources [12].
2. Another possibility considers a $M/G/\infty$ queue model, with clients inter-arrival times distributed like a Poisson process, and heavy-tailed service times [2]. Once again this model generates approximate traces, and a trade off between accuracy and computation length has to be met.
3. In [12] Paxon proposes a method based on the Discrete Time Fourier Transform. This model is more complicated from the mathematical viewpoint, but it does not require a lot of CPU resources keeping a good accuracy.
4. Other possible methods have been proposed, some of them are based on the wavelet transform, the ARIMA fractional process or the fractional Brownian process [12].

In the next section we will show the results we obtained by using the first of the above methods, while in the last part of this section we will determine the values of the parameters needed for the definition of the model.

We chose the Pareto distribution for the simulation of the heavy-tailness of the ON/OFF times. The Pareto distribution is heavy-tailed in its entire range, we preferred this function to the ones asymptotically heavy-tailed (for long enough times) because of its simplicity. In fig. 6 we will later show that the OFF times shorter than 10 seconds are not heavy-tailed, but in order to generate self-similar traces, the tails of the distributions are important, and the Pareto distribution is well suited for this target.

A distribution is heavy-tailed if

$$P[X \geq x] \propto x^{-\alpha}, \quad 0 < \alpha < 2 \qquad (2)$$

The probability mass function of the Pareto distribution is

$$p(x) = \alpha k^{\alpha} x^{-\alpha-1}, \quad \alpha, k > 0, \quad x \geq k \tag{3}$$

and its cumulative distribution function is

$$F(x) = P[X \leq x] = \int_{x_{min}}^{x} p(y)dy = 1 - \left(\frac{k}{x}\right)^{\alpha} \tag{4}$$

α is the important parameter in the formula. k will be dimensioned to satisfy the following relation

$$\int p(x)dx = 1 \tag{5}$$

We used a curve fitting algorithm to calculate α. If we define

$$\overline{F}(x) = 1 - F(x) \tag{6}$$

Then

$$\frac{dLog(\overline{F}(x))}{dLog(x)} = -\alpha \tag{7}$$

Therefore in order to determine α we plotted (7) in a log-log scale and then we fitted it with a linear regression algorithm.

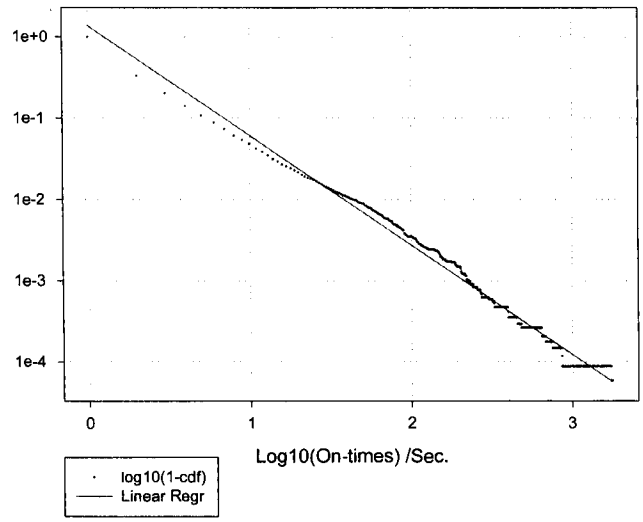


Fig. 5: Example of calculation of the Pareto parameter. The dotted line represent the $Log_{10}(\overline{F}(x))$ of the measured ON times and the straight line has been obtained with a linear regression algorithm. An analogous graph was made in order to determine α_{OFF} (see fig. 6).

In figure 5 we observe that the ON times have a constant α in their entire range, that we evaluated equal to 1.3483. On the other end, when we tried to calculate α_{OFF}, we found that in the range from 0 to 10 seconds (see fig. 6) α_{OFF} is by far not constant. We think that the machine generated OFF times are not at all heavy-tailed and should be described with some other type of distribution. The influence of the *active OFF times* stays strong until around the 10 seconds mark, for longer intervals the *thinking times* become dominant, and their heavy-tailed distribution becomes recognizable.

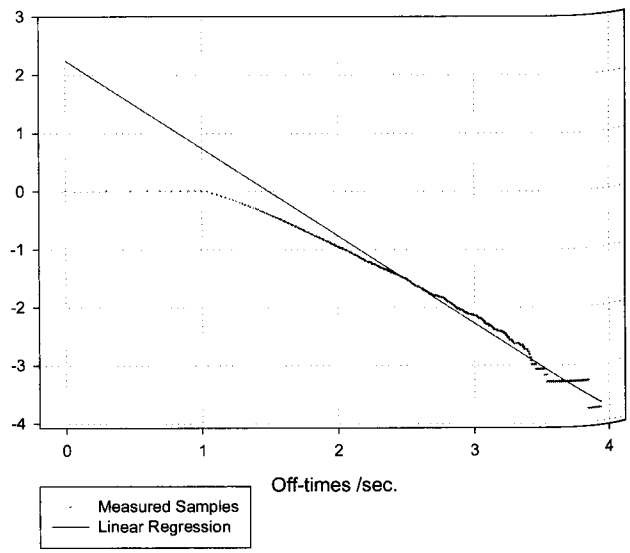


Fig.6: Linear regression for the calculation of α_{OFF}. Both axis are in logarithmic scale.

We did not try to model the OFF times in the 0 to 10 range, because the tails of the distributions are responsible for the self-similarity of the aggregate traffic. We calculated α_{OFF} of the OFF times longer than 10 seconds equal to 1.5057. In [6] Crovella et. al. observe that the machine generated OFF times may be described with a Weibull distribution.

5. GENERATION OF SELF-SIMILAR TIME SERIES

Now we have all the data we need to synthesize a self-similar series using the first algorithm outlined in section 4.

We ran two simulations, one with 1000 users simultaneously surfing the net for 10 minutes, the second with 10000 users and 3 minutes long. We did not generate longer times-series, because the algorithm is very slow and we also considered that in order to verify the grade of self-similarity, the duration of the series has very little importance. In this way we also avoided to generate series so long that they would not be representative of Internet traffic, being the real traffic not stationary after a certain duration threshold. To determine the grade of self-similarity we used a graphical heuristic method, known as *variance-time plot* [12]. If the series is aggregated by a factor m

$$X_t^{(m)} = \frac{1}{m} \sum_{i=m(t-1)+1}^{tm} X_i \tag{8}$$

then asymptotically the variance of the aggregated version falls of a factor $m^{-\beta}$ where $\beta = 2(1-H)$. A process is self similar if $0 < \beta < 1$ [3], the closer β to 0 the more the series is self-similar. Historically the Hurst parameter is normally calculated, in this case $\frac{1}{2} < H < 1$, as H approaches 1 the grade of self-similarity increases. With the *variance-time plot* test we can graphically calculate the H parameter, by

plotting the variance of $X^{(m)}$ against m on a log-log plot. The result should be a straight line with slope $(-\beta)$.

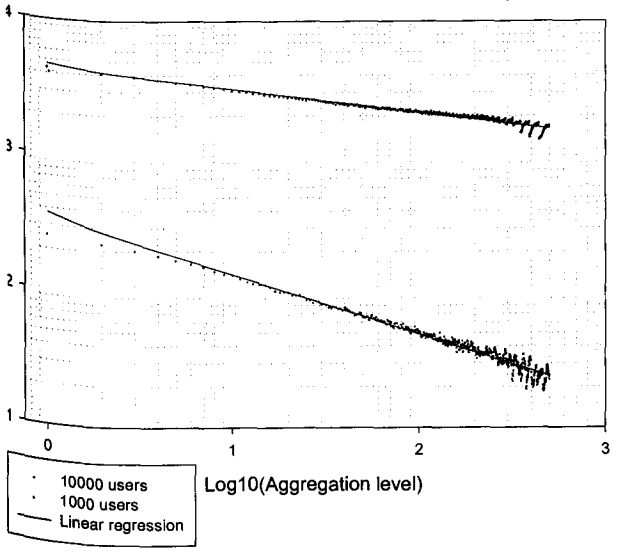


Fig. 7: *Variance-time test* for two sets of generated data. The higher plot was obtained by simulating 10000 ON/OFF processes, the lower plot was generated with 1000 processes. We did not normalize the values in order to draw a clearer graph for the reader. The straight lines represent the linear regression we used to calculate β.

From the observation of figure 7 we see that when more clients are simulated the grade of self-similarity increases, as a matter of fact the higher line drops clearly more slowly than the lower one. This is also confirmed by the theoretical viewpoint, as already pointed out in section 4, the generation of self-similar traces by means of multiplexing a high number of ON/OFF processes is only asymptotically valid: the more sources are multiplexed the higher becomes the grade of self-similarity. With 1000 clients we evaluated $\beta = 0.5163$ and H=0.74185, with 10000 clients $\beta = 0.1538$ and H=0.9231. According to (1) we expected to have H=0.82585, this is an asymptotic value that would be eventually reached with an extremely high number of multiplexed processes. Being this value right between the two that we graphically determined in fig. 7, we think this is a confirmation of the validity of our evaluations.

6. CONCLUSIONS

We began this paper with a quick overview on the status of research over Internet and in particular WWW traffic. By means of logs of single WWW clients [4] we calculated some significant parameters that characterize the behavior of WWW clients. With these results we could determine the parameters of a ON/OFF process that imitates the basic properties of WWW surfing. By multiplexing together a high number these processes we could evaluate the effects of web browsing on the Internet. The traces we generated showed to have self-similar properties, many previous studies pointed out that the Internet traffic has as well self-similar characteristics.

In conclusion we have implemented a generator of self-similar traffic that can be used to simulate aggregate traffic produced by web clients.

REFERENCES

[1] M. E. Crovella, M. S. Taqqu, A. Bestavros, "Heavy-Tailed Probability Distributions in the World Wide Web", submitted for publication in the book "A Practical Guide To Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions", 1996, Birkhauser, Boston.

[2] V. Paxon, S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling", IEEE/ACM Transaction on Networking 3(3), pp. 226-244, June 1995.

[3] M. E. Crovella, A. Bestavros, "Explaining World Wide Web Traffic Self-Similarity", Boston University technical report TR-95-015.

[4] C. A. Cunha, A. Bestavros, M. E. Crovella, "Chacteristics of WWW Client Traces", Boston University Computer Science Deparment, Technical Report TR-95-010, April 1995.

[5] K. Thompson, G. J. Miller, R. Wilder, "Wide-Area Internet Patterns and Characteristics", IEEE Networks, Nov/Dec 1997.

[6] M. E. Crovella, A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes", IEEE/ACM Transactions on Networking Vol. 5, Number 6, pp. 835-846, December 1997.

[7] S. Deng, "Empirical model of WWW document arrivals at access links", Proceedings of the 1996 IEEE International Conference on Communications, June 1996.

[8] W. Willinger, V. Paxon, M. S. Taqqu, "Self-Similarity and Heavy Tails: Structural Modeling of Network Traffic", appears on pp. 27-53 in the book "A Practical Guide To Heavy Tails: Statistical Techniques and Applications", 1996, Birkhauser, Boston.

[9] C. Matarasso, "Characterisation of World Wide Web Users Behaviour", paper presented at ECMAST'99, Madrid, 26-28 May 1999.

[10] C. Matarasso, F. Spataro, N. Bléfari-Melazzi, M. Lisanti, L. Secondiani, S. Vahid, B. Fan, "Criticality Analysis Result", ASSET deliverable AC326/DLR-MAT/DR-P/003/b1.

[11] A. Erramilli, O. Narayan, W. Willinger, "Experimental Queueing Analysis with Long-Range Dependent Packet Traffic", submitted to IEEE/ACM Transactions on Networking, 1994.

[12] V. Paxon, "Fast Approximation of Self-Similar Network Traffic", Lawrence Berkeley Laboratory and EECS Division internal report LBL-36750.

[13]W. Willinger, M. S. Taqqu, R. Sherman, D. V. Wilson, "Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level", IEEE/ACM transactions on Networking 5(1) pp.71-86, February 1997

Location Area Management for Mobile Satellite Systems Applying Diffusion Mobility Model

Gabriel Chávez, David Muñoz
Centro de Electrónica y Telecomunicaciones
Instituto Tecnológico y de Estudios Superiores de Monterrey
Suc. Correos J, 64849 Monterrey, N.L. México

Abstract - *In MSS (Mobile Satellite Systems) as in terrestrial systems, the importance of reducing different costs due to location area updating and paging procedures is a relevant issue. For incoming calls, factors like the delay generated by excessive paging steps could affect Quality of Service. In this work, a novel paging scheme and location area design for MSS is presented. The spot beam selection process is applied to the paging scheme and location area definition. Results are assessed in terms of paging success rates, paging-registration costs, delay-cost ratio as well as total signaling load.*

A simple mathematical formulation in terms of a diffusion process is presented. Diffusion constants are derived for subscriber classes. Results are presented and discussed for different mobility scenarios. This formulation enables us to analytically derive the paging and registration rates and in this way minimize the cost of the system through the definition of optimum location areas.

I. INTRODUCTION

The anticipated development of MSS is based on their capability in managing global coverage systems with a variety of users on land, sea, or air. MSS represent a choice to offer communications services to scarcely populated lands of broad areas where the implementation of a terrestrial mobile network is not possible to be deployed or the cost would become excessive. To make this possible, first, we have to address the existence of more bandwidth, higher frequency bands and much larger satellite constellations that, in the near future these networks will achieve.

Problem Definition

In mobile networks, the location of a Mobile Terminal (MT) has to be known by the network in order to properly route incoming calls. In cellular systems, for example, the system coverage area is divided into Location Areas (LA), which are collections of cells where a code (unique to this LA) is broadcast. The network keeps track of the current LA of every "switched on" MT. When a MT crosses the boundary between two LAs, it conducts a location update to inform the network of its LA change, thus the paging area is equal to the LA. The design of the LA involves a trade-off between the location update signaling (high for a small LA) and the paging signaling (high for a large LA).

This work has been partially sponsored by Nortel Networks .

II. PAGING AND LOCATION MANAGEMENT

The paging and location management models for MSS depend on characteristic parameters such as users' mobility indexes, population density, velocity rate, diversity aspects, desired Guaranteed Coverage Areas (GCA), etc.

Adequate mobility management provides the network with an efficient way of reaching active user terminals. When a terminal is turned on it must first register itself with the network. To do so, the terminal provides its current 'location' or 'position.' In [1], different mobility management approaches are described in some detail. As a user terminal travels, further location updates may be required during its active time depending on the distance it travels. Finally, when a user terminal is switched off, it should de-register itself from the network. This eliminates the possibility of the network paging a user terminal which is no longer responsive.

The GCA of a Fixed-Earth Station (FES) is proposed in [2]. This approach defines the service area of an earth station as the area around the FES where the terminal can connect to the FES through at least one satellite that is above the minimum elevation angle of the terminal.

Location Area and its Update. A location area is formed by an FES's instantaneous coverage. An MT will perform location update only if it loses the FES location area broadcast channel. From the network viewpoint, the location of the MT is "somewhere within reach of the FES." The FES (either on its own or with the help of the MT) may provide a sensible way of reducing the area over which it pages, in the event of an incoming call.

If an MT is in overlapping FES coverage and location updates to an FES only intermittently cover its location (because the MT is not in the FES's GCA), it will lose that FES's location area broadcast after a while and be forced to location update to another FES which properly covers its location.

Paging Implementation Options. As in terrestrial segments, any incoming call will be routed to the FES, which can guarantee that the MT, if it is working, is within the area that the FES is covering with its broadcast channel. With the FES in its simplest form, the FES would then transmit a paging message for the mobile through every spot beam in the GCA. If incoming calls occur infrequently, this may be acceptable, but otherwise, paging through these many spot beams is considered a signaling load and a waste of power and spectral resources.

A first approach to intelligent paging could be the FES identifying and recording the instantaneous size, shape and location (latitude and longitude) of the spot beam with which the MT last made contact (for location update, call set-up or any other reason) with a time-stamp. In the event of an incoming call, the FES would page only those spot beams required to completely cover the recorded area in the first instance. In case the mobile does not respond, the paging is repeated over a wider area, depending on the age of the time stamp and on the MT's mobility profile.

III. MOBILITY BASED ON DIFFUSION MODELS

When modeling location performance of MSS, some characteristic mobilities are assumed to be known. It is useful if these assumptions impose few restrictions and lead to analytically tractable models. In this paper we present a mobility model based on a diffusion function that is used to determine a location probability for a given area.

Given that the last unit's contact with the network occurred at time $t = t_0$ and place $x(t_0)$, the time varying distribution on location is described as $p_j(t, t_0, x(t_0))$, where j represents the unit locations. We assume that $p_j(t, t_0, x(t_0)) = \delta_{j,x(t_0)}$ where $\delta_{i,j}$ is unity for $i = j$ and zero otherwise [10]. The implication is that successful paging, registration or calls initiated by the mobile inform the system of the unit's location and then reset at t_0.

It is assumed that the paging process is much faster compared to the rate of a mobile's motion; that is, the mobile to be found does not change location during the paging process. This is because of the dimension of spot beam areas. Thus, if a page request arrives at time $t \geq t_0$, the distribution is assumed fixed at $p_j(t, t_0, x(t_0)) = p_j$.

Any mobile motion model could, in principle, be reduced to a corresponding time varying probability distribution. As a convenient, useful, and concrete example, we consider an isotropic motion process with drift [3,4]. Drift is defined as mean velocity in a given direction and can be used to model directed traffic such as vehicles along a highway. Zero mean velocity could be used to model a pedestrian motion.

A diffusion process is described in Equation (1). In this instance, we are interested in the two-dimensional case to view the context of the problem.

$$p_{x(t),x_0(t)}(\alpha,\beta) = \frac{1}{2\pi D t(1-\rho^2)} \exp\left[-\frac{(\alpha^2 - 2\rho\alpha\beta + \beta^2)}{2Dt(1-\rho^2)}\right] \quad (1)$$

Where D is the diffusion constant, t is the time since last contact was made to the MT [hrs.], ρ is the correlation coefficient $|\rho| \leq 1$ and r is the uncertainty radius [km].

Considering the polar transformation, the equation becomes the following:

$$P_{r,\theta}(r,\theta) = \frac{r}{2\pi D t(1-\rho^2)} \exp\left[\frac{-r^2(1-2\rho\sin\theta\cos\theta)}{2Dt(1-\rho^2)}\right] \quad (2)$$

Context for Mobility Model

As subscribers can exhibit a wide rage of mobility patterning, it may be convenient to group them into classes depending on the predictability of their daily routine. The system could treat each class differently to minimize the cost to the system. Users may be assigned to a class based on their past call history. Three possible classes include the following [5]:

- *Deterministic Users.* Follow a daily routine that the system knows.
- *Quasi-deterministic Users.* The time intervals vary slightly from day to day They may alternate routes on the way.
- *Random Users.* Display no orderly behavior whatsoever.

These three classes could be grouped into correlation ranges, and a mobility factor is assigned to a spread value.

We propose obtaining different values for the *diffusion constant* D, depending on the subscriber profile. That is, we want to find a *diffusion coefficient* $D=2Dt$ which, given an estimated level of *drift velocity*, can assure a confidence interval equal to 95% of probability success.

The calculus for confidence interval I is made over a region covering 95 percent of the diffusion *pdf,* as shown in Figure 1.

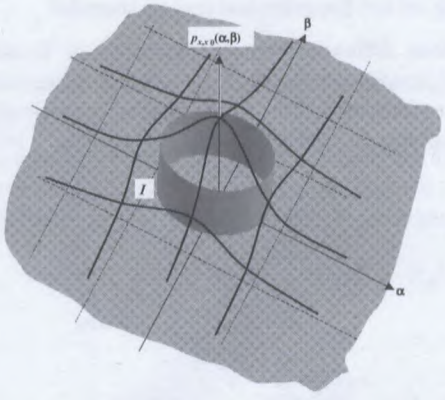


Figure 1 Confidence interval I for a two-dimensional diffusion pdf.

This comes from (1) with correlation parameter ρ set to zero.

$$P_c = \int_I p_{x(t)}(\alpha)\, d\alpha = 0.95,$$

$$= \iint_{I} \frac{1}{2\pi Dt} e^{-(\alpha^2+\beta^2)/2Dt} \, d\alpha \, d\beta, \qquad (3)$$

By making the change of variables, as in (2), the integration is easily carried out.

$$P_c = \int_0^d \int_0^{2\pi} \frac{r}{2\pi D\tau} e^{-r^2/2D\tau} \, dr \, d\theta = 1 - e^{-d^2/2D\tau}, \qquad (4)$$

The following shows the solution for diffusion constant D:

$$D\tau = \frac{d^2}{2\ln\left(\frac{1}{1-P_c}\right)}, \qquad (5)$$

The choice of probability criterion P_c affects the quality of service (1- P_c) represents the probability of an unsuccessful paging. If a user happens to be in this area when a call arrives, the call is lost. P_c can be chosen to be large enough so that the probability of being outside the paging-coverage area is negligible. However large a P_c means a larger PA and thus a higher cost. In fact, $P_c = 1$ implies an infinite paging cost because every location must be searched [6].

Since the diffusion constant depends on distance d and the *mean elapsed time between calls* is τ, it is possible to find a value which represents the mobility features of each user.

For example a mean radial shift d of 400 km is selected for high mobility users and an estimate of 10 hours as the *mean elapsed time* τ, according to [7]. Even though different values of τ will be analyzed, we aim to get the optimum value which minimizes total signaling load L in a location area. For medium mobility patterns, a mean radial shift value of 180 km is proposed, and for low mobility users a suitable value of a 50 km radial-shift is taken. These values represent a measure of mobility according to logical assumptions. However, other parameters can be considered, depending on the geographical region attended.

We propose different scenarios (shown in Table 1), for which *diffusion constants* were found. The correlation parameter indicates the user classes described, where correlation values near zero enclose random subscribers (unknown displacement direction). Quasi-deterministic classes are enclosed by a correlation range of 0.25 to 0.75 (there exists information about certain places of displacement), as is shown in Figure 2 below. A correlation equal to one is taken for deterministic mobility classes.

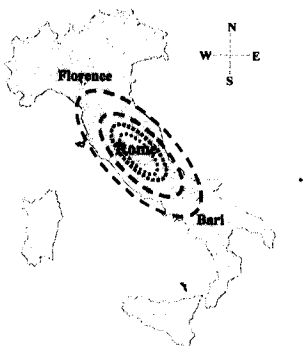


Figure 2 Example of correlation displacement in the diffusion *pdf*.

Table 1 Diffusion and correlation values for mobility scenarios

Scenario	Difussion value	Correlation ρ	Proportion Likelihood %
1	High	Deterministic	≈ 0
2	High	Semi-random	15
3	High	Random	≈ 0
4	Medium	Quasi-deterministic	15
5	Medium	Semi-random	10
6	Medium	Random	15
7	Low	Quasi-deterministic	≈ 0
8	Low	Semi-random	20
9	Low	Random	15

The proportion of the subscribers is likely a parameter that must be calculated according to the mobility profile of subscribers in the system (or geographical regions attended by the FES). Because some scenarios are more realistic than others (we do not expect high mobility with a displacement correlation near zero, for example) we assigned a greater percentage of that which could enclose the majority of subscribers.

Scenario 1 becomes a deterministic location procedure. As a specific example, consider highly directed and planned motion such as that of an aircraft or commuter train. If these conveyances follow exact courses at constant speeds, then the diffusivities would be identical to zero and the departure time and current time would be a completely determined location. In reality, the average arrival time is a random variable for a trip of any reasonable duration. We therefore try to infer the parameter of the equivalent diffusion model from the trip distance, mean duration and arrival-time variation.

IV. ANALYSIS AND RESULTS

In the event of an incoming call, the FES pages only those spot beams required to completely cover the recorded area in the first instance. In case the mobile does not respond, the paging is repeated over a wider area, depending on the age of the time stamp and the MT profile. Otherwise, the procedure is to analyze the spot beams and find which perform the paging success better, although this could imply a sequential paging and consequently increase the delay generated. This trade-off offers different location performance/cost choices from which operators or users can select.

Equating Probability of a Successful Paging

The mobile location at the time t_0 of last contact is assumed unknown and it can be considered randomly distributed on the Uncertainty Area (UA). At the time of an incoming call

$t_0 + t$, the estimated probability that the MT still places on the UA is as follows:

The diffusion process is characterized as time dependent *pdf*. Describing the mobility pattern for the subscriber classes, we obtain the *diffusion constant* for each mobility scenario, resolving (7) to obtain the probability of paging success given at any point over the UA, as follows:

$$P_{x(t),y(t)}(\alpha,\beta) =$$

$$\frac{1}{2\pi Dt\sqrt{1-\rho^2}}\exp\left[-\frac{(\alpha-x_0)^2-2\rho(\alpha-x_0)(\beta-y_0)+(\beta-y_0)^2}{2Dt(1-\rho^2)}\right], \quad (7)$$

Integrating (7) over the uncertainty region, we obtain the following;

$$P_{PS|a,\varphi}(r,\theta) =$$

$$\int_0^{2\pi}\int_0^{r'}\frac{r}{2\pi Dt\sqrt{1-\rho^2}}\exp\left[\frac{-r^2(1-2\rho\sin\theta\cos\theta)}{2Dt(1-\rho^2)}\right]dr\,d\theta, \quad (8)$$

$$=\frac{\sqrt{1-\rho^2}}{2\pi}\int_0^{2\pi}\frac{1}{1-\rho\sin 2\theta}\left(1-\exp\left[\frac{-r'^2(1-\rho\sin 2\theta)}{2(1-\rho^2)Dt}\right]\right)d\theta, \quad (9)$$

where,

$$r' = a\cos(\theta-\varphi)+\sqrt{a^2\sin^2(\theta-\varphi)+R^2}$$

Which can be solved numerically.

To determine the probability of paging success over the entire UA, it is necessary to consider all the points in the region. Considering a *probability density function* $p(a,\varphi)$ over the UA, we see the following:

$$E\left[P_{PS|a,\varphi}\right] = \int_{\varphi}\int_{a} p(a,\varphi)P_{PS|a,\varphi}\,da\,d\varphi \quad (10)$$

Figure 3 shows the probability of paging success averaged over the UA versus the call-arrival rate for different displacement correlation factors in the MT.

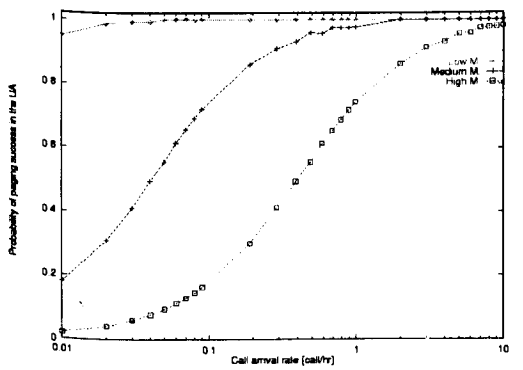


Figure 3. Probability of paging success averaged over the UA for low, medium and high mobility classes

Finally, obtaining a mean value of paging success rates for the mobility scenarios, enabling us to derive the definition of optimum location areas, which minimizes the system's cost.

$$E_{scenario}\left[P_{PS|m.s.}\right] = \sum_{i=1}^{9} p_{m.s.}(i)\cdot P_{PS|m.s}. \quad (11)$$

Equation (11) derives the mean value of probability of paging success in conjunction with the proportion of subscribers in the different scenarios described in Table 1.

Signaling Load

The paging rate yielded in two phases is the result of a multiple-step paging strategy. The first phase of paging yields the first term in (12), where the surface of the UA is multiplied by the user density (=UA *density*) and the incoming call rate λ_{in}. The second term represents the paging rate caused by the second paging phase, when the entire LA has to be paged in order to reach the subscriber that has moved out of the UA. $E[P_{PF|ms}]$ represents the mean value of the probability of paging failure over all mobility scenarios, resulted from (11). This term is finally multiplied by the rate of incoming calls in the entire LA and the proportion of user movement, as follows:

$$\lambda_p = \lambda_{in}\cdot\rho\cdot E[A_{PA}(t)] + E[P_{PF|ms}]\cdot\lambda_{in}\cdot\rho\cdot p_m\cdot\pi\cdot R_{LA}^2 \quad (12)$$

$$E\left[P_{PF|ms}\right] = 1 - E_{scenario}\left[P_{PS|ms}\right]$$

$E[A_{PA}(t)]$ is the average size of the paging area. We define this area as the area of the spot beam that touches the UA.

Defining the location update rate by characterizing the flow of traffic as a diffusion process in the same way as mentioned in [7], a similar analysis for MSS was obtained in the location update rates applied to a fluid model.

$$\lambda_{LU} = \frac{l\rho v}{\pi}.$$

Where l is the length of a location area perimeter, and v is the average velocity.

In order to evaluate the above equation, we define a location area of a circular area. With a single subscriber we have $\rho = 1/A$, and the average velocity is the variable to evaluate.

$$\lambda_r = \bar{\lambda}_{m.s.}\cdot\rho\cdot p_m\cdot\pi\cdot R_{SB}^2, \quad (13)$$

where,

$$\bar{\lambda}_{m.s.} = \sum_{\substack{mobility\\ scenarios}} cr\cdot p_{m.s.}$$

An average of velocity per mobility-scenario proportion gives an approximate rate of location update per speed of user.

Minimizing the Signaling Load

The overall signaling cost for the location area has the following additive structure of equations (12) and (13):

$$L = S_p \lambda_p + S_r \lambda_r, \qquad (14)$$

Where S_p and S_r are cost coefficients for paging and registration [signaling units/Event] respectively. A ratio of cost S_r/S_p =6.3 is assigned, accoding to [2]. The response message from the MT (and the following call set-up) is irrelevant for the optimization of location management schemes because location management schemes do not influence the number of paging requests (i.e. the number of incoming calls).

For each value of the spot beam radius and traffic parameters, there is an optimum location area value R_{LA} that corresponds, respectively, to a minimum of signaling load L. This is a tradeoff between a large R_{LA} that would cause a lot of paging and few location updates and a small one with the opposite effect.

Figure 4 depicts the paging, registration and total number of signaling loads versus the location area radius. The signaling loads are measured in signaling units per second [sups]. Where a user density ρ = 1.2 km^{-2} and a spot beam radius equal to 300 km has been selected, and the proportion of subscriber p_m is equal to one. The dashed lines represent the paging (increasing curve) and location update (decreasing curve) components. The registration does not depend on the incoming call rate, but only of the subscribers' mobility. In contrast, the paging rate is affected for both parameters.

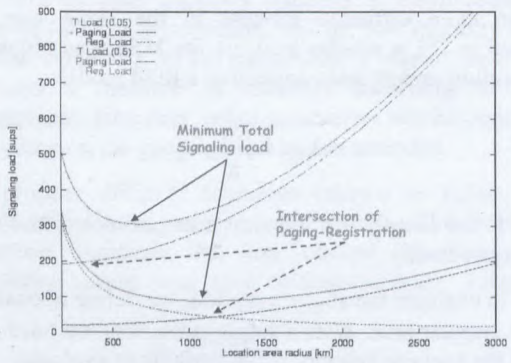


Figure 4. Minimum total signaling load for λ_{in} = 0.05 and 0.5 [call/hr.]

We observe that as incoming call rates increase the optimum location area size grows. Note that E[$P_{PF|ms}$] on Equation (12) tends to be zero as $\lambda_{in} \to \infty$ this is due to a more frequent registration rate. This is an important assumption because the goal of the systems providers is an increment of the arrival call rate. Because of all this, the definition of optimum location area is an important factor, as well as the paging procedure. In order to minimize the minimum signaling load, we take into consideration the following parameters:

- The mobility to call ratio.
- The registration-to paging-cost ratio.

The above two quantities are relevant because the optimum location area increases as either of the two quantities increases.

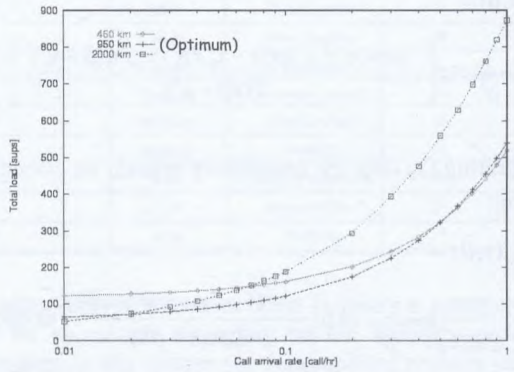


Figure 5. Total signaling load for fixed location area definition.

Figure 5 show three possible values of "optimum" location area radii. If we suppose an incoming call rate sub estimate, we will obtain a value of location area radius bigger than what should be obtained (e.g. a 2,000 km. radius is shown). Then if the incoming call rate grows in an unexpected way, the total cost will increase significantly. From the traces shown above, a value of 950 km. as a radius of LA represents an improvement in respect to others in the range of incoming call rates.

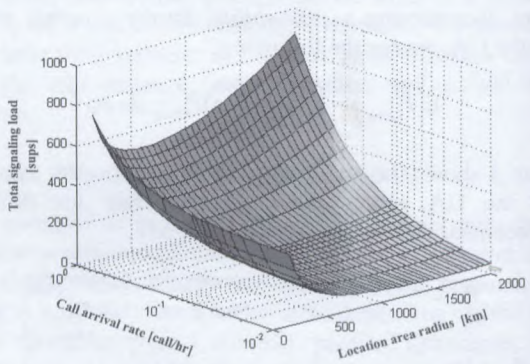


Figure 6 Total LA signaling load versus LA radius and call arrival rate.

Figure 6 illustrates a wide range for the call arrival rate, with a location area radius whose value presents a minimum of signaling cost for the system.

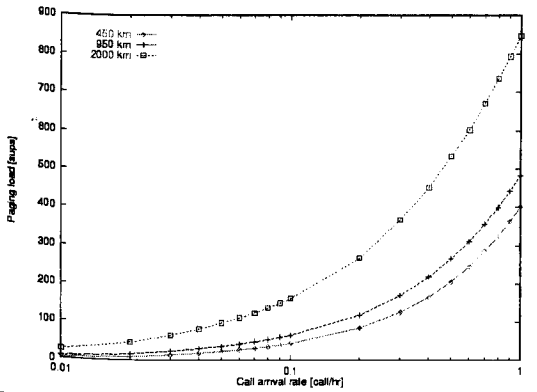


Figure 7 Increase of paging load due to call arrival rates.

Figure 7 shows the importance of proposing and selecting a paging procedure according to the characteristics of the system and, in this way, minimizes the signaling cost of the system.

V. CONCLUSIONS

In this work, a methodology for obtaining the cost of paging success using diffusion mobility model is presented. The result shows the flexibility of the mobility model enabling us to analytical derive a formulation to estimate paging success rates. The diffusion process as a tool of modeling the mobility of subscribers provides improvement in the analysis of location management. The Location Area Criterion LAC represent an alternative way to obtain the paging cost in an MSS, and in this way minimization of the system paging cost is achievable through the definition of optimum location areas. The advantages are listed as follows:

- LAC has the advantage of requiring smaller information network for the MT position.
- The numerical calculus to obtain the probability of paging success in order to evaluate whether the UA is paged on the first paging step or is realized in a single-step paging strategy.

The criteria represents a novel formulation in paging procedure for MSS's improved results, and these should be compared with other works.

ACKNOWLEDGMENT: Authors thank Payan Maveddat and Gerardo Donnis from Nortel Networks for multiple comments and suggestions.

REFERENCES

[1] Cullen C., Sammut A., Taffazolli R. and Evans B., "Networking and Signalling Aspects of a Satellite Personal Communication Network," *1st European Workshop on Mobile/Personal Satecomms (EMPS'94), Frascati,* 1994.

[2] Sammut A., Cullen C. and Tafazolli R., "Mobility Management Related Signaling for a MAGS-14 Based Satellite Personal Communications Network," *National Technology University of Athens Greece,* 1994.

[3] Papoulis A., *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill Inc., 1984.

[4] Feller W., *An Introduction to Probability Theory and its Applications*, John Wiley and Sons, Inc., 1971.

[5] Pollini G. P. and Chih-Lin I, "A Profile-Based Location Strategy and Its Performance," *IEEE JSAC,* Vol. 15, No 8, October 1997.

[6] Zhuyu L. and Rose C., "Probability Criterion Based Location Tracking Approach for Mobility Management of Personal Communications Systems," *IEEE Global Telecommunications Conference, Phoenix, AZ.*, November 1997.

[7] Hoiydi El A. and R. J. Finean, "Location Management for the Satellite - Universal Mobile Telecommunication System," *IEEE Int'l Conf. Universal Pers. Commun., Cambridge, Massachusetts,* Sept. 1997.

[8] Finean R.J., et al, "Impact of satellites on UMTS Network," *RACE Mobile Telecommunications Summit,* Vol. 2, 1995.

[9] Meenan C., Taffazolli R. and Evans B., "New Mobility Management and Call Routing Protocols for Dynamic Satellite Personal Communication Networks," 1998.

[10] Rose C., Yates R., "Location Uncertainty in Mobile Networks: A Theoretical Framework," *IEEE Communications Magazine,* February 1997.

Frequency Reuse Impact on the Optimum Channel Allocation for a Hybrid Mobile System

Tamer A. ElBatt, Anthony Ephremides
Electrical Engineering Department,
University of Maryland,
College Park, MD 20742, USA.
Email: telbatt@eng.umd.edu, tony@eng.umd.edu

ABSTRACT

In this paper we study the effect of the frequency reuse constraints in both layers on the optimum channel allocation for a multi-cell/multi-spot- beam hybrid system. We adopt a specific multi-faceted cost function that incorporates call-dropping due to unsuccessful hand-off attempts, and blocking of new calls. The minimization of the cost function is attempted by choosing the optimal split of the total number of channels between the cellular and the satellite layers. This complex optimization problem is solved by means of standard clock simulation techniques along with the adaptive partitioned random search global optimization technique and the ordinal optimization approach.

I. INTRODUCTION

Future mobile communication systems are expected to use land mobile satellite systems to enhance terrestrial cellular service. Recent studies on integrated satellite-terrestrial networks emphasize using satellites to provide "out-of-area" coverage to mobile users. However, with recent developments in satellite technologies, such as narrow beam antennas and switchable spot-beams for LEO and GEO systems, satellites can be used effectively to off-load localized congestion within the underlying cells.

In pure cellular networks, earlier studies have shown that efficient use of the system bandwidth can be achieved by reuse partitioning [1] and using hierarchical cell layout [2],[3] with larger macrocells overlaying small microcells. In [2], the authors applied the concept of *cluster planning*, via which the proposed sectoring arrangement allows microcells to reuse macrocell frequencies. This in turn achieves higher system capacity. However, users' mobility and hand-offs were not considered in that model. The problem of finding the optimum partitioning of the frequency spectrum between microcells and macrocells was also addressed in [3]. This work differs from our work in two aspects. First, the call assignment policy was assumed to be speed dependent. Second, identical frequency reuse patterns in the microcells and macrocells are assumed. Performance analysis of a hybrid satellite-cellular system with the satellite foot-prints forming the highest layer in the hierarchy was also studied [4]. However, the reuse profile for the satellite system was assumed to be the same as that for the terrestrial system.

This work is along the same line of the work done in [5]. It builds upon our earlier work [6] in which the frequency reuse effect in both layers was not considered in the model, but rather, only the propagation delay effect was considered. More specifically, we introduce a multi-dimensional Markov chain-based model for a hybrid network consisting of multiple cells overlaid by multiple spot-beams. In [6], we focused on showing the trade-off and solving the problem for a simple system of just two cells overlaid by one spot-beam. Here, we are extending the model to a more realistic case of multiple cells and spot-beams. It is worth mentioning here that the solution approach developed in [6] still holds, assisted by the Adaptive Partitioned Random Search (APRS) global optimization technique[7]. Our prime concern is to show how the optimum channel partitioning between the cellular and the satellite layers is affected by the frequency reuse constraints.

The paper is thus organized as follows: In section II, system assumptions and the mathematical model are given. This is followed by the problem formulation in section III. In section IV, the optimization approach is illustrated. Simulation results are given and discussed in section V. In section VI, the study is extended to large hybrid systems. Finally, the conclusions are drawn in section VII.

II. SYSTEM DESCRIPTION

A. Assumptions and Definitions

In order to investigate the frequency reuse impact on the optimal channel partitioning policy, we first make the following assumptions and introduce appropriate notation. The network under consideration consists of 8 cells, namely C_1, C_2,..., and C_8. In addition, there is a satellite emitting 4 spot-beams S_1, S_2, S_3, and S_4 covering the same area, and supported by on-board switching as shown in Figure 1. New calls arrive at cell C_i according to a Poisson arrival process with rate λ calls/min. The duration of each call is assumed to be exponentially distributed with mean $1/\mu$ min.

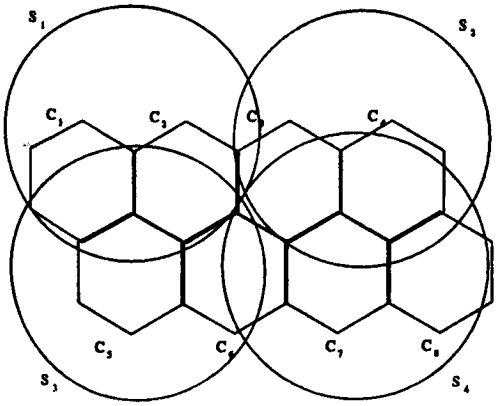


Figure 1. A Hybrid Mobile System of 8 Cells overlaid by 4 Spot-beams

We define K_s as the satellite reuse factor, that is the number of spot-beams per cluster, where all the spot-beams in a cluster use distinct frequency sets. Likewise, define K_c as the cellular reuse factor. Define K as the *relative reuse factor*, that is the ratio of the satellite reuse factor to the cellular reuse factor. The total number of *predesign* resources available to the system is M, where,

$$M = \Sigma_{i=1}^{K_c} M_i + \Sigma_{j=1}^{K_s} M_{s_j}$$

and,
M_i = number of channels dedicated to cell C_i.
M_{s_j} = number of channels dedicated to spot-beam S_j.

Define P_{ij} as the probability of assigning a call with one of the parties in cell C_i and the other in cell C_j to the nearest cells. Also, define (1-P_{ij}) as the probability of assigning a call with one of the parties in cell C_i and the other in cell C_j to the overlaying spot-beam(s), where i,j = 1,2,....,8, i $\leq$ j. Using this assignment rule, we restrict call routes to pure terrestrial and pure satellite routes, i.e. no hybrid routes. Nevertheless, it is straightforward to extend this assignment rule in order to take hybrid routes into consideration. It contributes to more call types and hence increases the dimensionality of the problem. We assume that the base stations, namely BS_i, i=1,...,8, communicate via a terrestrial wireline infrastructure. According to this assumption, each mobile-to-mobile call needs 2 duplex channels. A mobile user can access the satellite directly, not through its BS, using a *dual mode* satellite/cellular mobile terminal. All call types have the same priority and all calls considered in this model are mobile-to-mobile calls. BSs and spot-beams are assumed to be stationary. We define f as the fraction of calls that originate in a cell and are destined to any other cell. The interhand-off time of a mobile from cell C_i to a neighboring cell C_j is assumed to be exponentially distributed with mean $1/\lambda_h$ min, i,j=1,2,...,8. Accordingly, the interhand-off time of a mobile from spot-beam S_l to spot-beam S_k is also exponentially distributed with mean $1/\lambda_{h_s}$ min, where the handoff-rate is assumed to be *inversely* proportional to the cell/spot-beam radius and l,k=1,2,3,4. It is worth mentioning that the additional resources provided by the overlapping spot-beams in Figure 1 is not considered in this model. Finally, we assume that blocked calls are cleared.

B. System Model

The state of the system can be defined by the vector $(n_{11}, n_{12}, n_{13}, ..., n_{ij}, ..., n_{88}, n_{s_{11}}, n_{s_{12}}, ..., n_{s_{kl}}, ..., n_{s_{44}})$, where i,j = 1,2,3,...,8, i $\leq$ j and l,k = 1,2,3,4, l $\leq$ k. n_{ij} is the number of active calls of type 'ij'; that is, calls served by BS_i and BS_j, where one of the parties is in C_i and the other is in C_j. On the other hand, $n_{s_{kl}}$ is the number of active calls of type 's_{kl}'; that is, calls served by spot-beams S_k and S_l, where one of the parties is within foot-print 'S_k' and the other is within foot-print 'S_l'.

Accordingly, the system is modeled as a Continuous-time Markov Chain of 46 dimensions representing each call type. It is worth mentioning that all call types need 2 wireless channels/call. Therefore, the set of feasible states should satisfy the following state space constraints :

$$2n_{11} + n_{12} + n_{13} + n_{14} + ... + n_{18} \leq M_1$$

$$n_{12} + 2n_{22} + n_{23} + n_{24} + ... + n_{28} \leq M_2$$

$$n_{13} + n_{23} + 2n_{33} + n_{34} + ... + n_{38} \leq M_3$$

$$n_{18} + n_{28} + n_{38} + n_{48} + ... + 2n_{88} \leq M_8$$

$$2n_{s_{11}} + n_{s_{12}} + n_{s_{13}} + n_{s_{14}} \leq M_{s_1}$$

$$n_{s_{14}} + n_{s_{24}} + n_{s_{34}} + 2n_{s_{44}} \leq M_{s_4}$$

III. PROBLEM FORMULATION

The optimum channel allocation policy for a given call assignment rule and relative frequency reuse factor is obtained by solving the following minimization problem:

$$\min_{M_1, M_2, ..., M_8, M_{s_1}, M_{s_2}, ..., M_{s_4}} (P_b + \alpha . P_d) \qquad (1)$$

s.t.

$$M = \Sigma_{i=1}^{K_c} M_i + \Sigma_{j=1}^{K_s} M_{s_j}$$

where,
P_b = average new call blocking probability.
P_d = average call dropping probability.
α = weighting factor.

In the above formulation, the choice of the design parameter α is rather unguided, since there is no well-defined procedure for choosing it. The following formulation is equivalent and easier to implement. It consists of minimizing one component of the composite cost function above subject to the other component staying below a

pre-determined acceptable threshold, namely,

$$\min_{M_1,M_2,\ldots,M_8,M_{s_1},M_{s_2},\ldots,M_{s_4}} P_b \qquad (2)$$

s.t.

$$P_d \leq \beta$$
$$M = \Sigma_{i=1}^{K_c} M_i + \Sigma_{j=1}^{K_s} M_{s_j}$$

The quantity β is the alternative (equivalent) parameter in a one-to-one correspondence to the value of α.

IV. OPTIMIZATION APPROACH

Due to the sheer complexity of jointly optimizing the channel allocation and the call assignment policy[6], we chose here to solve for the optimum channel split between the satellite and the cellular layers given a call assignment policy. The formulation of the problem given in section III can be solved via *Discrete Exhaustive Search*. This optimization approach is not only complex, but infeasible as well. This is due to the large dimensionality of the Markov Chain, which in turn leads to an extremely large pool of channel allocation policies.

The numerical solution was infeasible, too, due to the Markov Chain being of 46 dimensions. Consequently, we had to resort to simulation. A simulation process was developed using C++ and run on SUN-ULTRA 1/2 workstations. To increase the efficiency of the simulation, we employed the so-called standard clock (SC) simulation method. Its basic principles are explored in more details in [9]. The essence of SC simulation is that it allows the simultaneous measurement of performance of multiple different control policies with a single simulation run. Moreover, as we are more interested in the relative ranking of the channel allocation policies, rather than in their actual performance values, and to further speed up simulations, Ordinal Optimization was employed. Ordinal Optimization has been applied in the literature using several approaches, namely short simulation runs, crude analytical models, and simplified, but imprecise simulation models [8]. In [6], we concluded that ordinal optimization, based on short simulation runs, is applicable to our problem. Accordingly, it is employed in this paper in conjunction with SC simulation.

As indicated earlier, the search space for this problem is very large. Hence, it is infeasible to search for the optimum in one phase. Therefore, a tree-search type of algorithms is employed. According to [7], the search region of the objective function is to be partitioned into certain number of sub-regions. Then, using the sampled function values from each sub-region, determine how promising each sub-region is. The most promising sub-region is then further partitioned. This global optimization technique is called the *Adaptive Partitioned Random Search* (APRS). The simulation results show that, while the APRS does not necessarily reach a global optimum, it is guaranteed to reach a near-optimal solution quickly. This is achieved at a computational cost much lower than discrete exhaustive search.

V. RESULTS

The hybrid network shown in Figure 1 was analyzed assuming the numerical parameters given in Table 1. It should be pointed out here that the following results were obtained with no constraint enforced on P_d while minimizing P_b, i.e. β was assumed to be 1 in (2).

Consider the problem of finding the optimum static channel split for a given call assignment policy and frequency reuse pattern. The optimum was determined for the frequency reuse factors given in Table 2 and the following call assignment probabilities:

$$P_{ij} = 0.5, \qquad i,j = 1,2,\ldots,8,\ i \leq j$$

For the first frequency reuse pair, i.e. $K_c = 4$, $K_s = 1$, the frequency reuse in the satellite layer was optimistic in the sense that neighboring, or even overlapping footprints, may use the same frequencies. This assumption is supported by the different satellite propagation characteristics which may permit a much denser frequency reuse pattern in the space segment.

Table 1. System Parameters

Total System Bandwidth (M)	40 channels
Call Arrival Rate per Cell (λ)	0.6 calls/min
Call Service Rate (μ)	0.6 calls/min
Call Hand-off Rate (λ_h)	0.5 calls/min
Fraction of calls originated in a cell and destined to any other cell (f)	0.125

Table 2. Frequency Reuse Factors

K_c	K_s	$K = \frac{K_s}{K_c}$
4	1	0.25
4	2	0.5
3	2	0.666
3	4	1.333

The development in antenna technology and careful evaluation of the propagation effects support this idea. On the other hand, the frequency reuse in the cellular layer was, relatively, conservative by assuming that each cell cluster has 4 cells. For this set of frequency reuse factors, the shared satellite resources assisted by the denser frequency reuse pattern in the space segment, give the superiority to the satellite layer. The pool of channel allocation policies is generally huge to search for the optimum in one phase, so the APRS global optimization technique was recommended to speed-up the search process as will be explained later. However, the spatial symmetry of the call arrival rates, service rates, and hand-off rates among the cells and spot-beams can be noticed from Table 1. Therefore, the search space was restricted to those policies having equal shares among cells and equal shares

among spot-beams, i.e. $M_{c_i} = M_c$, i=1,2,...,8, $M_{s_j} = M_s$, j=1,2,3,4. The simulation results shown in Table 3 indicate that the optimum policy (shown in bold font) is to assign all the resources to the satellite.

Table 3. Blocking and Dropping Performance of Channel Allocation Policies ($K_c = 4$, $K_s = 1$)

$(M_1, M_2, M_5, M_6, M_{s_1})$	P_b	P_d
(0,0,0,0,40)	**0.000007**	**0.000009**
(1,1,1,1,36)	0.000031	0.000020
(2,2,2,2,32)	0.000046	0.000027
(3,3,3,3,28)	0.000052	0.000036
(4,4,4,4,24)	0.000079	0.000057
(5,5,5,5,20)	0.000090	0.000070
(6,6,6,6,16)	0.000174	0.000080
(7,7,7,7,12)	0.000279	0.000183
(8,8,8,8,8)	0.001485	0.003517
(9,9,9,9,4)	0.019534	0.030689
(10,10,10,10,0)	0.080307	0.125289

Consider next the hybrid system having the second frequency reuse pair in Table 2, i.e. $K_c = 4$, $K_s = 2$. In this case, both layers have good, but not the best achievable frequency reuse patterns. Again, the shared capacity advantage of the space segment still wins and the *"All-Channels-to-Satellite"* allocation policy achieves the minimum blocking probability as shown in Table 4.

Table 4. Blocking and Dropping Performance of Channel Allocation Policies ($K_c = 4$, $K_s = 2$)

$(M_1, M_2, M_5, M_6, M_{s_1}, M_{s_2})$	P_b	P_d
(0,0,0,0,20,20)	**0.000029**	**0.000016**
(1,1,1,1,18,18)	0.000040	0.000028
(2,2,2,2,16,16)	0.000053	0.000044
(3,3,3,3,14,14)	0.000090	0.000053
(4,4,4,4,12,12)	0.000119	0.000061
(5,5,5,5,10,10)	0.000596	0.000708
(6,6,6,6,8,8)	0.002138	0.003385
(7,7,7,7,6,6)	0.007791	0.013216
(8,8,8,8,4,4)	0.020592	0.024475
(9,9,9,9,2,2)	0.052548	0.079817
(10,10,10,10,0,0)	0.081292	0.126263

For the third frequency reuse set, i.e. $K_c = 3, K_s = 2$, we assume an optimistic reuse pattern for the terrestrial layer. On the other hand, a good reuse pattern (but not the best) is assumed for the satellite. We expect that the best frequency reuse in the terrestrial layer might overcome the shared capacity advantage of the satellite, and this what actually happens. The simulation results, see Table 5, show that the optimum allocation policy, in terms of minimizing the blocking probability, is $M_1 = 2$, $M_2 = 2$, $M_3 = 2$, $M_{s_1} = 17$, $M_{s_2} = 17$.

Table 5. Blocking and Dropping Performance of Channel Allocation Policies ($K_c = 3$, $K_s = 2$)

$(M_1, M_2, M_3, M_{s_1}, M_{s_2})$	P_b	P_d
(0,0,0,20,20)	0.000029	0.000016
(1,1,1,18,19)	0.000026	**0.000011**
(2,2,2,17,17)	**0.000023**	0.000020
(3,3,3,15,16)	0.000034	0.000027
(4,4,4,14,14)	0.000068	0.000037
(5,5,5,13,12)	0.000080	0.000051
(6,6,6,11,11)	0.000104	0.000091
(7,7,7,10,9)	0.000537	0.000310
(8,8,8,8,8)	0.001812	0.004610
(9,9,9,6,7)	0.005413	0.008814
(10,10,10,5,5)	0.008635	0.016808
(11,11,11,4,3)	0.017908	0.0042381
(12,12,12,2,2)	0.036162	0.073978
(13,13,13,1,0)	0.051285	0.125041

Table 6. Blocking and Dropping Performance of Channel Allocation Policies ($K_c = 3$, $K_s = 4$)

$(M_1, M_2, M_3, M_{s_1}, M_{s_2}, M_{s_3}, M_{s_4})$	P_b	P_d
(0,0,0,10,10,10,10)	0.017	0.011
(1,1,1,9,9,9,10)	0.009	0.012
(2,2,2,8,9,9,8)	0.008	**0.005**
(3,3,3,7,8,8,8)	0.006	0.0053
(4,4,4,7,7,7,7)	**0.0058**	0.006
(5,5,5,6,7,6,6)	0.009	0.011
(6,6,6,5,6,6,5)	0.011	0.016
(7,7,7,5,4,5,5)	0.015	0.023
(8,8,8,4,4,4,4)	0.019	0.027
(9,9,9,3,3,4,3)	0.031	0.051
(10,10,10,2,3,3,2)	0.038	0.067
(11,11,11,2,1,2,2)	0.044	0.086
(12,12,12,1,1,1,1)	0.055	0.11
(13,13,13,1,0,0,0)	0.063	0.136

Finally, the last set of frequency reuse factors, $K_c = 3, K_s = 4$, indicates that bandwidth partitioning will be the optimum allocation policy since the terrestrial network has the best achievable frequency reuse factor, while the satellite layer has the worst one. The simulation results for this case are given in Table 6. It can be noticed that the optimum channel allocation policy in this case is $M_1 = 4$, $M_2 = 4$, $M_3 = 4$, $M_{s_1} = 7$, $M_{s_2} = 7$, $M_{s_3} = 7$, $M_{s_4} = 7$.

In order to reach the previous results, we made use of the spatial symmetry of the call arrival rates, call service rates, and call hand-off rates in limiting the search space. We restricted the search process to those policies having equal shares among cells and equal shares among spot-beams. For the general case, the search space will be extremely large and it would be impossible to search for the optimum in one phase. Therefore, we recommend employing a tree search type of algorithms, like the APRS global optimization technique. We applied this optimization technique on our system with $K_c = 4, K_s = 2$, and the same numerical parameters given in Table 1. To verify

our earlier results, shown in Table 4, we resolved the optimization problem without taking into account the spatial symmetry. Instead, we searched for the optimum in the whole space of 1,221,759 policies. In this case, the space of channel allocation policies was 6-dimensional. The search space was partitioned to 12 regions in each phase and a sample policy was picked randomly from each partition according to a uniform distribution. The partitioning was performed using hyperplanes parallel to the space axes. In each search phase, we marked the partition having the policy that gave the minimum bloking rate as the *"most promising"* partition, and it was partitioned further in the next phase. Tables 7 through 10 show the blocking and dropping performance of the sample policies in the four search phases performed. It should be pointed out that, in each phase, the *"most promising"* partition is shown in bold font.

It can be noticed from Table 10 that the partitioning process is approaching the optimum policy (0,0,0,0,20,20) given in Table 4. Therefore, we conclude that the APRS algorithm reaches a near-optimal solution quite fast as compared to exhaustive search. Hence, it is suitable for solving our complex optimization problem.

Table 7. Phase #1 ($K_c = 4$, $K_s = 2$)

Partition	P_b	P_d
(0-13,0-13,0-13,0-13,0-13,26-40)	0.016	0.035
(0-13,0-13,0-13,0-13,13-26,13-26)	**0.003**	**0.002**
(0-13,0-13,0-13,0-13,26-40,0-13)	0.117	0.095
(0-13,0-13,0-13,13-26,0-13,0-13)	0.036	0.029
(0-13,0-13,0-13,13-26,13-26,0-13)	0.215	0.184
(0-13,0-13,13-26,0-13,0-13,0-13)	0.013	0.012
(0-13,0-13,13-26,13-26,0-13,0-13)	0.054	0.054
(0-13,0-13,26-40,0-13,0-13,0-13)	0.222	0.530
(0-13,13-26,0-13,0-13,0-13,0-13)	0.069	0.122
(13-26,0-13,0-13,0-13,0-13,13-26)	0.012	0.020
(13-26,0-13,13-26,0-13,0-13,26-40)	0.050	0.140
(26-40,0-13,0-13,0-13,0-13,0-13)	0.070	0.174

Table 8. Phase #2 ($K_c = 4$, $K_s = 2$)

Partition	P_b	P_d
(0-6,0-6,0-6,0-6,13-19,13-19)	**0.000023**	**0.000017**
(0-6,0-6,0-6,0-6,19-26,13-19)	0.086920	0.051897
(0-6,6-13,0-6,0-6,19-26,19-26)	0.055596	0.037105
(0-6,0-6,0-6,6-13,13-19,13-19)	0.001575	0.001456
(0-6,0-6,6-13,0-6,13-19,19-26)	0.035220	0.041223
(0-6,0-6,6-13,0-6,19-26,19-26)	0.066947	0.067648
(0-6,6-13,0-6,0-6,13-19,13-19)	0.011133	0.015620
(0-6,6-13,0-6,6-13,13-19,13-19)	0.041521	0.064385
(6-13,0-6,0-6,0-6,13-19,19-26)	0.007141	0.013065
(6-13,0-6,0-6,0-6,13-19,19-26)	0.034991	0.069517
(6-13,0-6,0-6,6-13,19-26,13-19)	0.036735	0.107833
(6-13,0-6,0-6,6-13,13-19,13-19)	0.013944	0.032950

Table 9. Phase #3 ($K_c = 4$, $K_s = 2$)

Partition	P_b	P_d
(0-3,0-3,0-3,0-3,13-16,16-19)	0.000239	0.000111
(0-3,0-3,0-3,0-3,16-19,16-19)	**0.000019**	**0.000021**
(0-3,0-3,0-3,3-6,13-16,16-19)	0.000089	0.000037
(0-3,0-3,3-6,0-3,16-19,13-16)	0.000526	0.000638
(0-3,0-3,3-6,3-6,13-16,3-6)	0.000308	0.000160
(0-3,3-6,0-3,3-6,16-19,13-16)	0.000698	0.000765
(0-3,3-6,3-6,3-6,13-16,13-16)	0.002556	0.003067
(0-3,3-6,3-6,3-6,16-19,3-6)	0.024001	0.048456
(3-6,0-3,3-6,0-3,13-16,16-19)	0.000182	0.000369
(3-6,0-3,3-6,3-6,13-16,3-6)	0.007087	0.011455
(3-6,3-6,0-3,0-3,16-19,13-16)	0.003145	0.007962
(3-6,3-6,3-6,3-6,13-16,13-16)	0.001834	0.002189

Table 10. Phase #4 ($K_c = 4$, $K_s = 2$)

Partition	P_b	P_d
(0-2,0-2,0-2,0-2,18-19,18-19)	**0.000068**	**0.000027**
(0-2,0-2,0-2,2-3,16-18,18-19)	0.000093	0.000053
(0-2,0-2,0-2,2-3,18-19,18-19)	0.000117	0.000083
(0-2,0-2,2-3,0-2,18-19,18-19)	0.000250	0.000176
(0-2,0-2,2-3,2-3,18-19,16-18)	0.000430	0.000389
(2-3,2-3,0-2,2-3,16-18,18-19)	0.000749	0.000531
(0-2,2-3,2-3,2-3,18-19,16-18)	0.000081	0.000058
(0-2,2-3,2-3,2-3,16-18,16-18)	0.000641	0.001064
(2-3,0-2,0-2,2-3,16-18,18-19)	0.000085	0.000047
(2-3,0-2,2-3,2-3,16-18,16-18)	0.000102	0.000079
(2-3,2-3,2-3,0-2,18-19,16-18)	0.000096	0.000080
(2-3,2-3,2-3,2-3,18-19,16-18)	0.000101	0.000064

VI. LARGE HYBRID SYSTEMS

In this section, our objective is to emphasize the significance of the relative frequency reuse effect on the optimal channel allocation policy for hybrid networks reflecting practical environment. Therefore, we consider a network of 4 spot-beams overlaying 100 terrestrial cells and assume the numerical parameters given in Table 11. Our major concern is to demonstrate that partitioning the channels between the satellite and the cellular layers outperforms, under certain frequency reuse conditions, the *"All-Channels-to-Satellite"* allocation policy. This is due to the denser cellular frequency reuse factor, as compared to the satellite reuse factor, which in turn overcomes the shared capacity advantage of the space segment. For this large system, discrete event simulation is the only feasible performance evaluation approach. Therefore, simulation studies were conducted on the Object Oriented Hybrid Network Simulation (OOHNS)[11] testbed developed at the University of Maryland.

Table 11. Large Hybrid Network Parameters

Number of Cells	100
Number of Spot-beams	4
Total System Bandwidth (M)	100 channels
Call Arrival Rate	0.333 calls/min
Call Service Rate	0.333 calls/min
Cellular Frequency Reuse Factor (K_c)	3
Satellite Frequency Reuse Factor (K_s)	2

We compared the performance of two policies, namely policy π which allocates all the channels to the satellite and policy $\bar{\pi}$ which partitions the total number of channels equally between the satellite and the cellular layers. For the numerical parameters given in Table 11, the blocking and dropping probabilities for policy π turned out to be 0.052 and 0.06 respectively. On the other hand, the blocking and dropping probabilities for policy $\bar{\pi}$ are 0.009 and 0.006. From these results, we emphasize the major role the relative frequency reuse factor plays in the design of real hybrid systems.

VII. CONCLUSIONS

In this paper we studied the effect of the relative frequency reuse factor on the optimal static channel split for a multi-cell/multi-spot-beam hybrid network. The objective was to show how the optimal channel allocation policy is affected by varying the frequency reuse factors in both layers. This was achieved via minimizing a multi-faceted cost function composed of the call blocking and dropping probabilities for a given set of frequency reuse factors in both layers. We have shown, via simulations, that the optimal channel allocation policy is the *"All-Channels-to-Satellite"* policy if the terrestrial frequency reuse pattern is not dense enough to overcome the shared capacity advantage of the space segment. On the other hand, when the terrestrial reuse pattern is denser than the satellite reuse pattern, partitioning the channels between the two layers turns out to be the optimum policy. Therefore, it can be concluded that the relative frequency reuse factor plays a major role in the design of hybrid systems. Finally, we found out that our results carry for large hybrid networks reflecting practical environment.

REFERENCES

[1] J. Zander and M. Frodigh, "Capacity Allocation and Channel Assignment in Cellular Radio Systems Using Reuse Partitioning," In *Electronics Letters*, vol. 28, no. 5, pp. 438-440, Feb 1992.

[2] L. Wang, G. Stuber, and C. Lea, "Architecture Design, Frequency Planning, and Performance Analysis for a Microcell/Macrocell Overlaying System," *IEEE Transactions on Vehicular Technology*, vol. 46, no. 4, pp. 836-848, Nov. 1997.

[3] K. Yeung, and S. Nanda, "Channel Management in Microcell/ Macrocell Cellular Radio Systems," *IEEE Transactions on Vehicular Technology*, vol. 45, no. 4, pp. 601-612, Nov. 1996.

[4] L. Hu and S. Rappaport, "Personal Communication Systems using Multiple Hierarchical Cellular Overlays," *IEEE Journal on Selected Areas in Communication*, vol. 13, no. 2, pp. 406-415, Feb 1995.

[5] D. Ayyagari and A. Ephremides, "Blocking Analysis and Simulation Studies in Satellite-Augmented Cellular Networks," *Proceedings of the 7th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications PIMRC'96*, vol. 2, pp. 437-441, 1996.

[6] T. ElBatt and A. Ephremides, "Optimization of Connection Oriented, Mobile, Hybrid Network Systems," *IEEE Journal on Selected Areas in Communication*, vol 17, no. 2, Feb. 1999.

[7] Z. Tang "Adaptive Partitioned Random Search to Global Optimization," *IEEE Transactions on Automatic Control*, vol. 39, no. 11, pp. 2235-2244, Nov. 1994.

[8] Y. Ho, R. Sreenivas, and P. Vakili, "Ordinal Optimization of DEDS," *Journal of Discrete Event Dynamic Systems*, 2, pp. 61-88, 1992.

[9] J. Wieselthier, C. Barnhart and A. Ephremides, "Ordinal Optimization of Admission Control in Wireless Multihop Integrated Networks via Standard Clock Simulation," *Naval Research Laboratory*, NRL/FR/5521-95-9781, 1995.

[10] D. Bertsekas and R. Gallager, *Data Networks*. New Jersy:Prentice-Hall Inc., 1987 (2nd Ed. 1992).

[11] J. Baras, G. Atallah, R. Karne, A. Murad, and K. Jang, "Object Oriented Hybrid Network Simulation, A Functional Description" *Technical Report CSHCN TR 94-2, Center for Satellite and Hybrid Communication Networks, University of Maryland, College Park*, 1994.

ATM QoS provisioning in Broadband Satellite Networks

I. Mertzanis, G. Sfikas, R. Tafazolli, B. G. Evans
Centre for Communication Systems Research,
University of Surrey, Guildford, Surrey, GU2 5XH.
Email: {I.Mertzanis,G.Sfikas}@ee.surrey.ac.uk

ABSTRACT

One of the main areas of interest in broadband satellite engineering is the provision of multimedia services with Quality of Service (QoS) guarantees to a large number of users, while maintaining high system resource utilisation. In a broadband satellite network, a two level control algorithm must be considered; one at the Medium Access Control (MAC) layer and the other at the Asynchronous Transfer Mode (ATM) layer. This paper concentrates on the queuing performance and buffer dimensioning of a Satellite-ATM (S-ATM) network model, assuming that the network arrivals at the edges of this network are self-similar processes. A simulation model was developed in OPNET modeler and Fractional Gaussian Noise (FGN) samples were used to represent the input traffic. Simulation data analysis was performed to derive the coefficients of a formulae that approximates the buffer overflow probability. Finally, the ATM multiplexer input traffic characteristics were investigated and results for the ATM buffer dimensioning are presented.

INTRODUCTION

A huge increase in the demand for satellite multimedia communications has been noticed in the last few years and several different systems are under development worldwide. These systems are designed to provide high quality multimedia services to remote fixed, transportable or even mobile user terminals. Based on the assumption that the future broadband satellite networks [1] will accommodate ATM compatible equipment, queuing models can be used to study the network performance and to perform link and buffer dimensioning optimisation. Traditional methods of studying the network performance assume Poisson arrivals for the network arrival process. However, most studies on wide-area traffic patterns conclude that the assumption of exponentially distributed packet inter-arrival times is not always valid. In fact, a large number of articles [2,3,4,5,6] suggest that modelling the network traffic arrivals as Poisson processes, underestimates the burstiness of the traffic and the required buffer space in the intermediate nodes of a packet switched communication network. Some of them [5,6] have even provided evidence that the actual network traffic is self-similar, or in other words its bursty nature is kept over a wide range of time scales. It is reported that Poisson models are only valid when modelling the arrivals of TELNET or File Transfer Protocol (FTP) user sessions but not the FTP data bursts. If the correlation between neighboring exclusive blocks of a stationary process does not asymptotically vanish when the block size is increased, then this process is called long range dependent. As a result, long range dependence is quite important to be considered in buffer dimensioning studies. A self-similar process is long range dependent, as long as it has any positive correlations; these remain the same at all time scales. A lot of research has been carried out in order to develop methods that produce self-similar traffic. In [7], a sort description of some of the methods presented in the literature is given, highlighting their main advantages and drawbacks. Among them is the aggregation of n ($n \to \infty$) number of ON/OFF sources in which the ON (active) and OFF (silence) periods follow "heavy tailed" distributions (e.g. Pareto) or the generation of sample paths of a fractional ARIMA [8] process. However, in order to create traces of a true self-similar process, these methods require a very large amount of computer processing power and therefore are very slow. Fractional Gaussian Noise is a type of self-similar process that can be approximated by fast algorithms and used in network simulations to model packet arrivals. In the rest of the paper a S-ATM network model is described first, followed by the estimation of the FGN traffic descriptors and the buffer dimensioning study at both the MAC and the ATM layers.

S-ATM QUEUING MODEL

The main objective of this investigation is to study the performance of a satellite-ATM network model, assuming that the input traffic that enters the system though a variable number of terminals experience self-similar characteristics. FGN provides a useful and practical way to model aggregated Local Area Network (LAN) traffic that exhibits self-similar characteristics. The approach suggested in [7] for generating approximated FGN sample paths was followed and using the C code provided in [9], a set of 5 different sample traces were created having a target Hurst parameter of 0.50, 0.60, 0.70, 0.80 and 0.89. This method is presented in [10] using a Fast Fourier Transform (FFT) algorithm and it is very fast and quite accurate.

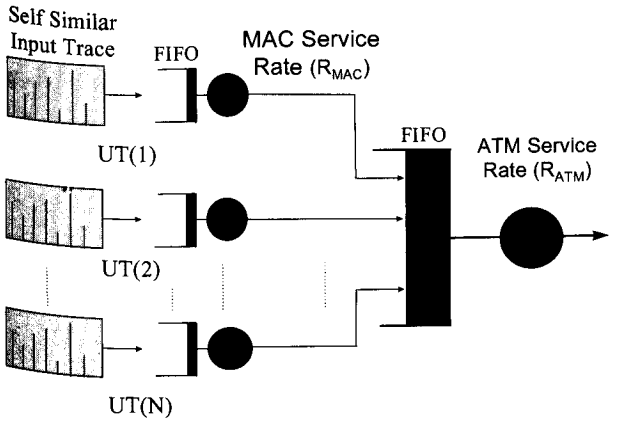


Figure 1: S-ATM queuing model

The block diagram of the queuing model that was used to represent a S-ATM network is shown in Figure 1. A simulation model was developed in OPNET modeller and is able to execute different input traffic scenarios. Each User Terminal (UT) consists of an input arrival process, an infinite First-In First-Out (FIFO) order MAC buffer with a fixed service rate R_{MAC}. The UT input process is a self-similar FGN process with a given mean *(m)*, variance *(v)* and Hurst parameter *(H)* and represents aggregated traffic that enters the network at each service access point. The self-similar traffic arrives at the input of the MAC queue in bursts of fixed packets (S-ATM cells) per time unit. The traffic that comes out of each queue, enters the ATM multiplexer which represents an on-board satellite switch output buffer. This is also a FIFO order infinite buffer that accepts traffic from all the simulated Uts. Each packet is served at a fixed ATM service rate (R_{ATM}). A common time unit of 1 second was used for all the simulation parameters; therefore all the simulation results are normalised to a common time reference. The representation of the ATM switch as a FIFO queue server model is accurate only for single service traffic. In this study, either non-real time Variable Bit Rate (nrt-VBR) or Available Bit Rate (ABR) service classes could be considered.

TESTING FGN SAMPLES FOR SELF-SIMILARITY

The variance-time (v-t) plot is a heuristic test for self-similarity which shows the rate of change of the variance when it is calculated for different aggregation sizes. For example, for a process that consists of set of samples $\{X_t\}$, t=0,1,2,.. the aggregated process $\{X_t^{(m)}\}$ is the process where:

$$X_t^{(m)} = \frac{1}{m} \times \sum_{i=(t-1)m+1}^{tm} X(i), t = 1,2,...$$

For very bursty traffic (i.e. H>0.75) the v-t plot test underestimates the Hurst parameter. Whittle's estimator is considered to be a very accurate method to estimate the Hurst parameter but it is not a test for self-similarity. Therefore, both methods the v-t plot and Whittle's estimator were used to characterize five different FGN sample paths and the results are shown in Table 1.

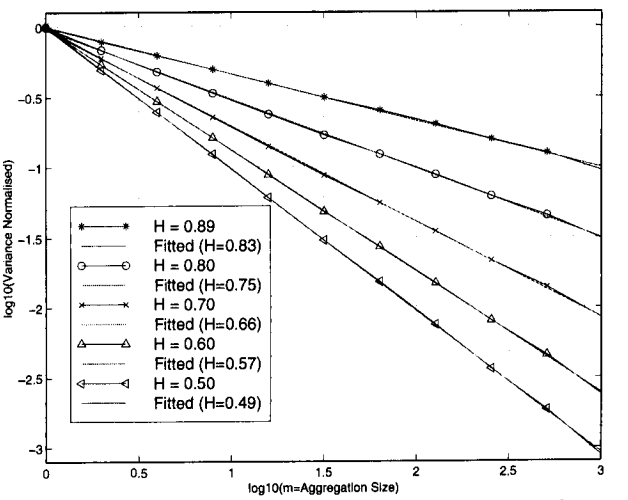


Figure 2: v-t plots of self-similar FGN sample paths

The 'Target' Hurst parameter is the parameter that was used in the FGN sample generation program and the 'Test' Hurst parameter is the one that was recorded from both tests. In both methods, the 'Test' Hurst parameters are always slightly lower than the 'Target' values. For a wide range of H values the 'Test' values are very close to each other. Figure 2, shows that as the trace aggregation size *(m)* increases from 1 to 1,000 the variance of the five different FGN traces decays by the expected factor of $m^{-2(1-H)}$. This decay is shown as a straight line in a log-log plot and it illustrates the long range dependence of the traffic that enters the MAC buffer of the S-ATM queuing model.

QUEUING PEROFRMANCE AT THE MAC SHAPER

The queuing performance at the MAC buffer is examined in this paragraph, using FGN as the input process. According to [11], there is no exact formula for the queue length distribution of the fractional Brownian motion. Therefore, a lower bound approximation solution that was followed by Norros resulted in the following expression for the buffer

Table 1: Parameters of the generated self-similar FGN sample paths

Trace No	Target Hurst Parameter	Test v-t plot Hurst Parameter	Test Whittle's Estimator (95% Confidence)	Sample Mean	Sample Variance
1	H=0.89	0.832	0.8536±1.49E-03	3.99	16.141
2	H=0.80	0.752	0.7548±7.77E-04	3.99	16.137
3	H=0.70	0.658	0.658±4.59E-04	3.99	16.127
4	H=0.60	0.566	0.571±3.84E-04	3.99	16.125
5	H=0.50	0.493	0.4977±3.76E-04	3.99	16.115

overflow probability:

$$P(X > x) \approx \exp\left[-\frac{(C-m)^{2H}}{2k(H)^2 a \cdot m} \cdot x^{2-2H}\right] \quad (1)$$

$C = R_{MAC}$ is the mean output rate, m is the mean input rate, v is the input variance, α is the variance coefficient, H $\in$ [1/2, 1) is the Hurst parameter and $k(H)$ is given by:

$$k(H) = H^H (1-H)^{1-H} \quad (2)$$

It should be noticed that in general, the Cell Loss Ratio $CLR \neq P(X > x)$ since CLR assumes a finite buffer space and the buffer overflow probability assumes infinite buffer space. Nevertheless, as explained in [12], this is not a bad approximation if extreme cases are excluded such as when a very large burst arrives at once and then for a very long period there is no traffic at all. As a result, we define:

$$CLR \equiv P(X > x) \quad (3)$$

Impact of the MAC utilisation

As explained in the previous paragraph, the exact calculation of the CLR (or the buffer overflow probability) in the presence of self-similar traffic is not a straight forward procedure. Therefore, several simulation runs were performed in order to capture and analyse the combined affects of both the MAC output rate and the Hurst parameter on the queue length distribution.

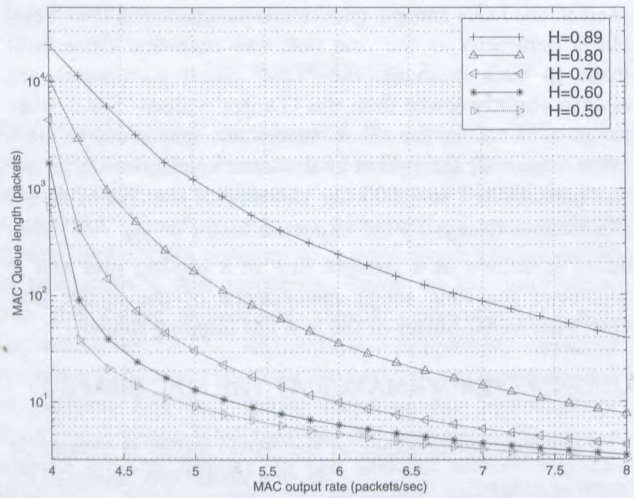


Figure 3: Average MAC Queue length as a function of the MAC service rate

In Figure 3, the average queue length of a MAC shaper as a function of the service rate for different 'target' H (see *Table 1*) is presented. In these graphs, the MAC utilisation (m/R_{MAC}) varies from 0.5 to 1. It is clearly shown that for low values of the Hurst parameter (i.e. H < 0.65) the average queue length is less sensitive to the changes on MAC utilisation than it is when the Hurst parameter becomes higher than 0.70.

However, although the average values give an indication of the buffer space requirements, what is also important to extract from the simulation statistics is the buffer overflow probability which approximates the expected CLR. This is shown in Figure 4 and Figure 5 for MAC utilisation of 0.80 and 0.67 respectively.

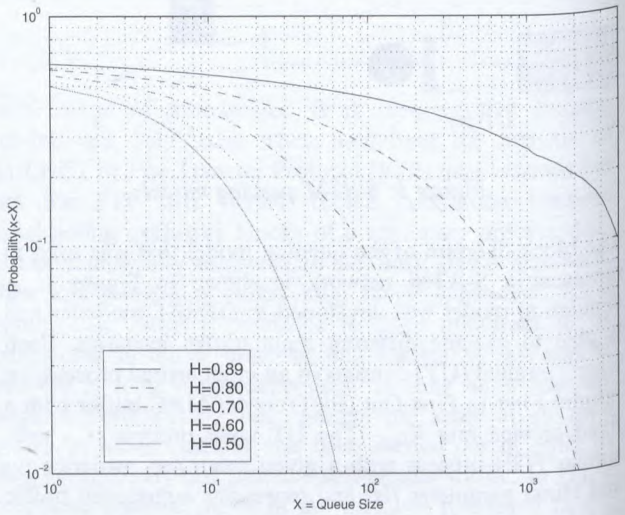


Figure 4: MAC buffer overflow probability for MAC utilisation = 0.80

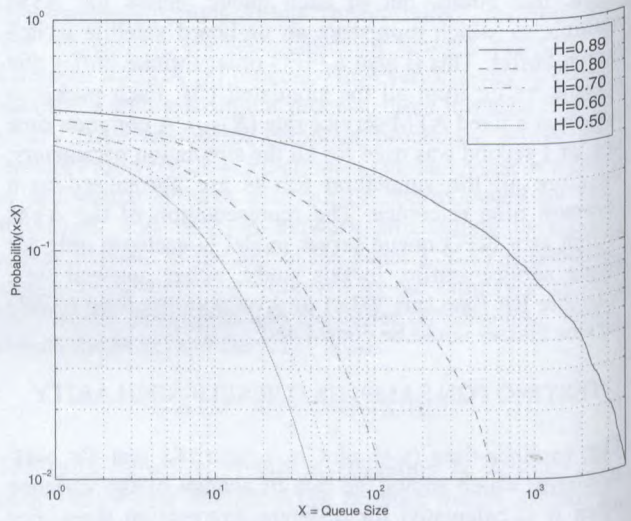


Figure 5: MAC buffer overflow probability for MAC utilisation = 0.67

In Table 2, a summary of the buffer space requirements for all the simulated traffic streams at $CLR \cong 10^{-2}$ is given. When compared to the values that were recorded for Poisson packet arrivals (mean=4 packets/time unit), it is clearly shown that buffer dimensioning becomes an entirely different problem. The methodology that was followed in order to approximate the CLR at much larger buffer space follows.

Table 2: Buffer space requirements for CLR=10^{-2}

MAC Utilisation	Trace No	Target H (see Table 1)	Buffer Space
0.80	1	0.89	> 10,000
	2	0.80	2,000
	3	0.70	300
	4	0.60	100
	5	0.50	60
	Poisson		10
0.67	1	0.89	4,000
	2	0.80	600
	3	0.70	110
	4	0.60	50
	5	0.50	40
	Poisson		5

LARGE BUFFER SPACE APROXIMATION

The buffer space requirements presented in Table 2 represent a very high value of CLR assuming low and medium MAC buffer space. However, in many ATM traffic streams the CLR requirements are for much lower values (i.e. < 10^{-10}). Therefore, an approximation method is needed which can predict the CLR at much higher buffer space. The approximation given by Norros provides a lower bound of the $P(X>x)$ for large values of **x**. So from equation (1) we can write:

$$\ln(P(X>x)) \approx \ln(\exp\left[-\frac{(C-m)^{2H}}{2k(H)^2 v\cdot m}\cdot x^{2-2H}\right]) \Rightarrow$$

$$\ln(P(X>x)) \approx \left[-\frac{(C-m)^{2H}}{2k(H)^2 v\cdot m}\cdot x^{2-2H}\right]$$

By using the following expression for **x**:

$$x=\left[-\ln[P(X>x)]\cdot\left[\frac{(C-m)^{2H}}{2k(H)^2 v\cdot m}\right]^{-1}\right]^{\frac{1}{2-2H}} \Rightarrow$$

$$\ln(x)=\ln\left[\left[-\ln[P(X>x)]\right]^{\frac{1}{2-2H}}\right]+\ln\left[\frac{(C-m)^{2H}}{2k(H)^2 v\cdot m}\right]^{-\frac{1}{2-2H}} \Rightarrow$$

$$\ln(x)=\frac{1}{1-2H}\cdot\ln\left[-\ln[P(X>x)]\right]-\frac{1}{1-2H}\cdot\ln\left[\frac{(C-m)^{2H}}{2k(H)^2 v\cdot m}\right] \quad (4)$$

we define:

$$\aleph=\ln(x),\Psi=\ln\left[-\ln[P(X>x)]\right] \quad (5)$$

So (4) can be written as:

$$\Psi=\alpha\cdot\aleph+\beta, \text{ where } \alpha=2-2H,\ \beta=\ln\left[\frac{(c-m)^{2H}}{2\cdot k(H)^2\cdot v\cdot m}\right] \quad (6)$$

By applying the transformation shown in (5) to the simulation data, we can use the least-squares method to find the coefficients α and β. As a result, a new formula can be derived directly from equation (1) that takes into account the adjustments applied by α and β to the curves that represent the approximated lower-bounds. In such as way, we can extend the $P(X>x)$ plots for much larger values of **x**.

Table 3: Polynomial coefficients used in (6)

Hurst parameter	R_{MAC}	α	β
H=0.85	5	1.51E-01	-8.98E-01
	6	1.52E-01	-2.74E-01
	7	1.55E-01	6.57E-02
	8	1.53E-01	2.99E-01
H=0.75	5	2.50E-01	-8.81E-01
	6	2.49E-01	-2.70E-01
	7	2.58E-01	2.28E-02
	8	2.59E-01	2.47E-01

In the figure below, the analytically approximated lower-bounds of CLR and the graphs derived from equation (6) are presented for different values of H and R_{MAC}. The coefficients α and β are presented in Table 3.

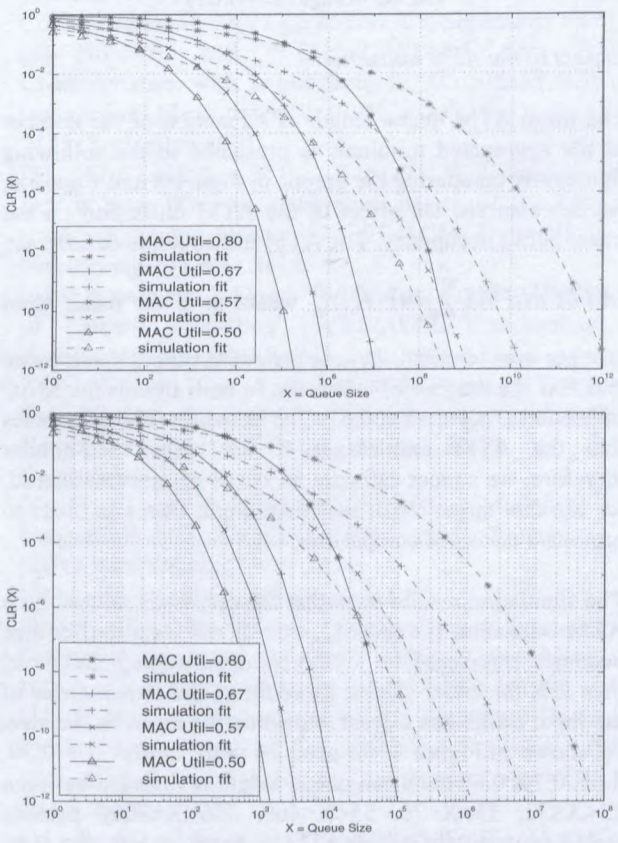


Figure 6: Lower-bound (plain lines) and simulation data extrapolation (dashed lines) using Equation (6) for Trace No.1 (top) and Trace No. 2 (bottom)

QUEUING PERFORMANCE AT THE MULTIPLEXER

A variable number of self-similar MAC shaped sources contribute to the packet inter-arrival process of the ATM multiplexer. As shown in Figure 7, the input at the ATM-multiplexer has no longer self-similar characteristics. This observation remains valid from a single MAC shaped trace up to an arbitrary high number of aggregated self-similar processes.

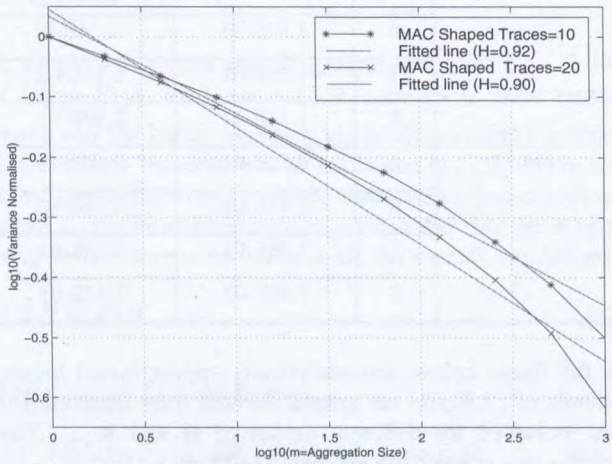


Figure 7: v-t plots for aggregated MAC shaped traces (MAC utilisation =0.67)

Impact of the ATM utilisation

The mean ATM queue length as a function of the number of the aggregated terminals is presented in the following figures. By comparing the graphs in *Figure 8* and *Figure 9*, we can also see the affect of the ATM utilisation on the mean buffer occupancy. The ATM utilisation is defined as:

$$ATM\ util = (\sum_{i=1}^{N} m) / R_{ATM}$$ where m is the mean input rate per user terminal, R_{ATM} is the ATM output service rate and N is the number of terminals. In both figures the MAC utilisation is selected at 0.67. The shaped traffic that comes into the ATM multiplexer is no longer self-similar therefore, we cannot estimate its Hurst parameter. Instead, we use the 'target' H of the MAC input traces in order to provide a means of comparison.

The simulation results show that for very high values of the ATM utilisation (i.e. ATM_{util} >0.90) the mean buffer size becomes less sensitive to the terminal aggregation level than it is for lower values. In addition, the H parameter of the input traffic has a great impact on the mean buffer size. As shown in Figure 8, for samples with 'target' H = 0.50, 0.60, 0.70, 0.80 the mean queue length is bounded between 23<X<33, 33<X<70, 55<X<400, 300<X<3000 packets (cells) respectively. When ATM_{util} becomes less than 0.90 (see Figure 9), for samples with 'target' H greater than 0.70 and large buffer space, the mean buffer size drops rapidly with the number of terminals, which indicates that a high ATM multiplexing gain can be achieved. For H less than 0.70 the mean buffer size is kept less than 10 packets.

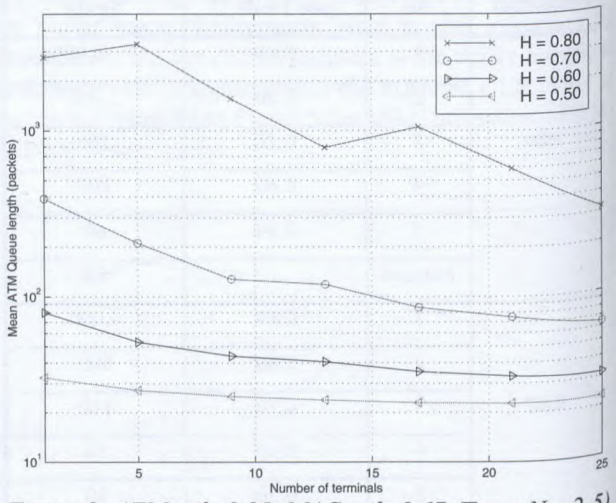


Figure 8: ATM util=0.95, MAC util=0.67 (Trace No. 2-5)

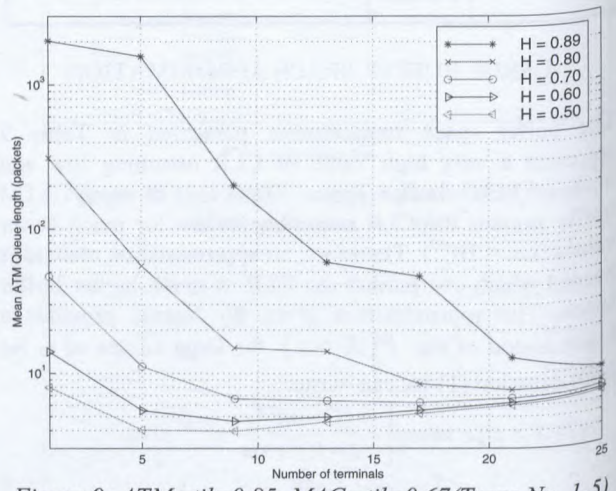


Figure 9: ATM util=0.85, MAC util=0.67(Trace No. 1-5)

ATM buffer overflow probability

In the previous paragraph, the impact of the input traffic and the ATM utilisation on the mean ATM queue length was presented. The next step is to examine the ATM buffer overflow probability which is calculated from the simulated ATM buffer occupancy histograms. The probability $P(X > x) = 1 - CDF(X)$ is plotted as a function of the queue size X, where CDF is the Cumulative Distribution Function of X. The results of two of the network traces shown in Table 1 are presented for ATM utilisation 0.75 and 0.85. For ATM utilisation greater than 0.85, the buffer overflow probability decays very slowly and it is less sensitive to the number of active terminals. However, when moving to lower values of the ATM utilisation (i.e. 0.75) a very sharp decay of the CLR is expected with the aggregation of more than 17 terminals. As shown in Figure 12, the buffer space requirements assuming CLR=10^{-4} are slightly higher than 20 packets (cells).

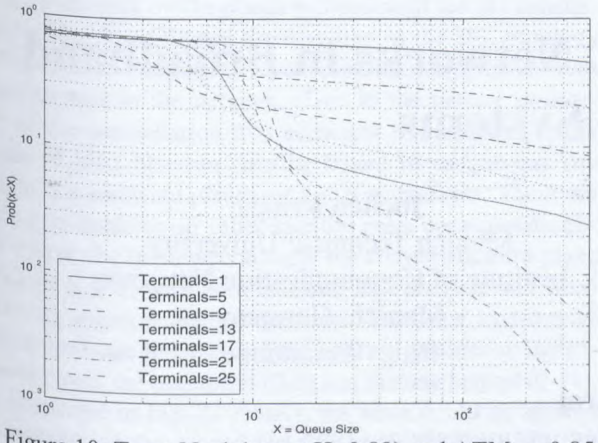


Figure 10: Trace No 1 (target H=0.89) and ATM_{util}=0.85

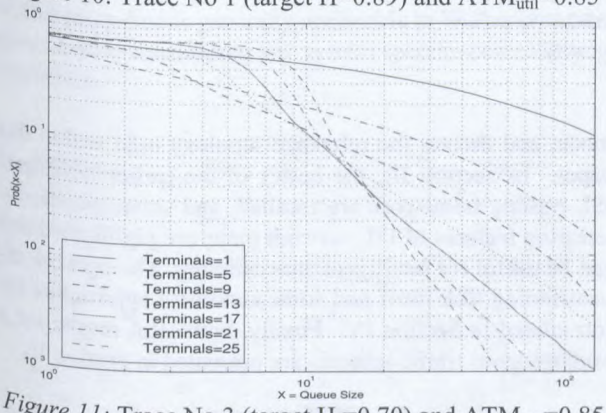


Figure 11: Trace No 3 (target H =0.70) and ATM_{util}=0.85

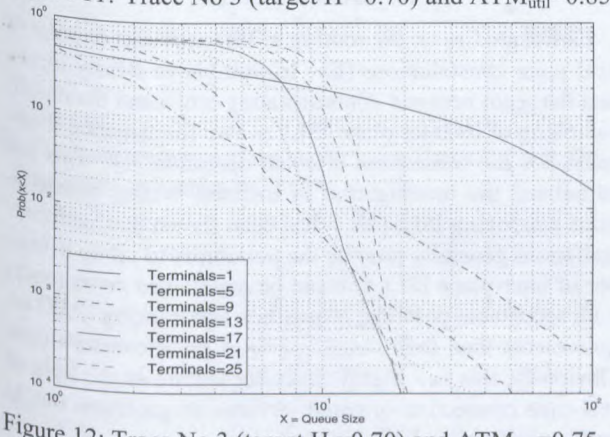


Figure 12: Trace No 3 (target H =0.70) and ATM_{util}=0.75

SUMMARY AND CONCLUSIONS

In this paper the queuing performance of a satellite-ATM network model is investigated, assuming that the input traffic that enters the system through a variable number of terminals experience self-similar characteristics. A simulation model was developed using OPNET modeler and the correction coefficients of an approximation formulae for the buffer overflow probability were derived from the simulation data. It was observed that the network arrival process at the on-board satellite switch is no longer self-similar, since it is affected by the MAC shapers at the satellite network access points. A certain multiplexing gain can be achieved at the on-board satellite output buffers provided that the MAC utilization is kept low (i.e. 0.67). These results are based on the assumption that self-similarity is observed at the edges of the satellite-ATM network. However, there might be still cases where the network arrival process at the on-board satellite switch exhibits self-similar behavior. Therefore, the MAC buffer dimensioning results can be used for the on-board satellite switch dimensioning. Nevertheless, due to the space segment hardware limitations it is suggested that traffic shaping at the satellite network access points should be always performed in order to minimize the possibility of introducing self-similar traffic at the on-board satellite switch.

REFERENCES

[1] I. Mertzanis, G.Sfikas, R.Tafazolli, B.G.Evans, "Protocol Architectures for Satellite-ATM Broadband Networks", IEEE Communications Magazine, Vol.37, No.3, March 1999. special issue on Satellite-ATM Network Architectures.

[2] R. Jain and S.Routhier, "Packet Trains-Measurements and a new model for Computer network traffic", IEEE JSAC, 4(6), pp.986-995, September 1986.

[3] R.Gusella, "A Measurement Study of Diskless Workstation Traffic on an Ethernet", IEEE Transactions on Communications, 38(9), pp.1557-1568, September 1990.

[4] H.Fowler and W.Leland, "Local Area Traffic Characteristics, with Implications for Broadband Network Congestion Management", IEEE JSAC, 9(7), pp.1139-1149, September 1991.

[5] W.E.Leland, M.S. Taqqu, W.Willinger and D.V.Wilson, "On the self-similar nature of the Ethernet traffic (Extended Version)", IEEE/ACM Transactions on Networking 2 (1994), 1-15.

[6] V.Paxson and S.Floyd, "Wide-Area Traffic: The failure of Poisson Modelling", IEEE/ACM Transactions on Networking, 3(3), pp.226-244, June 1995.

[7] V. Paxson, 'Fast Approximate Synthesis of Fractional Gaussian noise for generating Self-Similar network traffic', Computer Communications review 27(5), pp. 5-18, October 1997.

[8] M. Garret, W.Willinger, "Analysis, Modelling and Generation of Self-Similar VBR traffic", Proceeding of SIGCOMM'94,September 1994.

[9] http://www.acm.org/sigcomm/ITA/

[10] P. Flandrin, "Wavelet Analysis and Synthesis of Fractional Brownian Motion", IEEE Transactions on Information Theory, 38(2), pp.910-917, March 1992.

[11] I. Norros, "On the use of the Fractional Brownian Motion in the Theory of Connectionless Networks", IEEE Journal on Selected Areas In Communications, VOL 13, NO.6, August 1995.

[12] R. G. Addie et al, "Broadband Traffic Modelling, Simple Solutions to Hard Problems", IEEE Communications Magazine August 1998.

Capacity Dimensioning of ISL Networks in Broadband LEO Satellite Systems

Markus Werner, Frédéric Wauquiez
German Aerospace Center (DLR)
Institute for Communications Technology
P.O. Box 11 16, D-82230 Wessling, Germany
Markus.Werner@dlr.de

Jochen Frings
Munich Technical University
Institute of Communication Networks
Munich, Germany
frings@ei.tum.de

Gérard Maral
Ecole Nationale Supérieure des Télécommunications (ENST)
Toulouse, France
maral@tlse.enst.fr

ABSTRACT

This paper considers the capacity dimensioning of intersatellite link (ISL) networks in broadband LEO satellite systems, where the major challenge is the topology dynamics. First, a general method to design convenient ISL topologies for connection-oriented operation is presented, and a reference topology for numerical studies is derived. A permanent virtual topology is then defined on top of the physical one, thus forming a framework for discrete-time dynamic traffic routing. On this basis, heuristic and optimization approaches for the combined routing and dimensioning task, operating on discrete time steps, are presented and their performance is numerically compared. It is shown that minimizing the worst case link capacity is an appropriate target function, which can be formulated as linear minmax optimization problem with linear constraints. Using linear programming (LP) techniques, the dimensioning results are clearly better than with simple heuristic approaches.

I. INTRODUCTION

Future broadband LEO satellite communication systems will increasingly rely on an intersatellite link (ISL) trunk network with time-variant topology. Within the network planning process, the routing and dimensioning tasks are in general closely coupled and do essentially influence both installation and operation costs of a system. The routing of complex global traffic flows over dynamic network topologies has already been addressed in former work [1], [2], [3] to a level of detail that allows to provide necessary input for the dimensioning task. For the capacity dimensioning of ISL networks one may use approaches, algorithms and tools known from terrestrial (ATM) networks to a certain extent, but has to take into account specific additional constraints like the time-variance of the topology and also potentially different target functions adapted to the LEO ISL scenario.

The remainder of this paper is organized as follows: Section II first presents a pragmatic method to design convenient ISL topologies for connection-oriented operation mode, and derives the reference topology used within this paper. In Section III, the basics of the earlier developed ISL routing framework are recalled, and some specific interesting features of ISL network paths are highlighted that can be useful for intelligent dimensioning. The network dimensioning task itself and some candidate approaches are formulated in Section IV. Finally, numerical results for a homogeneous traffic scenario are presented in Section V.

II. ISL TOPOLOGY

Earlier studies on the routing in ISL networks of polar or near-polar constellations (like Iridium) have clearly identified the seam between contra-rotating orbits and the on/off-switching of its inter-plane ISLs as two fundamental drawbacks for the connection-oriented operation [1]. This has stimulated the investigation of inclined Walker constellations employing ISLs [4]. It has been shown that such constellations generally provide the possibility to set up a number of inter-plane ISLs that can be maintained *permanently* with acceptable pointing, acquisition and tracking (PAT) requirements than their counterparts in polar constellations. Obviously, this is a highly desirable feature in the light of real-time connection-oriented services, as problems due to path switching can be avoided.

M-Star [5] has been one of the first commercial system proposals aiming at the promising combination of Walker orbits and ISLs. A proper design of the ISL topology to be implemented is a first important step to guarantee efficient networking in the operational system. In the following we present a pragmatic approach to the ISL topology design in Walker constellations, deriving a reference topology for M-Star as an example. The relevant constellation parameters are summarized in Table I.

TABLE I. M-STAR CONSTELLATION PARAMETERS.

Orbit altitude	1350 km	Number of orbits	12
Orbit period	113 min	Inclination	47°
Number of satellites	72	Phasing factor	5

A closer look on the planar projection of the 3D constellation, Fig. 1, facilitates the first step in the ISL topology design, which is to identify potential links to be reasonably implemented in the network. Due to the perfect symmetry of the constellation it is sufficient to consider generic types of ISLs between satellite 0 and its neighbours (on "right",i.e. eastward planes) at $t = 0$ as example. These are then applicable to all other satellite pairs correspondingly. Of course, the implementation of the convenient intra-plane ISLs (0-1, 0-5) with both constant link distance and fixed pointing angles is beyond question. Then one would intuitively envisage links toward the next neighbours on the adjacent plane, the one ahead (0-6) and the one behind (0-11), as illustrated in Fig. 2. Finally, the same could be applied to partners on the second but next plane (0-12, 0-17). So far, this procedure is generally applicable to Walker constellations *before* considering any system specific constellation parameters.

Fig. 1. Planar projection of the M-Star constellation at $t = 0$. Distances and angles are *not* true to scale.

Fig. 2. Potential generic ISLs in a regular Walker constellation.

In fact, the feasibility of all these envisaged links has still to be proven taking into account geometrical and technological constraints for a specific constellation. The diagrams in Fig. 3 display important geometrical data for this purpose. One can notice that both links toward the adjacent orbit show relatively little variation in distance and pointing angle, and thus establishing these ISLs in permanent mode does not introduce severe problems. On the contrary, ISL 0-17 is less attractive but still possible, whereas Earth shadowing completely prevents implementation of ISL 0-12.

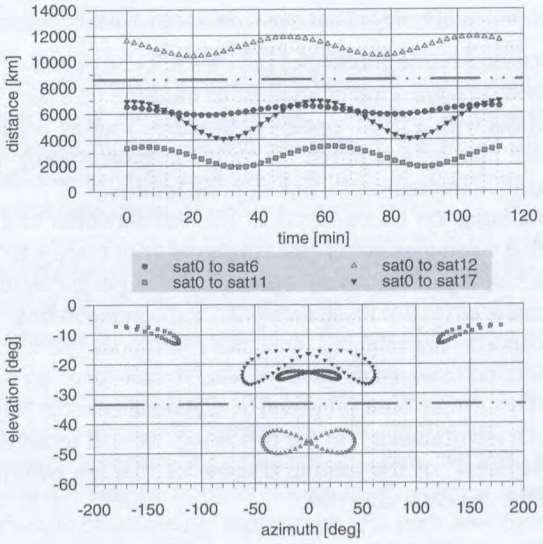


Fig. 3. Geometrical feasibility of ISLs in M-Star: (a) time variation of ISL distance; (b) pointing diagram. The dash-dot line represents the upper bound with respect to Earth shadowing.

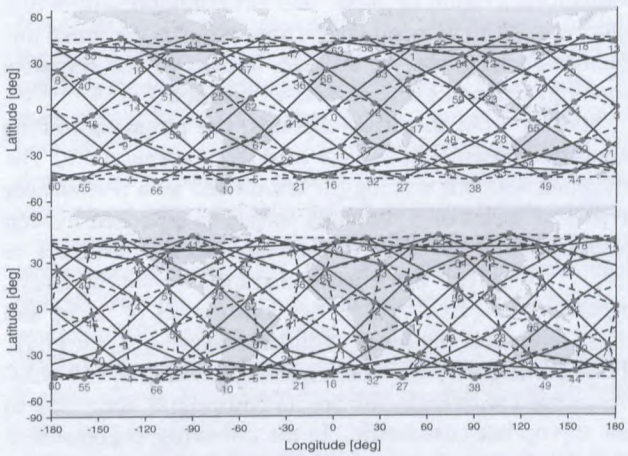


Fig. 4. Considered M-Star ISL topologies T1 (top) and T3 (bottom).

For the purpose of this paper, we have selected as reference topology T1 the case of only implementing generic ISL type 0-6 besides the intra-plane ISLs (resulting in 4 bidirectional ISL's per satellite). Most numerical results will be presented for T1. A topology T2, with generic ISL 0-6 replaced by the alternative 0-11, has also been investigated but does not show significantly different results. Finally, a topology T3 has both mentioned generic inter-plane ISL's employed at the same time, thus yielding 6 bidirectional ISL's per satellite. The resulting higher degree of meshing does certainly have impact on routing performance and dimensioning; therefore some representative results for T3 will also be presented in comparison with T1. Fig. 4 displays snapshots of T1 and T3, respectively.

III. ISL Routing Framework

Classical routing strategies have traditionally more or less focused on some kind of shortest or multiple path search in networks with a fixed topology. Dynamic routing capabilities are then only required for traffic adaptive routing or in reaction to unpredictable link or node failures.

The situation encountered in ISL subnetworks of LEO satellite constellations is quite different with respect to network topology. Permanent topological changes are an inherent characteristic of those networks, and corresponding routing concepts are required to enable continuous operation in connection-oriented mode. *Dynamic Virtual Topology Routing (DVTR)* has been proposed as a general concept for use in such environments. In the following, we will recall those "ingredients" of the routing framework that are correlated with the dimensioning task.

A. Discrete-Time Network Model and Routing Concept

Consider an arbitrarily meshed network topology with a constant number N of network nodes (satellites in the constellation). The network topology is in the most general case subject to changes due to (a) discrete-time activation/deactivation of links, and (b) continuous-time distance variations between nodes; depending on the actual implementation,(a) may be avoided. Moreover, the complete topology dynamics is periodical with period T.

Starting from these assumptions, the proposed routing concept is based on a discrete-time topology approach: The dynamic network topology is considered as a periodically repeating series of S topology snapshots separated by step width $\Delta t = T/S$. Each of the snapshots at $t = s\Delta t, s = \{0, \ldots, S-1\}$, is modelled as a graph $G(s) = (V, E(s))$, where $V = \{1, \ldots, N\}$ is the constant set of nodes and $E(s)$ represents the set of undirected links $(i,j)_s = (j,i)_s$ between neighbouring nodes i and j, existing at $t = s\Delta t$. Associated with each link are its costs $c_{ij}(s)$ according to an appropriate cost metric. In the following, it is assumed that the link costs are mainly determined by node distance respectively propagation delay.

Throughout the period T a constant set W of $\frac{N(N-1)}{2}$ unordered origin-destination (OD) node pairs is given. A set $P_w(s)$ of up to K distinct loopless paths is now assigned to each OD pair $w, w \in W$. Every path $p(s) \in P_w(s)$ consists of a unique sequence of links $(i,j)_s$. A *link occupation indicator* catches the relationship between a link and a path,

$$\delta^{p(s)}_{(i,j)_s} = \begin{cases} 1 & \text{if } (i,j)_s \text{ belongs to path } p(s) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The assignment procedure reflects the *setup of an instantaneous virtual topology* $VT(s)$ upon $G(s)$. This task is performed by an K-shortest path search algorithm (KSPA) for every OD pair. The single shortest path search task can be formulated as finding the least-cost path $p(s)$, i. e. the path with minimum path cost

$$\min_{\forall p(s)} C_{p(s)} = \min_{\forall p(s)} \sum_{i,j} c_{ij}(s)\, \delta^{p(s)}_{(i,j)_s} \quad (2)$$

Performing this path search for all $s = \{0, \ldots, S-1\}$ completes the set-up of a discrete-time dynamic virtual topology. It consists of an ordered set of K alternative paths for any OD pair at any step s. From these K alternatives, the $k \in \{1, \ldots, K\}$ best ones can be effectively used by the routing, where k could be simply identical for all OD pairs. However, for a fixed k together with a strict ordering within the K-path sets, it can happen that a path does not belong to the reduced k-path set in subsequent steps with obvious consequences for the continuity of connections using it. Fortunately, the typical path (delay) characteristics of the considered ISL topologies offer a nice opportunity to cope with this problem without discarding the cost-oriented ordering of sets. This will be shown in the following subsection.

B. Path Grouping (PG) Concept

Studying the cost values of all ordered paths in a K-set for various OD pairs and various steps, one observes that the paths can be easily grouped according to cost ranges, as illustrated for an example OD pair of topology T1 in Fig. 5. Typically, one cost range corresponds to a certain number of ISL hops forming the respective paths; the paths belonging to the first two groups of the considered example are displayed in Fig. 6, and the relationship between hop count and path group becomes obvious.

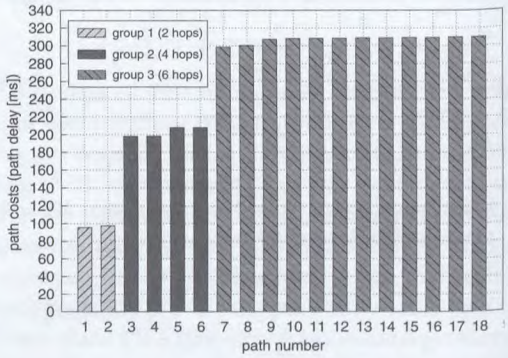


Fig. 5. Numerical example for clear path group separation by cost ranges (OD pair sat0-sat7 at step 0).

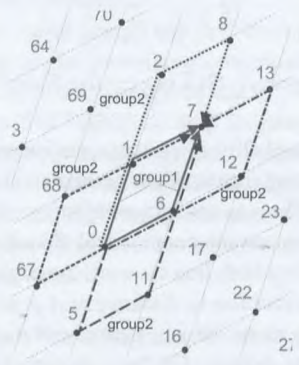


Fig. 6. Illustration of hop-based path grouping: first two path groups for OD pair sat0-sat7 at step 0.

It is now of great importance to realize that (in the case of T1) this clear path grouping extends over all steps, with characteristic group sizes and cost ranges for the respective OD pairs. In other words, the cost ranges of once identified path groups at a certain step do never overlap in any of the other steps of the topology period T as well, and therefore it is guaranteed that always the same paths belong to a specific group. The ordering of paths resulting from KSPA can only vary *within* a group.

One obvious solution for the potential path continuity problem addressed above is therefore simply to select a unique k for each OD pair, such that always *complete* path groups are contained in the k-path set; with this, the same paths per OD pair will be available for routing over all steps.

IV. ISL Network Dimensioning

A. Overall Approach and Assumptions

The "classical" network design process for terrestrial ATM WANs typically aims at minimizing some total or cumulative network costs (often: implementation costs), where the key optimization parameters are (a) the definition of the VPC (virtual path connection) topology, (b) the VPC capacity assignment, and (c) the routing rules for OD traffic.

With respect to an appropriate target function for the ISL network dimensioning, there is an important difference compared to the terrestrial case, which is due to the dynamic topology encountered. With the complete satellite constellation periodically orbiting the Earth and thus the source traffic demand, each satellite and each ISL will face the same worst case requirements (in terms of traffic load) "one day" and has hence to be dimensioned and built with respect to this unique value; as a consequence, all satellites in the constellation, including especially the ISL equipment, will be identical. Capacity requirements on an ISL translate into bandwidth and thus RF power requirements, whereas capacity requirements for a satellite node mainly drive processing power and buffer sizes. Alltogether, this translates into DC power requirements as well as into size and weight of single on-board components and finally of the whole satellite, the latter being a major cost factor of the satelite constellation.

Based on these considerations it is straightforward to formulate the two most appropriate target functions for the ISL network dmensioning:

TF1: Minimize the worst case link (WCL) capacity, which is the maximum capacity required on any link at any time

TF2: Minimize the worst case node (WCN) capacity (being correspondingly defined)

Of course, some combined metric is possible as well. In this paper, we consider the WCL target exemplarily. Keeping in mind this minmax type target function as a key feature, the ISL network dimensioning process comprises the following steps and additional assumptions.

Time-discrete approach. In general, every routing-dimensioning cycle is performed in time-discrete manner for a series of topology and traffic demand snapshots. This corresponds to what is known as *multi-hour network design* from terrestrial ATM network planning (cf. e.g. [6]). The additional complexity usually coming in with a multi-hour scenario is in our case immediately releaved by the minmax type target fuction: we can break down the whole ISL network dimensioning task into a number of completely independent dimensioning tasks, each minimizing the WCL capacity *per step*, since the overall minimized WCL capacity (our final target value) can be simply filtered out as the maximum of all the minimized per-step WCL capacities.

Permanent virtual topology. On top of the time-variant physical ISL network, a VPC topology is overlaid which is permanent over time; this has been discussed in the two previous sections. In contrast to most approaches for fixed topology design, in the considred ISL case the VPC topology is not subject to optimization itself, since it is simply the result of assessing appropriate OD path sets based on superior criteria (delay, delay jitter, permanence) aiming at a good performance of connection-oriented operation.

Routing and VPC capacities. The routing/distribution of given demand pair traffics on the available VPCs is either performed heuristically, according to fixed rules, or it can be treated as an optimization problem which is formulated and solved using linear programming (LP) techniques. Assuming that a limited set of alternative VPCs may be used for splitting the traffic between a specific pair of end nodes, the main optimization parameters are then the splitting factors. Instead of Erlang traffics, we directly operate with given *OD demand pair capacities* or *bandwidths* assuming that these values have been calculated before from the Erlang traffics [1] – independently for each demand pair according to the *virtual trunking concept* [7], [8]. The VPC capacities result directly from the splitting of the demand pair capacity on all available VPCs of the OD pair.

Link capacities. The required capacity of a single link at a given step is determined by simply summing up all VPC capacities crossing it.

The "heart" of the presented approach, namely the routing/splitting of OD capacities, is looked at in more detail in the following two subsections.

B. Heuristic Approach

A simple but pragmatic approach is based on observations made in earlier research [9] that investigated some intuitive rules for traffic routing/distribution in an ISL network. For a simple reference dimensioning we use *Equal Sharing (ES)*, that is, each OD traffic will be equally distributed on the k best paths, k being fixed for all OD pairs in the topology and over time. Intuitively, one can expect that the ES approach leads to a decrease of WCL traffic load with increasing k, just by "somehow" smoothening link load peaks mainly through dividing peak path loads by k. However,

[1] This calculation could for instance be based on the Erlang-B blocking formula or simply assume a proportional relationship between Erlang traffic and bandwidth, being linked by the average bit rate of the considered services.

such an effect can not be guaranteed in general, and a potential gain cannot be forecast as a systematic relation between the ES routing rule and the WCL target value does simply not exist.

Besides the pure ES approach, we have also investigated the combination with path grouping – *Equal Sharing using Path Grouping (ES/G)*. In this case, from the network-wide fixed k a resulting k^* is determined for each OD pair, which is simply the smallest integer value larger than or equal to k that completes a path group. Using the example from Fig. 5 again, $k = 4$ translates into $k^* = 6$, completing path group 2 of the considered OD pair. The positive effect we expect from introducing path grouping knowledge – besides the advantages already commented – is that the optimization potential may be significantly increased with the number of used paths, whereas the additional paths do not introduce much higher costs than encountered on the last path in the k-set, which is already used without path grouping.

C. Optimization Approach

In contrast to the heuristic approaches presented above, one may consider a dedicated optimization of a given target function, which is in our case the minimization of the overall WCL capacity. As already discussed in Section IV-A, the major part of this optimization consists of S dimensioning subtasks, namely minimizing the WCL capacity for all steps s independently, and the overall WCL capacity is then taken being the maximum of all minimized per-step WCL capacities. Using the formulation of the network model and dynamic routing concept presented in Section III-A, we consider one of these subtasks for a given step in the following, without explicit indexing with s in order to enhance the readability of the notation.

First of all we now restrict ourselves to the case of permanent physical links in the ISL topology, which means that we have a fixed set of links l over all steps, $l \in \{1 \dots L\}$; for T1, $L = 2N = 144$, for T3, $L = 3N = 216$. The offered capacity n_w per OD demand pair w is distributed among the k alternative paths effectively available, so that each path carries a certain share n_p of the total demand pair capacity,

$$n_w = \sum_{p, p \in P_w} n_p \qquad (3)$$

The required bandwidth n_l of a physical link l, which is our central target value for the optimization, is obtained as the sum of the bandwidths n_p of all paths containing this link,

$$n_l = \sum_{w, w \in W} \sum_{p, p \in P_w} \delta_l^p n_p \qquad (4)$$

Our objective to minimize the maximum required bandwidth on a single physical link (the worst case link WCL) can be formulated as the following linear minmax optimization problem:

$$\max_{\forall l}\{n_l\} = \max_{\forall l}\{ \sum_{w, w \in W} \sum_{p, p \in P_w} \delta_l^p n_p\} \rightarrow \min \quad , \qquad (5)$$

subject to the (linear) constraint formulated in Eq. (3).

The optimization parameters are the shares of total OD capacity carried by each path belonging to the OD pair, or equivalently, the *splitting factors* that determine the OD capacity split onto its correlated paths.

So far, we have not introduced any specific constraint on the share of the total traffic that one path is allowed to carry. As a consequence, a single path may convey the complete offered OD traffic alone, whereas other paths may remain empty. We refer to this approach as *Full Optimization (FO)* in the following.

Although FO certainly leads to the maximum possible WCL load reduction per step, there are some reasons – e.g., consequences for operation in failure situations, potentially enormous load variations on single links from step to step, etc. – to introduce an additional linear constraint in form of an upper bound α for the normalized share of the OD traffic one path is allowed to carry,

$$0 \leq n_p \leq \alpha n_w, \quad \alpha \in [0 \dots 1] \quad . \qquad (6)$$

This approach is refered to as *Bounded Optimization (BO)* (with parameter α) in the remainder of the paper.

Both, FO and BO can additionally be combined with the path grouping concept; these approaches are identified by the corresponding acronyms FO/G and BO/G, respectively.

Concluding this subsection, it is again emphasized that all the above considerations on optimization are valid for a single step in the dynamic scenario, and are consequently applied to each single step independently. An advanced but highly complex approach would be to include the time dependence of subsequent steps in the optimization; this would lead to a dynamic optimization problem, which is clearly beyond the scope of a first dimensioning approach as adopted in this work.

V. Numerical Studies

A. Scenario and Assumptions

All numerical studies presented in this section are based on a homogeneous traffic scenario, i.e., each OD demand pair offers traffic of one bandwidth unit over all steps. In the following, the terms traffic (load), capacity and bandwidth are used synonymously.

The main motivation to start with such an unrealistic traffic scenario for a first ISL network dimensioning approach has been to isolate and study influences resulting from the topology characteristics on the dimensioning process and the dimensioning results. A heterogeneous traffic scenario is likely to mask all or the most of such topological influences – at least it seems impossible to really separate the effects from the results later.

Extensive dimensioning runs have been performed for all topologies, T1, T2, and T3, where $K = 18$ (for T1 and T2) and $K = 10$ (T3) best paths have been in the complete set resulting from KSPA. Representative results are presented and discussed using T1 as example, after a first comparison between T1 and T3.

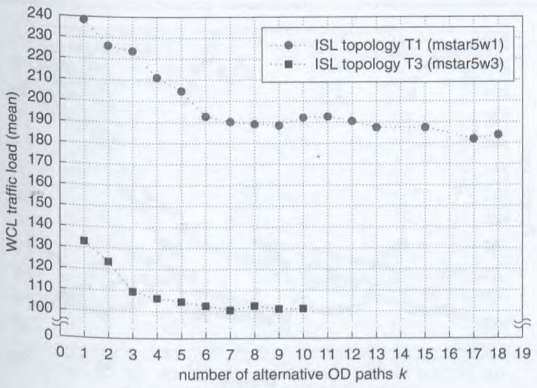


Fig. 7. Reduction of mean WCL traffic load through ES with growing k: Performance comparison between topologies T1 and T3.

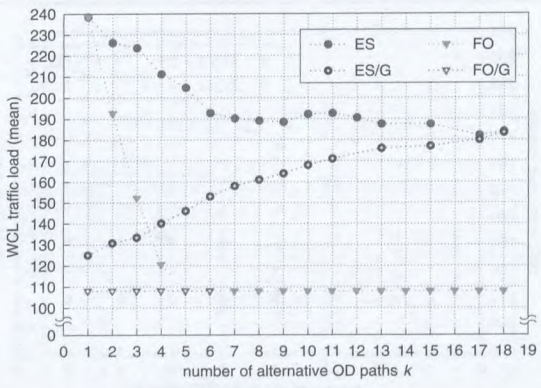


Fig. 8. Mean WCL traffic load versus k: Performance comparison between ES, ES/G, FO and FO/G.

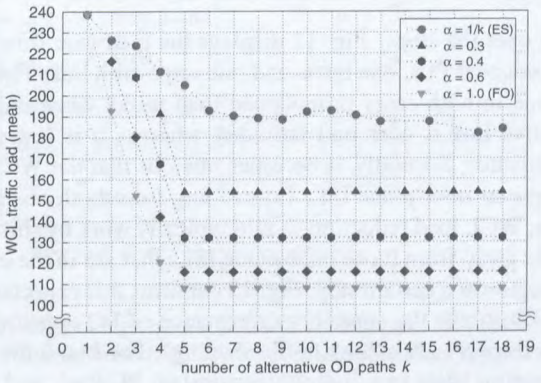


Fig. 9. Mean WCL traffic load versus k: Performance comparison for different values of α in the range between the two extreme cases, ES and FO.

B. *Worst Case Link (WCL) Traffic Load*

First of all it is interesting to study the performance of the different dimensioning approaches in terms of the traffic load on the most loaded physical link (worst case link, WCL), since this has been our primary objective value. The dimensioning has always been performed over all steps, and one important observation is that the variation of the WCL load over time is very small. Already without any systematic traffic distribution or optimization – if all OD traffic is routed over the respective shortest path alone – the WCL load variation over the steps stays within a range of less than 10%. And any approach that effectively reduces the WCL value (per step) leads to a significant further reduction of its variation over time as well – the lower the instantaneous WCL target value, the lower also its variation. Based on this observation we can restrict ourselves to presenting and discussing mean values (calculated from 50 single-step values) in the following.

Fig. 7 shows the WCL load results for the simple equal sharing (ES) approach, comparing T1 and T3. As expected, the values for T3 are significantly lower due to the availability of much more links to distribute the traffic on. With growing k, the WCL target value decreases nearly monotoneously, but in both topologies "saturation" of the optimization is more or less achieved for $k = 7$. The maximum WCL load reduction of ES is then roughly 25% with respect to the case of only routing over the shortest path ($k = 1$).

Fig. 8 displays at a glance the performance results for the "extreme" approaches ES, ES/G, FO and FO/G, concentrating on T1. The first impressive comparison is between the pure ES and FO curves, where FO achieves an *additional* improvement of 45% compared to ES in the saturation zone. With $k = 5$ it is already possible to use the full optimization potential within T1. When combining the two methods with path grouping, one faces an interesting difference: Whereas FO/G achieves the lowest possible WCL value independently of k, ES/G shows significant improvements with respect to pure ES for small k, but then the WCL target monotoneously grows with increasing k, which is a result of the fact that ES forces the equal distribution on all available paths, however they are selected. Obviously, from the WCL perspective alone, the only reasonable implementation of ES/G is then for $k = 1$.

Coming back to the pure approaches without path grouping, Fig. 9 illustrates how the BO method performs compared to the extreme cases ES (BO with $\alpha = 1/k$) and FO (BO with $\alpha = 1$). As expected, the additional constraint reduces the optimization potential with respect to the WCL target value. However, with moderate values of α the results come pretty close to the FO ones in the saturation; for $\alpha = 0.6$, the WCL load is less than 10% higher than for FO.

Concluding the observations of the WCL load target, Fig. 10 shows the excellent performance of BO when path grouping comes in again: BO/G with a moderate $\alpha = 0.4$ achieves nearly the same WCL load as FO/G. Combining as many advantages as possible, BO/G with $k = 5$ and $\alpha = 0.4$ is consequently a top candidate for our ISL dimensioning problem.

C. *Physical Link (PL) Traffic Load*

In order to understand *how* the various traffic distribution and optimization approaches operate on the topology, it is helpful to study the load variation on single physical links

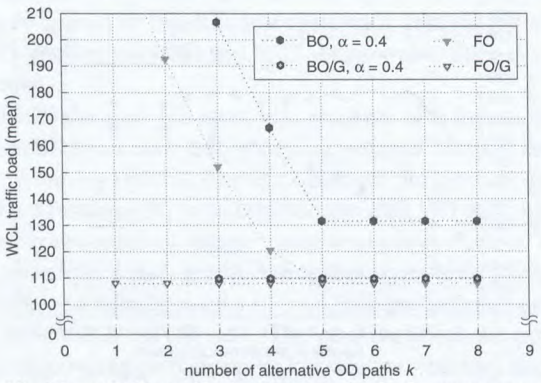


Fig. 10. Mean WCL traffic load versus k: Performance comparison between FO/G and BO/G.

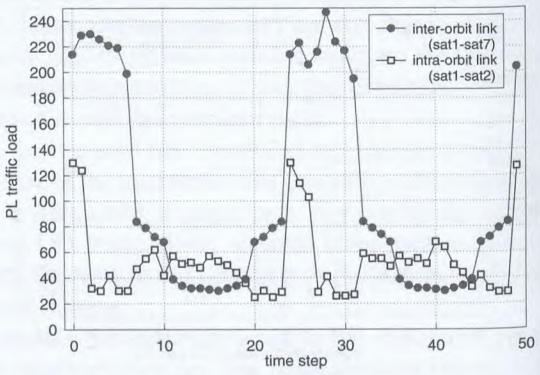


Fig. 11. Traffic load on selected physical links (PL) over one constellation period.

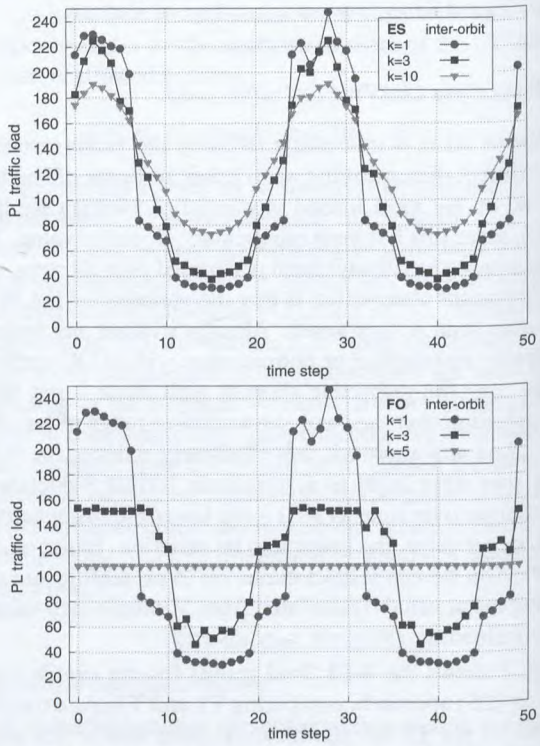


Fig. 12. Traffic load shaping on an inter-orbit PL through the (a) ES and (b) FO approaches with varying k.

(PL) over the steps. Fig. 11 displays the load over time for two selected PLs, one intra- and one inter-orbit link. Firstly, we see in both cases pronounced load peaks when the respective link is over mid latitudes, whereas it is low near the equator. Secondly, it becomes obvious that the WCL is always an inter-plane ISL. Concluding from both observations, WCL load reduction should typically work by shifting traffic away from those inter-plane ISLs that are in the critical region at a certain step. Fig. 12 confirms this expectation and illustrates the superior performance of FO approaches with respect to this "link traffic shaping". For $k = 5$ the optimization leads to a completely constant PL load, and one can already expect from observation of this example link that this is the final limit for any optimization working on this topology – a conclusion which is in line with the shape of the FO curve in Fig. 9. Another impressive confirmation of these considerations is given by the PL load distributions over all PLs of a certain step, as displayed in Fig. 13 for step 0 (the shape of the curves being nearly identical for other steps, by the way). It becomes obvious how traffic from higher loaded links is "shifted" to lower loaded ones, and a complete balance is achieved for the inter-plane ISLs when FO with $k \geq 5$ is used. In reality, of course, a certain traffic is not just shifted from one link to another but OD traffic is shifted to other paths – sometimes to paths with more hops, so that we can expect a growing *average* PL load in the network with increasing k. This is one of the prices to be paid for WCL load reduction/minimization, and it is quantified in Fig. 14. Again, the superior performance of FO becomes obvious; whereas ES with a reasonably high $k = 10$ leads to an increase of more than 20%, the effect is nearly negligible for the FO counterpart.

D. Path Costs

The most important trade-off from a user or QoS perspective is of course between the WCL load minimization target (which is mainly important from a network performance and system cost viewpoint) and the correlated degradation in terms of path delay encountered. The average costs in terms of both, number of hops and path delay in ms, are summarized in Fig. 15 for selected approaches. One basically observes that with increasing optimization freedom (from ES over BO to FO) not only the WCL load target value, but also the average path delay can be reduced for a given k. Moreover, looking on the two most promising candidates identified above, namely FO/G and BO/G,$\alpha = 0.4$, both for $k = 5$, the increase of path delay is fairly low around 4-7% with respect to the reference case of routing each OD traffic completely over the shortest path alone.

VI. Conclusions

A systematic approach to perform the capacity dimensioning in LEO ISL networks with dynamic topology has been presented. Inclined Walker constellations have been

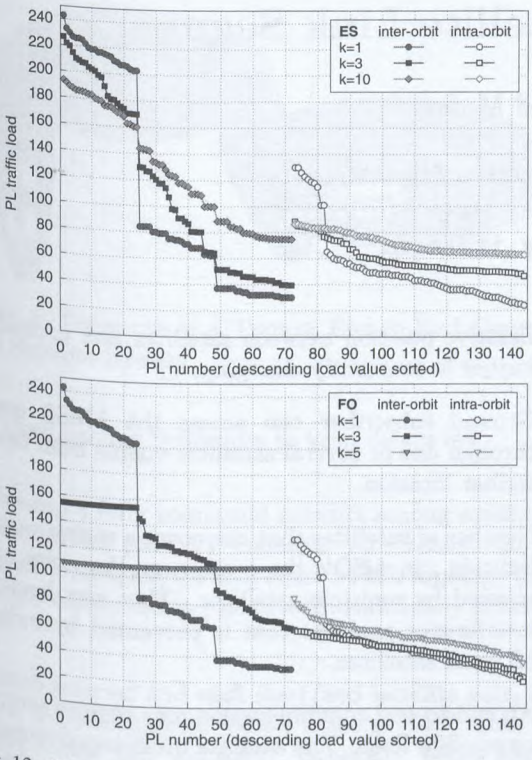


Fig. 13. PL traffic load distribution in the network at step 0 for (a) ES and (b) FO with selected k.

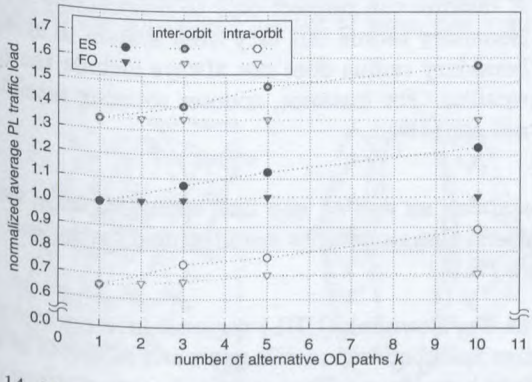


Fig. 14. Normalized average PL traffic load versus k, and corresponding splits into averages for inter- and intra-orbit links, respectively.

identified as appropriate candidates to implement convenient ISL topologies for connection-oriented operation. Defining a discrete-time virtual topology on top of the orbiting physical one, the combined routing and dimensioning task can be tackled with similar approaches as known from terrestrial ATM network design. Minimizing the worst case link capacity is an appropriate target function, which can be formulated as linear minmax optimization problem. Various heuristic and optimization approaches have been numerically compared under a given homogeneous traffic scenario. Optimization approaches clearly outperform heuristic methods with respect to the minimum WCL capacity target, while the increase of both, average load per link and average

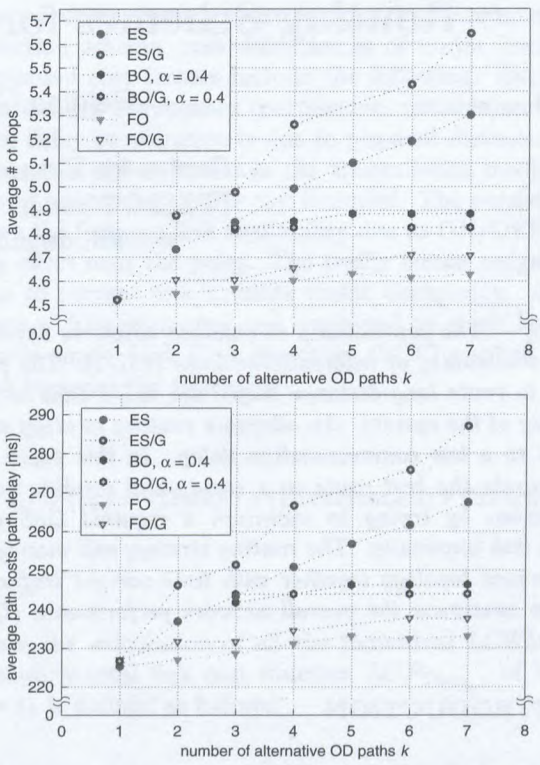


Fig. 15. Cost comparison in terms of (a) hops and (b) path delay for different (WCL) optimization approaches versus k.

path delay, is limited to acceptable values.

References

[1] M. Werner, C. Delucchi, H.-J. Vögel, G. Maral, and J.-J. De Ridder. ATM-based routing in LEO/MEO satellite networks with intersatellite links. *IEEE Journal on Selected Areas in Communications*, 15(1):69–82, Jan. 1997.

[2] M. Werner. A dynamic routing concept for ATM-based satellite personal communication networks. *IEEE Journal on Selected Areas in Communications*, 15(8):1636–1648, Oct. 1997.

[3] M. Werner and G. Maral. Traffic flows and dynamic routing in LEO intersatellite link networks. In *Proceedings 5th International Mobile Satellite Conference (IMSC '97)*, pages 283–288, Pasadena, California, USA, June 1997.

[4] M. Werner. Analysis of system connectivity and traffic capacity requirements for LEO/MEO S-PCNs. In E. Del Re, editor, *Mobile and Personal Communications, Proceedings 2nd Joint COST 227/231 Workshop*, pages 183–204, Florence, Italy, Apr. 1995. Elsevier.

[5] Motorola Satellite Systems, Inc. Application for authority to construct, launch and operate the M-Star system. FCC filing, Washington, D.C., USA, Sept. 1996.

[6] T. Bauschert. Multihour design of multi-hop virtual path based wide-area ATM networks. In V. Ramaswami and P. E. Wirth, editors, *Teletraffic Contributions for the Information Age, Proceedings 15th International Teletraffic Congress (ITC-15)*, pages 1019–1029, Washington, DC, USA, June 1997. Elsevier.

[7] B. Doshi and P. Harshavardhana. Broadband network infrastructure of the future: Roles of network design tools in technology deployment strategies. *IEEE Communications Magazine*, 36(5):60–71, May 1990.

[8] R. Siebenhaar. Optimized ATM virtual path bandwidth management under fairness constraints. In *Proceedings IEEE GLOBECOM '94*, pages 924–928, San Francisco, California, USA, Nov./Dec. 1994.

[9] K. Burchard. Application of virtual path concepts to the broadband LEO satellite system M-Star. Master's thesis, Technical University Munich, Institute of Communication Networks, Munich, Germany, July 1997.

Routing Schemes for Intersatellite Link Segment

Galdino Gutiérrez, David Muñoz

Center for Electronics and Telecommunications
ITESM, Monterrey, Nuevo León, 64849 México
e-mail: dmunoz@campus.mty.itesm.mx

Abstract – *The connectivity in satellite networks depends on the availability of intersatellite links (ISL's). The possibility to route long-distance traffic via ISL's adds to the flexibility of the system. An adequate routing strategy contributes to a low communication delay. In this paper we discriminate the best route in a end-to-end satellite communications by trying to maintain a required QoS and keeping link continuity. The routing strategy will manage a time-variant topology together with time-variant traffic in order to maximize the overall network performance (QoS and GoS).*

INTRODUCTION

[1] We can refer to the routing algorithm as the network layer protocol that guides packets through the communication subnet to their correct destination [1]. There are many forms to classify routing algorithms [2]. If we use static routing the path used by the session, in a end-to-end communication, is fixed regardless of the traffic conditions, and they will only change in response to a node failure. The adaptive algorithms try to choose a route by avoiding congested segments and nodes in the network. Costwise, the concatenation of segments that exhibits the overall minimum length must be selected. Note that the length of each link reflects not direct physical distance but also parameters such as congestion, delay, channel availability, etc.

II. DYNAMIC ROUTING IN LEO INTERSATELLITE LINKS NETWORKS

With emerging low-earth orbit (LEO) satellite communication technologies, it will be feasible to give global personal broadband communication. Nevertheless, there are some issues not yet addressed adequately. In this paper we address the issue of global traffic flow over dynamic network topologies. Networks with intersatellite links and polar orbits can implement routes under the concept of Virtual Paths (VP's) which exist in ATM networks and are called fully connected.

Among the principal aspects to be taken into account by the routing algorithm in these kinds of networks are :

- Relative position between satellites and ground subscriber is constantly changing.
- Ground subscriber can access the LMSS network through one or several satellites visible from the subscriber location.
- Number of satellites that can cover a region vary with latitude. In LEO's the areas near the poles can be covered by multiple satellites. This may lead to an interference problem that is prevented by switching off some satellites.
- An end-to-end communication may require several intersatellite links. The satellite mesh is connected (i.e. there is always a succession of links connecting any two satellites.
- A satellite can connect only to neighboring satellites. Boundary radius can vary from system to system. A boundary radius does not always include the nearest satellite. For instance, counter rotating satellites are not connected.

To establish an end to end link communication between two users, three segments are identified, as we can see in Figure 1:

1. The Up/Downlink (UDL) segment incorporates the respective links between mobile users and satellites as well as the links between satellites and fixed earth stations (Gateway GW's).

2. Intersatellite Link (ISL) segment.

3. Terrestrial Network Link (TNL) segment.

This study focuses on the ISL segment where system perform ON/OFF switching function near the poles [5] and counter-rotate satellite are not connectable.

The strategy followed to achieve optimal routing tackles the main issues, which we can list in this way:

- Establish ISL topology.
- Path search to establish a route among satellites.

[1] This work has been partially sponsored by Nortel Networks.

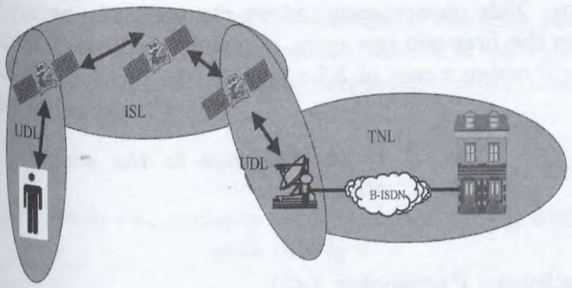
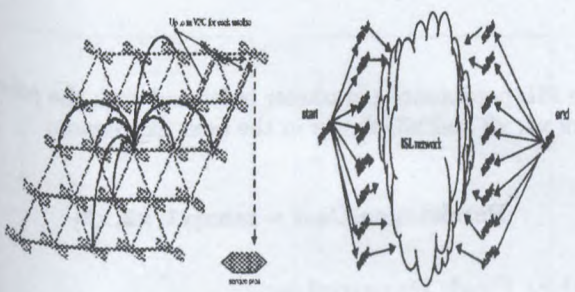


Figure 1: Segments of a Typical End-to-End Connection LEO Satellite System.

- Optimization procedure to keep delays low.

To develop a fully connected network among satellites, we use the virtual path VP concept. To establish the set of different topologies, we discretize the ISL topology changes during a constellation period with sufficiently small time-steps.

For each interval and each start/end satellite pair, a path search is performed to define a set of VPC links based on the actual ISL topology. Generally, every start/end satellite pair is connected via up to m link VPC's , m is the number of satellite links that can be established for each satellite, m is dependent of a connectivity of the underlying topology at the considered instant of time. See Figure 2.



(a) Path search

(b) Up to m Link for each satellite

Figure 2: Path Search in the ISL Network

Among the different approaches to establish the shortest path in the ISL networks, the most well known are Dijkistra Shortest Path Algorithm and Bellman-Ford algorithm [1].

These algorithms are widely used on terrestrial (computer) networks. They can be modified to be used in an MSS network. These two algorithms find the shortest paths from a given source node to all other nodes. In this paper, the Dijistra algorithm was considered adequate.

Since routing algorithms are based on a minimum cost path selection scheme, cost definition is of major importance. Proposed cost factors include the following: Delay propagation;ISL Persistence (permanent, variable)and Traffic. The delay propagation is due to physical distance. An information signal travels in the transmission medium. On board processing time is also included. The persistence relates the foreseen link availability due to ON/OFF switching effect near the poles. The traffic factor assigns a cost due to current link/satellite traffic occupancy. All these costs will be evaluated and combined on a per link basis, where single Link Cost Function (LCF) applied to each link becomes the following:

$$\mathbf{LCF(t_k) = (Delay) * (Persistence) * (Traffic).} \quad (1)$$

Once the calculation of all the individual LCF's is completed, a total link cost function $LCF_{Tot_{(t_k)}}$ of VPCj at $t = t_k$ is defined as followa:

$$\mathbf{LCF_{Tot_{(t_k)}} = \sum_{n=1}^{hops} LCF_{n_{(t_k)}}.} \quad (2)$$

These composed costs are used to evaluate and substantiate the optimal path selection process. The delay that is had in the switch is not significant, and we can include in the propagation delay.

Discussion of the Link Cost Function

It must be noted that link costs are time dependent as, for instance, traffic on a link will change from time to time; also, on/off switching affect the link availability, depending on the current satellite location. Delay also tends to be latitude dependent. Additionally, in the case of counter-rotating satellites, interconnecting links are prevented by assigning heavy penalizations, as these links exhibit tracking difficulties not present in co-orbiting planes.

Optimization Procedure

The fundamental objectives of the optimization procedure ([6],[7]) are as follows:

- Provide at a minimum cost and continuous end-to-end communication service, regardless of topological changes.
- The link assignment should be done in such a way as to maximize the overall network performance in coordination with current traffic demand.

- Handover between predefined or pre-selected VPC's must be guaranteed in order to avoid forced-connection termination. The number of handovers must be as low as possible.

All the effects and cost assigning function for each of the cost parameters must be characterized. We know, that an LEO's network has a periodical performance related to the orbital period (on/off switching, for instance). In order to simplify simulation time scale is discretizied. The expected availability or persistence become an important factor. In consequence, where satellite diversity is feasible, lower costs are assigned to the link with more time in line-of-sight (better visibility).

ISL Persistence Parameter

The ISL persistence parameter reflects the effect of network dynamism. It involves factors relative to the links and why they are permanent or time-dependent links, etc.

ON/OFF Switching Parameter (x1)

Persistence is also low for satellites near the poles where the on-off operation is performed (see [3]).Thus, it is advisable to avoid routing near polar locations. This is achieved by appropriate cost function, as shown in figure 5.

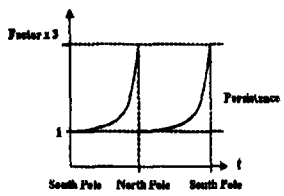


Figure 3: Function to Represent Persistence Parameter.

Since turned off satellites are not utilizable, an on-off parameter x1 can be defined as

$$\mathbf{x1} = \begin{cases} 1 & , \ liminf < \phi_1 < limsup \ , \\ & , \ liminf < \phi_2 < limsup, \\ \infty & , \ otherwise \end{cases}$$

Seam Parameter (x2)

The seam parameter concerns the counter-rotating effect. This is a technological problem because of the difficult task of tracking a satellite that moves in the opposite direction. This poses a major constraint because it generates a discontinuity that divides our network, affecting the delay factor. This phenomenon, as we can see, only occurs between the first and last orbit of our network. In this case, we will assign a cost of X2 :

$$\mathbf{x2} = \begin{cases} \infty & , \ if \ link \ belongs \ to \ the \ seam, \\ 1 & , \ otherwise. \end{cases}$$

Continuity Parameter (x3)

It takes charge to give the highest cost to links with less time in line-of-sight and to protect a call to be interrupted by switching off the satellites near poles. The function used for this respect is exponential function 3 and 4. The coef1 and coef2 parameters are used to give a soft change and to limit the function used.

$$func1 = \frac{e^{coef_1 * \phi_1}}{e^{coef_1 * 180}} * coef2 + 1, \quad (3)$$

$$func2 = \frac{e^{coef_1 * (-\phi_1 + 180)}}{e^{coef_1 * 180}} * coef2 + 1. \quad (4)$$

The $coef_1$ and $coef_2$ parameters establish the function of the cost to evaluate the continuity parameter as is shown in Figure 3.

$$\mathbf{x3} = \mathbf{func}_{1,2}(\theta_{1,2}).$$

The ISL persistence parameter is formed with the parameters x1, x2 and x3, shown in the next expression:

$$\mathbf{Persistence\_Cost} = \mathbf{max}[\mathbf{x1}, \mathbf{x2}, \mathbf{x3}]. \quad (5)$$

Delay Cost Parameter

The delay cost parameter is obtained by calculating the equivalent delay to the distance between two satellites in line-of-sight. It is clear that this delay is latitude dependent. A linear cost relationship is assumed.

$$\mathbf{Delay\_Cost} = \sqrt{(\mathbf{z2} - \mathbf{z1})^2 + (\mathbf{y2} - \mathbf{y1})^2 + (\mathbf{x2} - \mathbf{x1})^2}. \quad (6)$$

Traffic Parameter

The traffic parameter is assigned according to the periodically reported congestion of the network is conducted from those paths with greater available capacity. A function

that permits evaluation of the amount of capacity in the complete route takes into account the ATM concept of admission control with dynamic separation of services, as in the following:

$B_j(n_j) = b_j * n_j$ capacity used by n_j VC type j, with peak rate b_j

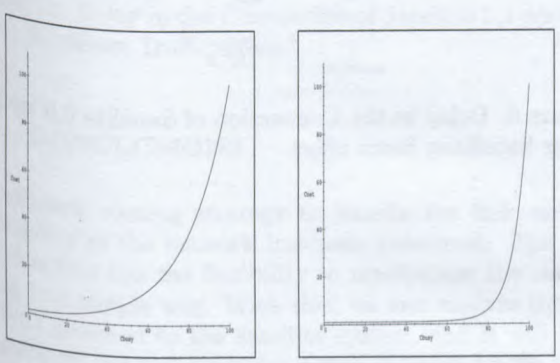
We accept other service, if the following inequality is valid:

$$B_j(n_1) + \quad ... + B_j(n_j + 1) + ... + B_k(n_k) \leq C_{VPC}.$$

The above equation evaluates the cost route of each link, based on the fact that selected path exhibits the highest average capacity (available circuits) in all the links that make up the path. The following proposed cost function is used.

$$\text{Traffic\_Cost} = \frac{e^{coeftraf_1 * (\%C_{busy})}}{e^{coeftraf_1 * 100}} * coeftraf_2 + 1. \quad (7)$$

The parameter $coeftraf_1$ and $coeftraf_2$ are used to limit and achieve the desired shape. The traffic cost function is illustrated in Figure 4, where $coeftraf_2 = 100$.



(a) $coeftraf_1 = 0.05$ (b) $coeftraf_1 = 0.15$

Figure 4: Functions Related to Traffic Parameter.

III. RESULTS

Proposed Scenario

In order to illustrate the routing scheme, a hypothetical network with 10 orbits and 10 satellites per orbit is considered. Proposed methodology can be applied to other ISL networks. In this network, we consider the start of the seam at zero longitude wehre the first and the last orbital planes are supposed to coincide. The seam will move west to east. That is, for each orbital period the seam moves east to another zone of coverage on the earth.

Traffic Effect

The principal results are in the aspect of avoiding congested areas, as we can see in the following examples. We are presenting the results obtained when certain links in the VPC used to connect two earth zones are blocked. The coefficients were selected to assign a high value to the links that have a busy capacity of over 80 percent. Here the algorithm clearly shows how to avoid the congested area by not using the blocked links, as is shown in Figure 5. We are supposing blocked links together in a route because we are simulating a congested area on earth that is grouped in certain links.

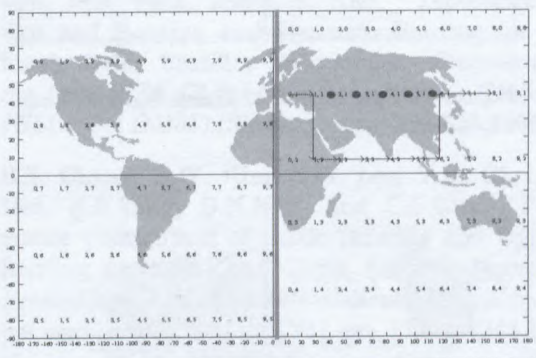


Figure 5: Connection of Zone 0,1-8,1, with Blocked Links.

Seam Effect

The effects of the discontinuity as a result of the seam effect can be appreciated when, in a certain step the seam moves and the path selected for a specific link changes.

This means that, according to the movement of the seam, the availability and frequency in the use of the links changes. This is due to the seam problem that makes certain connections that have to cross the seam use more hops to establish the connection. This is an important effect because, based on our simulations, we can predict when a zone is going to have a traffic overload. The prediction is based on the fact that the seam causes certain links to be used more often, and when the links are over a congestion area we have to manage the resources in the ISL network more efficiently.

The seam effect can be appreciated when we have to route through zones between orbit plane "0" and "9" in a certain longitude. This effect is not always presented. This

is when we are establishing a connection between points that have almost the same longitude, but for connections like zone 8,8-2,2, the conditions change because we have an increment in the number of hops used, as is shown in Figure 6. If the seam can be avoided and the links between satellites belonging to the seam were possible the connection will have a noticeable decrease in the expected delay, as we can see in Figure 7.

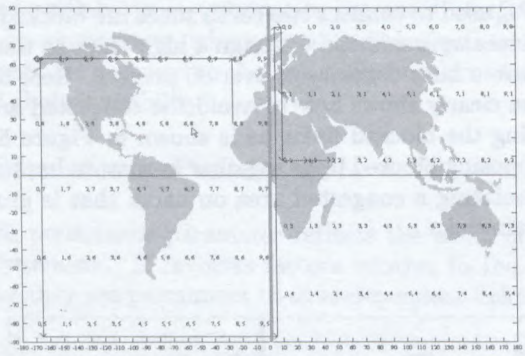


Figure 6: Connection of Zones 8,8-2,2, k=0

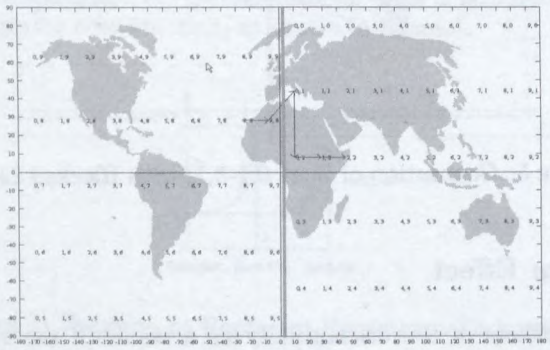


Figure 7: Connection of Zone 8,8-2,2, without Seam, k=0

Delay under Different Constraints

This section shows the change in the delay that is produced when we apply our algorithm under different constraints. This change exists because the least congested route can not be the path with minimum delay, and in the same context the route that presents the most continuity will not always be the route with less delay. If the ISL network did not have the time variant link availability, we would have an almost symmetrical network and the delay produced by the propagation in the network would be proportional to the minimum distance between two points in an almost direct path.

The graph in Figure 8 shows how the seam effect increases the delay produced when certain satellite connects with all other satellites. The point in the connection that show the increment is bacause for the seam effect, the hops used to reach the target are increased and the maximum delay is produced when satellites of different orbits are connected and the link avoids the seam effect.

Figure 9 shows the increase in the expected delay by trying to avoid the switching off and by keeping link continuity. The effect of the permanence effect has on links which use satellites that are in a short period of time is that the links are going to disappear because of the movement of the satellite.

This increase is more visible in satellites that are near the poles and that travel in direction of the pole. The delay in both cases is in range [3] of the delay found under VPC-HO constraints.

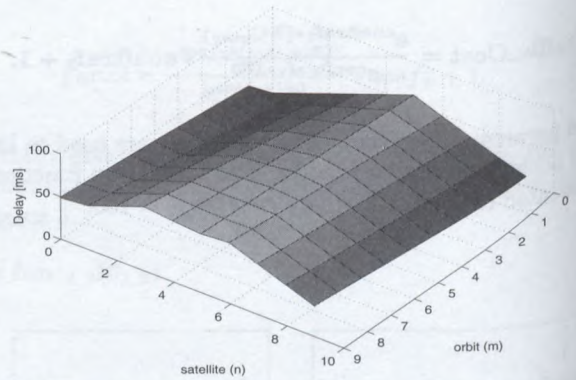


Figure 8: Delay in the Connection of Satellite 0,0 with all other Satellites; Seam effect

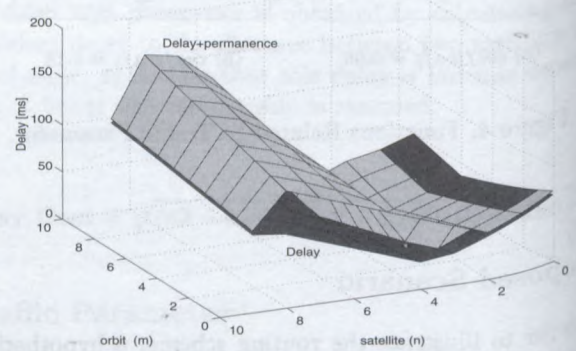


Figure 9: Delay in the Connection of Satellite 6,5 with all other satellites; Persistence effect

If we apply traffic to the ISL network, as was done in the traffic effect section. We can see that due to traffic constraints the times used in a certain link increase, because the traffic will be diverted to alternative routes avoiding the congested links. Figure 10 shows the delay suffered when the links are blocked, Figure 5 shows how the routing algorithm avoid the congested zone. In that case, the connection shown was the VPC between zones 0,1 and 8,1. Now we are going to connect the same zone cover by satellite 0,1 to all other satellites. In initial conditions the areas of connection on the earth are covered by a satellite with the same nomenclature as the zone to cover. In Figure 10, we can appreciate that certain VPC's that use the blocked links have an increase in the delay expected, and this delay increases in the points covered by the satellites that have that link connection.

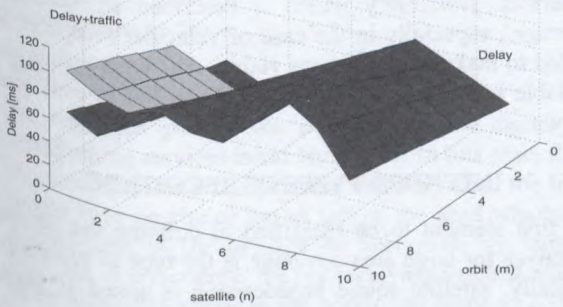


Figure 10: Delay in the Connection of Satellite 0,1 with all other Satellites; Traffic Effect.

IV. CONCLUSIONS

An efficient routing strategy to handle the link variant availability of the network has been presented. The link cost function has the flexibility to manipulate the cost of a link in a simple way. With this, we can analyze the capacities attached to the satellite system, and it will help us evaluate the concerns about link delay and how often the links are used under different constraint scenarios.

Based on the traffic measurement and location of the satellite the best route is selected, based on a cost link function that group the time-variant link availability.

ACKNOWLEDGMENT: Authors thank Payan Maveddat and Gerardo Donnis from Nortel Networks for multiple comments and suggestions.

REFERENCES

(1) Dimitri Bersekas, Robert Gallager, *Data Networks*, 1987 PRENTICE-HALL. INC.

(2) C-K Toh, *Wireless ATM and AD-HOC Networks*, 1997, Kluwer Academic Publishers.

(3) M. Werner, C. Delucchi, H.-J Vogel, G. Maral, and J-J. De Ridder. "ATM-based Routing in LEO-MEO Satellite Networks with Intersatellite Links," *IEEE Journal on Selected Areas in Communications*, Vol.15, No.1, pp.69-82, January 1997.

(4) M. Werner, G. Berndl, and B. Edmaier, "Performance of Optimized Routing in LEO Intersatellite Link Networks," *Proceedings of the 47th International Vehicular Technology Conference (VTC97)*, pp. 246 - 250, 1997.

(5) M. Werner,O. Kroner, and G. Maral,"Analysis of Intersatellite Links Load in Near Polar LEO Satellite Constellation," *International Mobile Satellite Conference (IMSC97)*,pp 289-294,1997.

(6) H.S. Chang, B.W. Kim, C.G. Lee, Y.H. Choi, S.L. Min, H.S. Yang, and C.S. Kim, "Topological Design and Routing and Dynamic Routing for Low-Earth Orbit Satellite networks," *Proceedings of the 1995 IEEE Global Telecomunications Conference (GLOBEC OM95)*,(Singapure),pp.529-535,1995.

(7) H.S. Chang, B.W. Kim, C.G. Lee, Y.H. Choi, S.L. Min, H.S.Yang, D.N.Kim, and C.S.Kim, "Performance Comparison of Static Routing and Dynamic Routing for Low-Earth Orbit Satellite Networks," *Proceedings of the 46th International Vehicular Technol ogy Conference (VTC96)*, pp. 1240 - 1243,1996.

Large Area Coverage for Digital Radio Broadcasting

Gérald Chouinard
Communications Research Centre
3701 Carling Ave. Box 11490, Station H, Ottawa, Ontario, Canada K2H 8S2
Email: gerald.chouinard@crc.ca

ABSTRACT

A review of the various types of system implementation for large area coverage by digital radio broadcasting is made. It ranges from pure satellite coverage to satellite coverage complemented by terrestrial re-transmitters, through to large area coverage provided by a network of terrestrial transmitters all fed from a satellite. The technical rules and constraints needed to secure a 'seamless' coverage in the case of a satellite coverage augmented by terrestrial re-transmitters are described. Using as references the satellite broadcasting systems proposed by WorldSpace (Afrisat, etc.), ESA (MediaStar) and the Association of Radio Industries and Business in Japan, ARIB (BSS sound at 2.6 GHz), the three main transmission techniques (TDM, FDM and CDM) are reviewed and compared to identify their respective merits. Finally, an optimum arrangement of transmission and receiver characteristics is proposed for a satellite broadcasting service complemented by terrestrial re-transmitters in view of providing the 'seamless' coverage required for portable and vehicular reception over large areas.

INTRODUCTION

Radio broadcasting from satellite has been studied since the early 1970's to provide coverage to large areas such as Canada, continental US or Europe. At the time, analog modulation (compressed FM) was considered. Later, digital modulation was introduced and proved to allow for better performance at lower satellite power. These services were to cover fixed, portable and vehicular reception.

As a result of almost two decades of work in considering the sound Broadcast Satellite Service (BSS) in the ITU-R, three frequency bands were allocated for this service at the World Administrative Radio Conference of 1992 (WARC'92). Some 40 MHz were allocated on a worldwide basis at 1.5 GHz except for the USA; 50 MHz at 2.3 GHz in the USA, India and later Mexico (WARC'97); and 120 MHz at 2.6 GHz in China, Japan, India, Russia and eight other countries in Asia.

As of now, some 35 sound BSS systems have been notified to the ITU. Three of the proposed systems are in active development, one intended to provide international coverage, especially for low latitude countries (WorldSpace) at 1.5 GHz, and two being developed to cover continental USA at 2.3 GHz (XM Satellite Radio and CD-Radio). A fourth system had reached an advanced state of development and was planned to provide a pan-European coverage at 1.5 GHz (MediaStar) but the European Space Agency (ESA) has put it on hold for the time being. There is also a system being developed in Japan by ARIB (i.e., Toshiba) for use in the 2.6 GHz band. In all these cases, there is recognition that a minimum level of service availability which is described as 'seamless coverage', especially in the case of vehicular reception, is needed to make these systems viable. Different means are available in trying to reach this goal and this is explained below.

INTENDED TYPES OF RECEPTION

The first element to be identified in defining the service objectives for large area coverage is the type of reception. Typically, satellite sound broadcasting is aimed at fixed, portable and vehicular reception.

Fixed reception

Reception by fixed installation is the least demanding requirement for sound BSS. It is relatively easy to find a place on the roof or the side of a building where line-of-sight reception to the satellite can be secured. In fact, higher frequency bands such as Ku band, used for TV-DBS for which no allowance for blockage is made in the satellite link budget, would be quite appropriate for such fixed reception since a more directional receiving antenna (e.g., parabolic antenna) can be permanently installed. Some satellite sound broadcasting systems directed at fixed receivers exist in various areas of the world using sub-carriers on existing C-band and Ku-band TV satellites.

Portable reception

When the service is to be directed at portable receivers, the assumption has to be that the receiving antenna will have a limited directivity and gain to reduce the need for accurate aiming. This is more demanding in terms of satellite power and, in order to maximize the effective aperture of the receive antenna for such a low directivity required for vehicular reception, a lower operating frequency is preferable. This is why the range 1.5-2.6 GHz was preferred for this type of service. Furthermore, the amount of possible blockage in the case of portable reception varies widely. For a "cooperative listener" (e.g., short-

wave type listening) who would take the necessary steps to bring his receiver in line-of-sight with the satellite, the fade margin on the satellite link can be reduced to a few dB's (e.g., 5 dB) to allow reception through typical tree canopy. In such case, the satellite power required is still within a reasonable range.

However, if the service is to be provided to listeners equipped with portable receivers for which line-of-sight reception is not always possible, the fade margin needs to be much larger (e.g., 25 dB) to allow the receiver to use signal reflections from neighboring buildings. If indoor reception is to be secured for these portable receivers, an additional fade margin needs to be included (e.g., 15 dB typical building penetration losses). Furthermore, if the service is to be provided to wearable devices (e.g., "Walkman™") for which the receiver RF front-end has to go through more stringent compromises, this would result into some further fade allowance (e.g., 5 dB). It is clear that in this case, the power requirement from the satellite would be excessive and other means to provide the required service availability have to be used as described in the following sections.

Vehicular reception

An important part of the market for such digital radio broadcasting services is car reception. In this case, there is no possibility for a "cooperative listener" since the listener would have no choice but to follow the road whether it is line-of-sight to the satellite or shadowed by buildings, hills, etc. Furthermore, the car-receiving antenna will have to be omni-directional in the horizontal plane to allow for consistent reception independent of the direction of the vehicle. This results in a limited antenna gain except for the cases where the satellite could be received from zenith in all locations. Vehicular reception in a city core is another level of difficulty where little can be done to improve the situation. Here again, a satellite alone cannot provide for adequate reception in all cases. A study sponsored by the FCC in the US [1] confirmed that the sound BSS systems have to provide for a 'seamless' coverage over the entire service area to allow for reliable vehicular reception.

IMPROVING SERVICE AVAILABILITY

Increasing the elevation angle at which the satellite signal is received is the main factor allowing a reduction of shadowing and consequently an improvement of service availability. This can be done, in the case of a low latitude country, by locating a geostationary (GEO) satellite at a longitude as close as possible to that of the center of the area to be covered. In the case of a higher latitude country, satellites on highly elliptical orbits (HEO) for which the apogee is located over the area to be served are more appropriate. However, a high elevation angle is difficult to maintain over extended coverage areas since lower elevation angles are experienced at the edges away from the satellite apogee. Even in the best case, i.e., signal reception from zenith, there is still a residual service outage that would be caused by highway overpasses, tunnels and downtown core, not to mention indoor reception [2].

In the work of the ITU-R, an approach to keep the satellite power at a reasonable level has been to include a fade margin of some 5 dB to cover for most of the fades caused by foliage, and to assume that other means would be used to compensate for attenuation due to blockage [3]. These other means are the positioning of the satellites at high elevation angles, the use of redundant satellites with a relatively large orbital separation to build diversity in the transmission, and finally, the use of terrestrial re-transmitters. Increasing attention is being put on the latter, since it is the only means by which these specific reception problems can be completely resolved. The advantages and constraints related to the use of terrestrial re-transmitters are described below.

SYSTEM CONCEPTS FOR LARGE AREA COVERAGE

A number of system concepts for large area coverage with digital radio broadcasting can be identified. They range from a satellite directly covering a service area (providing for fixed reception, some portable reception and limited mobile reception), to a satellite that is only used to provide the signal to a large network of terrestrial transmitters. In this latter case, any Fixed Satellite Service band can be used to feed these terrestrial transmitters.

In between, the satellite broadcast coverage is provided first over a large area and then complemented by a network of terrestrial repeaters progressively implemented to improve the coverage in shadowed areas (e.g., city and mountainous environments) and even for indoor reception. This has been called the 'hybrid' satellite/terrestrial approach. This concept was initially described by Ratliff et al in 1990 [4] and later refined at CRC [5,6,7,8]. The coverage can also be secured by the initial deployment of a terrestrial network of transmitters which is later complemented by satellite broadcasting where it no longer makes economic sense to do it terrestrially (e.g., rural).

Although the optimum arrangement depends on the size of the area to be served, the demographics, the topography and other aspects, there is currently a trend towards implementing digital radio broadcasting over large areas according to the third case, i.e., implementation of a broadcasting satellite complemented with terrestrial repeaters. We concentrate on this hybrid approach and identify the advantages and limits in its applicability, as well as the different technologies proposed for its implementation by the main players, in the following sections.

RULES AND CONSTRAINTS FOR SEAMLESS COVERAGE WITH HYBRID SATELLITE/TERRESTRIAL SYSTEMS

There are two ways to implement a hybrid satellite/ terrestrial digital radio broadcasting system: a) both satellite and terrestrial repeaters transmit on the same

frequency, and b) satellite and terrestrial repeaters transmit on two different frequencies in the same broadcast band.

Same frequency for satellite and terrestrial transmission

The two key advantages of this approach are that it is spectrum efficient and that it allows for the use of a single receiver connected to both the satellite and the terrestrial receive antennas to provide a seamless reception. This is possible if the receiver can operate adequately in the 'multipath environment' created by this double reception. For the purpose of this discussion, it is assumed that the EU-147 DAB system is used. This system was developed especially to remove inter-symbol interference (ISI) caused by signal multipath using a Coded-Orthogonal-Frequency-Division-Multiplex modulation (COFDM) which includes a Guard Interval (GI) at the beginning of each transmitted symbol (containing most of the ISI), during which the received signal is discarded [3, 11]. The performance of this modulation scheme is compared to two other transmission techniques in the following sections.

There is an intrinsic limitation on the terrestrial repeater power to avoid affecting the satellite coverage in fringe areas. The cause of the potential problem is that there is a distance from the terrestrial repeater where the signal generated by this repeater can impair the reception of the satellite signal. This occurs when the repeated signal creates an echo of sufficient amplitude falling outside of the GI, therefore creating an excessive level of ISI. The area over which this phenomenon can occur is a crescent-shaped area near the terrestrial repeater in the direction of the satellite [3,5].

TABLE 1: Maximum ERP from a terrestrial on-channel repeater using an omnidirectional re-transmit antenna to avoid interference into the satellite coverage.

EHAAT [m]	MODE II (GI= 62 µs)		MODE IV (GI = 123 µs)	
	Elevation angle		Elevation angle	
	30°	60°	30°	60°
25	110 W	300 W	4000 W	10000 W
50	20 W	70 W	1000 W	2300 W
100	4 W	18 W	250 W	700 W
200	< 1 W	3 W	40 W	180 W

The way to eliminate this destructive area is to keep the power of the terrestrial repeater in the general direction of the satellite below a specified level. Typical maximum effective radiated power (ERP) values for terrestrial repeaters using omnidirectional transmit antennas were developed at CRC and are given in Table 1 as a function of the Effective Height Above Average Terrain (EHAAT) of the transmit antenna, the elevation angle at the satellite and the width of the GI, which is related to the transmission mode in the case of the EU-147 DAB system [6]. Much more powerful on-channel repeaters can be used if directional transmit antennas aimed away from the satellite are used as shown in Figure 1 [9]. The feeding of these repeaters can be either from the satellite directly or from the closest repeater in the direction of the satellite through either RF pickup, RF microwave link or fiber optic distribution.

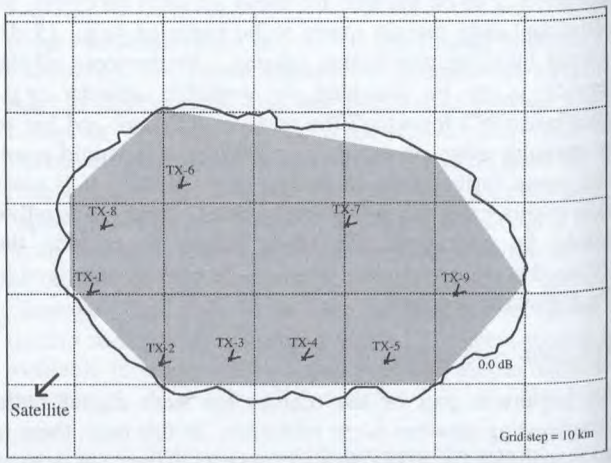


FIGURE 1: Service coverage of a given area for Eureka-147 DAB at 2.3 GHz requiring 9 directional repeaters of 8 kW ERP each at 100 metre EHAAT.

There is also a requirement for a very large isolation between the receive and transmit antennas of the terrestrial repeater. Such isolation can be achieved through a careful combination of local blockage from the surrounding physical structures as well as through controlled sidelobes for the transmit and receive antennas. Isolation of the order of 86 dB was successfully achieved at CRC using standard off-the-shelf equipment and it is felt that isolation greater than 105 dB could be achieved with custom developed antennas and careful siting [8]. As an ultimate measure, the isolation can be improved further by locating the receive antenna at a different site than the transmit antenna and providing the signal to the repeater through a microwave or fiber optic link [7].

Different frequencies for satellite and terrestrial transmissions

In this case, the spectrum requirement is twice that of the previous case and each receiver needs two RF front-ends tuned to two different channels. However, the power from the repeaters does not need to be restricted, as described above, to preserve the satellite coverage in the fringe areas of the repeaters. The repeater implementation is simpler because of the possible isolation improvement through channel filtering. This can even be avoided if the repeaters are fed by another satellite in a different frequency band. Although less spectrum efficient, this allows complete

decoupling between the feeding of terrestrial repeaters and the satellite coverage. When multiple gap-fillers are used, the same rules as for the implementation of terrestrial single frequency networks need to be applied [3].

Operating the satellite and the terrestrial repeaters on different frequencies also brings the possibility of using different transmission techniques on the two different broadcast channels. We will see later that this may actually be the overwhelming reason for using different frequencies for the satellite and the terrestrial portions of the system.

OPTIMUM SIGNAL TRANSMISSION FOR SATELLITE REPEATERS

A number of systems have been developed or are still being developed for satellite sound broadcasting. Studies conducted in the ITU-R and in WorldDAB* lead to the comparison of the best technologies to be used in the context of a sound BSS augmented by terrestrial repeaters. The main technologies considered are:

- Single carrier modulation using QPSK and time-division multiplexing (TDM) [10]
- Multicarrier modulation using QPSK and frequency-division multiplexing (FDM) [3,11], and
- Single carrier modulation using QPSK and code-division multiplexing (CDM) [12].

The merits of each of these technologies are summarized in Table 2. As can be seen, there is a clear advantage in using TDM over satellite because of its near constant amplitude QPSK modulation. The quasi-constant envelope modulation allows the transponder to be maintained close to saturation and therefore maximizes the power available at the output of the transponder.

The difference in required transponder output power between TDM and FDM can be explained as follows [13]:

- Difference between coherent and differential detection: 2.9 dB
- Presence of the guard interval in EU-147 DAB system: 1.0 dB
- Use of Reed-Solomon (223, 255) block coding in the TDM system: 0.6 dB
- Reduction in transponder output back-off for the single-carrier modulation: (TWTA) = 2.6 dB, (linear HPA) = 2.2 dB

This results in total differences of 7.3 dB and 6.9 dB for a TWTA and a perfectly linearized HPA respectively. The advantage of the CDM modulation as compared to FDM is reckoned to be about 3.9 dB because of the presence of coherent detection and the absence of a guard interval, in spite of the fact that the signal envelope of the CDM RF signal varies with time like in the case of FDM.

* *International Forum whose mandate is the promotion of the EU-147 DAB system (ETSI Standard ETS 300 401) for terrestrial and satellite sound broadcasting.*

However, the situation is quite different in the case of the terrestrial repeaters. There is a clear advantage for the FDM approach (EU-147 DAB system) because reliable reception in a multipath environment can be obtained by a much simpler receiver. As indicated in Table 2, COFDM only requires time synchronization that is sufficiently accurate to ensure that most of the energy from the various signal echoes (passive and active) fall within the GI. In the case of the two other techniques, complete channel estimation in amplitude, phase and excess delay is needed to allow signal recovery. In a vehicular situation, this complete channel estimation needs to be updated in real time and can require excessive computational power at the receiver considering the current state of technology.

Furthermore, in the TDM case where a time-domain channel equalizer is required, it is known that a single echo needs to be at least 6 dB higher than any other received signal to allow a typical decision directed feedback equalizer algorithm to converge [14]. Such condition cannot be maintained in a moving vehicle and reception drop-out is likely to occur. On this basis, it seems clear that FDM (EU-147 COFDM) is the best transmission technique to be used for the terrestrial re-transmitters.

Although the multicarrier modulation is sub-optimal in terms of satellite transmission due to its non-constant envelope signal and its differential detection, if one can afford the sizeable satellite power, FDM is the best overall transmission technique for satellite and terrestrial repeaters from the spectrum efficiency point of view. On the other hand, if the satellite power requirement is considered to be excessive and only one frequency is available, CDM becomes preferable since it allows for a reduction in satellite power at the cost of more complex receivers. However, if two different frequencies can be used, the best arrangement is for TDM transmission to be used over satellite and FDM transmission to be used for the terrestrial repeaters.

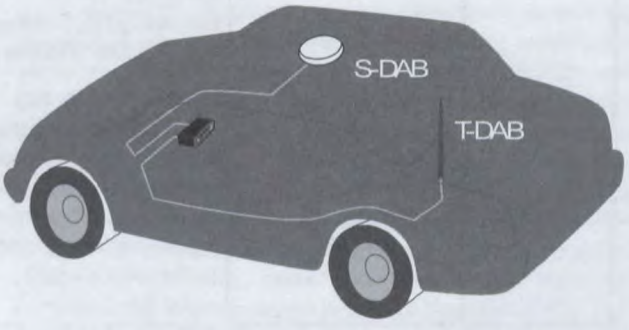


FIGURE 2: Vehicle equipped for S-DAB and T-DAB reception

RECEIVER GATEWAY ANTENNA

To allow the reception of two different modulations by a Digital Audio Broadcasting (DAB) receiver in the context

TABLE 2: Comparison of the merits of the three proposed modulation systems for satellite and terrestrial Digital Radio Broadcasting

	Parameters	WorldSpace Single-carrier modulation	MediaStar EU-147 DAB multi-carrier modulation	Toshiba-ARIB Coded division modulation
Satellite reception	Modulation	QPSK-TDM	QPSK-COFDM	QPSK-CDM
	Data throughput	1 584 kbit/s	1 152 kbit/s (FEC=1/2)	40 * 236 kbit/s
	Channelling bandwidth	1 840 kHz	1 536 kHz	25 MHz
	Required Eb/No in AWGN channel (BER=10^{-4})	2.7 dB	7.2 dB	4 dB
	Advantage of the modulation on the satellite link budget	7.3 dB (TWTA) 6.9 dB (linear HPA) (constant envelope modulation, coherent demodulation)	0 dB (reference point) (multi-carrier modulation, guard interval, differential demodulation)	3.9 dB (coherent demodulation)
Terrestrial reception	Impact of terrestrial repeater on receiver complexity.	Need complex channel estimation and time equalizer before coherent QPSK demodulator [9].	COFDM requires a partial FFT in the receiver. It was optimized for terrestrial multipath propagation.	Need de-spreading code acquisition and tracking, complex channel estimation and RAKE receiver [9].
	Level of channel estimation required in the receiver (complexity).	Complete channel estimation (amplitude, phase, excess delay) is needed for the operation of the channel equalizer in a multipath environment.	Adequate timing of the start of the symbol is needed for FFT windowing. Most of the energy from the impulse response of the multipath channel has to be contained within the guard interval.	Complete channel estimation (amplitude, phase, excess delay) is needed for the operation of the RAKE receiver in a multipath environment.
	Vehicular reception	Computation complexity increases rapidly with variability of the channel (Doppler spread proportional to vehicle speed and carrier frequency) and extent of equalization window [9].	Possible trade-off between immunity to Doppler spread (for a given vehicle speed and carrier frequency) and extent of the guard interval in selecting the system transmission mode.	Computation complexity increases rapidly with variability of the channel (Doppler spread proportional to vehicle speed and carrier frequency) and time window covered by the RAKE receiver.
	Use of multiple on-channel repeaters	Possible but creates instability in the equalizer if repeater signals reach the receiver at the same amplitude. Maximum repeater spacing is related to the extent of the equalizer window.	Maximum repeater spacing is related to the extent of the guard interval.	Possible but there could be ambiguity in echo excess delays when it is beyond the symbol period. Maximum repeater spacing is related to the extent of the time window covered by the RAKE receiver.

of a hybrid sound BSS system, DASA (Daimler-Benz Aerospace AG) proposed a 'gateway antenna' to provide for the coupling of a TDM satellite transmission to a terrestrial DAB receiver optimized for FDM terrestrial transmission using the EU-147 DAB system [15]. Such a 'gateway antenna' would be an active antenna containing all the electronics to demodulate the QPSK signal from the satellite, re-multiplex part of the received data and re-modulate it into a EU-147 DAB compatible signal receivable at VHF by a Terrestrial DAB (T-DAB) receiver.

Further work was done on this approach at CRC and was contributed to the WorldDAB Module 4 discussions. The

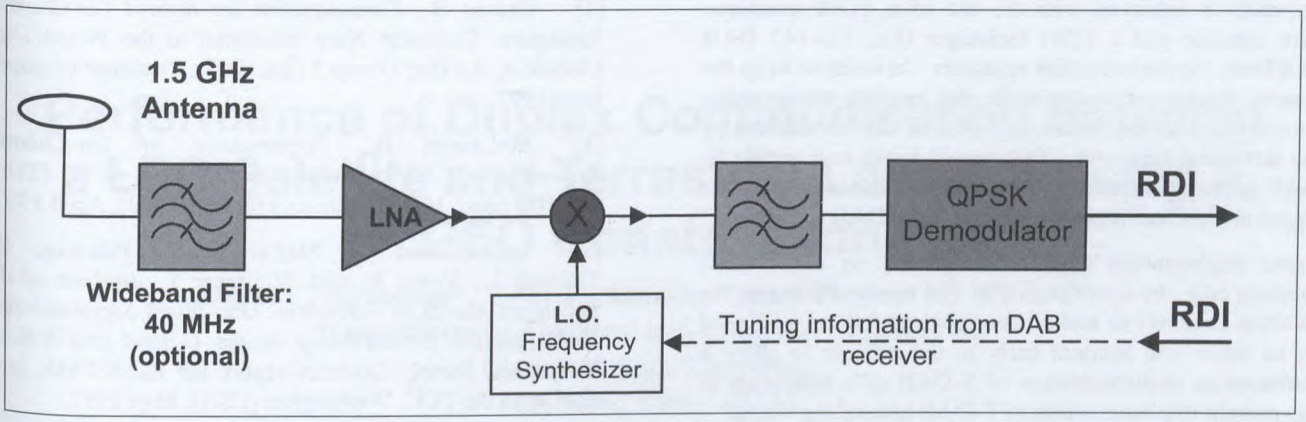


FIGURE 3: Simplified S-DAB gateway antenna assuming common multiplex structure on S-DAB and T-DAB and baseband data input to the T-DAB receiver

aim was to review this 'gateway antenna' proposal and try to reduce its complexity to a minimum [16]. Figure 2 shows how the reception of a hybrid satellite/terrestrial broadcast service impacts the implementation of the receive antennas on a car. Figure 3 shows a block diagram of the necessary electronics for the proposed simpler version of the 'gateway antenna', which is to be physically located at the base of the Satellite DAB (S-DAB) phased array antenna.

The terrestrial repeaters are to use the EU-147 DAB transmission format to counter multipath in the terrestrial environment, and it is assumed that the same data multiplex structure will be used over the satellite channel including FEC, time interleaving and the Fast Information Channel [3]. This will allow the gateway antenna that is to be added in front of the T-DAB receiver for S-DAB reception to simply become a coherent QPSK demodulator.

To make this simpler gateway antenna possible, it is also assumed that the T-DAB receiver will accept the bit stream from the gateway antenna directly through its bi-directional Radio Data Interface (RDI). This way, the data stream can be fed directly to the T-DAB receiver at baseband or through an infrared link. A second requirement is that the T-DAB receiver can provide for S-DAB channel tuning through its man-machine interface and send the control back to the gateway antenna, therefore ensuring proper remote control of the satellite channel tuning. The T-DAB receiver also needs to supply DC power for the operation of the active gateway antenna.

If these interface requirements can be met in the manufacturing of the T-DAB receivers, the S-DAB gateway antenna can become a relatively simple optional add-on that could be acquired later when the satellite service becomes available. Conversely, for the S-DAB providers, the launch of the service would be facilitated since they could rely on the fact that the T-DAB receivers already on the market could be easily upgraded for satellite reception through the addition of this gateway antenna.

It should also be noted that the use of the EU-147 DAB multiplex on the single carrier QPSK signal transmitted over the satellite allows for a comparatively simple implementation of the terrestrial repeaters since the demodulated data stream only needs to be re-modulated through a FFT before being frequency converted and transmitted.

CONCLUSION

The most challenging reception environments for sound BSS are the portable, indoor and vehicular reception. Beyond trying to secure the highest possible elevation angle from the satellite over the whole service area, the reception availability can only be improved further by the use of complementary terrestrial repeaters. Seamless coverage can only be achieved through hybrid satellite/terrestrial operation.

If spectrum is at a premium, the optimum solution for providing wide area coverage from satellite in a seamless fashion is the use of the multicarrier modulation technique (i.e., EU-147 DAB) over the satellite and its re-transmission by terrestrial repeaters. This costs some 7 dB in satellite power compared to a single carrier modulation but allows the use of the same frequency on satellite and terrestrial transmissions and the same receiver to capture the program from either satellite or terrestrial transmission. The other possibility is the use of the code-division multiplex (CDM) system which would provide a possible reduction of 4 dB in satellite power over the multicarrier technique at the cost of more complex receivers. In such cases, care has to be taken in implementing the terrestrial repeaters because of the fact that the same frequency is being re-used. There are some general rules that need to be applied with respect to the maximum allowable ERP of the terrestrial repeaters, the directivity of their transmit antennas, their physical location and the isolation that needs to be provided between the receive and transmit antennas.

If two different frequencies can be used for satellite transmission and terrestrial augmentation, the overall

optimum is achieved with the use of a TDM technique over satellite and a FDM technique (i.e., EU-147 DAB COFDM) for the terrestrial repeaters. In order to keep the system implementation simple, the satellite transmission should carry the same data multiplex as that transmitted by the terrestrial repeaters. This would result in a simple S-DAB gateway antenna that would demodulate the satellite signal and deliver it at baseband to the T-DAB receiver.

Some requirements need to be placed on the T-DAB receiver (e.g., bi-directional RDI and remote tuning of the satellite channel) to make it possible, and these will need to be taken into account early in the process to allow a harmonious implementation of S-DAB as a follow-up to the current implementation of T-DAB around the world.

REFERENCES

[1] Report of the Satellite Digital Audio Radio Service Pioneer's Preference Review Panel: Request for comments to the FCC. "*Evaluation of Pioneer's Preference Applications to the FCC that were Submitted by Three DARS Applicants*", Report No. SPB-67, November 19, 1996

[2] Hoeher P. et al, *Helicopter emulation of Archimedes/Mediastar satellite DAB transmission to mobile receivers*, International Journal of Satellite Communications, Vol. 15, pp. 35-43 (1997)

[3] ITU-R Special Publication, *Terrestrial and satellite digital sound broadcasting to vehicular, portable and fixed receivers in the VHF/UHF bands*, Radiocommunication Bureau, Geneva, Switzerland, 1995.

[4] Ratliff P.A., Pommier D. and Meier-Engelen E., *The convergence of satellite and terrestrial system approaches to digital audio broadcasting with mobile and portable receivers*, EBU Review-Technical, Nos. 241-242, June/August 1990, pp. 82-94.

[5] Chouinard G., Voyer R. and Paiement R., *Coverage Concepts for Digital Radio Broadcasting at 1.5 GHz*, Second International DAB Symposium, Toronto (Canada), March 1994.

[6] Voyer R. and Breton B., *Study of the Hybrid Coverage Concept at 1.5 GHz*, Document submitted to the ITU-R WP 10-11S (doc. 70), Geneva (Switzerland), November 1994.

[7] Breton B., *Consideration on Hybrid On-Channel Repeaters*, Technical Note submitted to the WorldDAB, Module 4, Ad Hoc Group 3 (doc. 007), Toulouse (France), June 1997.

[8] McLarnon B., *Performance of On-Channel Repeaters at 1.5GHz*, Document submitted to the ITU-R WP 10B (doc. 10B/40), Geneva (Switzerland), April 1997.

[9] Chouinard G., McLarnon B., Paiement R., Thibault L., Voyer R. and Whitteker J., *Analysis of the Technical Merits of Terrestrial Gap-fillers Supplementing DAR Satellite Broadcasting in the L-Band and S-Band Frequency Range*, Contract report for EIA/CEMA later filed with the FCC, Washington (USA), May 1997.

[10] Document 10-11S/27, 10 January 1998, *Digital System X*, USA contribution to the JWP 10-11S meeting held in Hawaii, January 1998.

[11] LeFloch B., Habert-Lassalle R and Castelain D., *Digital sound broadcasting to mobile receivers*, IEEE Transactions on Consumer Electronics, Vol. 35, No. 3, August 1989.

[12] Document 10-11S/85, 2 October 1998, *The Broadcasting-Satellite Service (Sound) Using 2.6 GHz Band in Japan*, Japan contribution to the JWP 10-11S, Geneva, October 1998.

[13] Chouinard G. and Le M., *Comparison between link budgets for Eureka-147 DAB system and the WorldSpace system for geostationary satellites*, Contribution to the WorldDAB Module 4 meeting (doc. M4/AH3/153), Cologne, March 1998.

[14] Proakis, John G., *Digital Communications*, McGraw-Hill series in electrical engineering. Communications and information theory. ISDN 0-07-050927-1, 1983.

[15] Kuhlen, H., *Gateway Antenna for Eureka-147 S-DAB, Proposal for consideration*, DASA, Contribution to the WorldDAB Module 4 meeting (doc. M4/160), Cologne, March 1998.

[16] Chouinard G., *Study of possible alternatives for the Antenna Gateway for S-DAB reception through the T-DAB receiver*, Contribution to WorldDAB Module 4 (doc. M4/AH3/162), September 1998.

Performance of Duplex Communication Between a LEO Satellite and Terrestrial Location Using a GEO Constellation

Daryl C. Robinson*, Vijay K. Konangi*, and Thomas M. Wallett**
*Department of Electrical and Computer Engineering
Cleveland State University
Cleveland, Ohio 44115

** Satellite Networks and Architectures Branch
NASA Glenn Research Center
Cleveland, Ohio 44135

ABSTRACT

A network comprised of a terrestrial site, a constellation of three GEO satellites and a LEO satellite is modeled and simulated. Continuous communication between the terrestrial site and the LEO satellite is facilitated by the GEO satellites. The LEO satellite has the orbital characteristics of the International Space Station. Communication in the network is based on TCP/IP over ATM, with the ABR service category providing the QoS, at OC-3 data rate. The OSPF protocol is used for routing. We simulate FTP file transfers, with the terrestrial site serving as the client and the LEO satellite being the server. The performance characteristics are presented.

INTRODUCTION

The International Space Station (ISS) is a LEO (low earth orbit) satellite which needs continuous communication with a terrestrial location so that services such as communications, tracking, telemetry and data acquisition can be provided. An extensive worldwide network of tracking and communication ground stations could provide this type of service. Since each ground station can communicate for very brief periods of time when the ISS is in line of sight, an elaborate terrestrial network of ground stations is necessary for global coverage [1]. The cost of maintaining, operating and upgrading this worldwide network is prohibitive.

An alternate approach to facilitate communication between the terrestrial location and the ISS is to use a constellation of GEO (geosynchronous earth orbit) satellites. The Tracking and Data Relay Satellite System (TDRSS) used by NASA represents such a system [2]. The TDRSS consists of three GEO satellites and a ground terminal facility located at White Sands, New Mexico. The system can transmit and receive data, and track a LEO user spacecraft for 100 percent of its orbit.

In this paper, we consider a constellation of three GEO satellites, with orbital characteristics similar to the TDRSS satellites, which can provide 100 percent global coverage. Unlike the TDRSS satellites, which are bent-pipe systems, the GEO satellites in this paper function as routers in a network. Our GEO satellites are also assumed to have inter-satellite links. A LEO satellite, such as the ISS, can communicate with the White Sands Ground Terminal (WSGT) via the GEO constellation [3]. Our objective in this paper is to determine the performance characteristics for communication between the ISS and the WSGT, using the GEO constellation. The communication is based on TCP/IP over ATM at OC-3. We present a comprehensive set of simulated performance characteristics -- throughput, end-to-end delay and server utilization -- for a range of FTP file sizes.

SATELLITE NETWORK

The network consists of a ground terminal at the White Sands Ground Terminal (WSGT), White Sands, New Mexico; three GEO satellites which provide worldwide coverage and the International Space Station in a LEO orbit. The WSGT is responsible for the command, telemetry, tracking, and control of the GEO constellation and the ISS. The three GEO satellites, GEO-1, GEO-2 and GEO-3 are

positioned over the Equator at 41° West, 275° West and 174.3° West longitude, respectively. These satellites are at an altitude of 22,300 statute miles (35,888 kilometers) and orbit geosynchronously. The GEO-1 satellite is in direct line-of-sight communication with WSGT. The simulation of the ISS, which is a LEO satellite, is based on the following orbital characteristics: semi-major axis = 6734.32 km, eccentricity = 0.0014064, inclination = 51.66°, right ascension of the ascending node = 243.89°, mean anomaly = 222.30° and argument of perigee = 137.91°. The network is illustrated in Figure 1

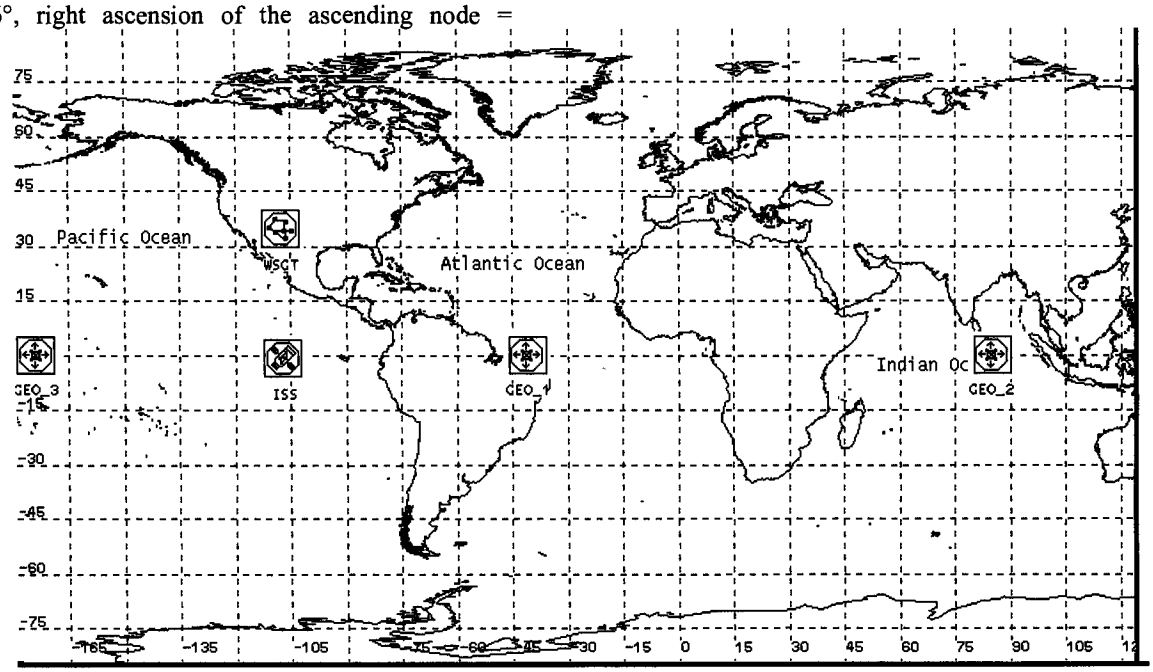


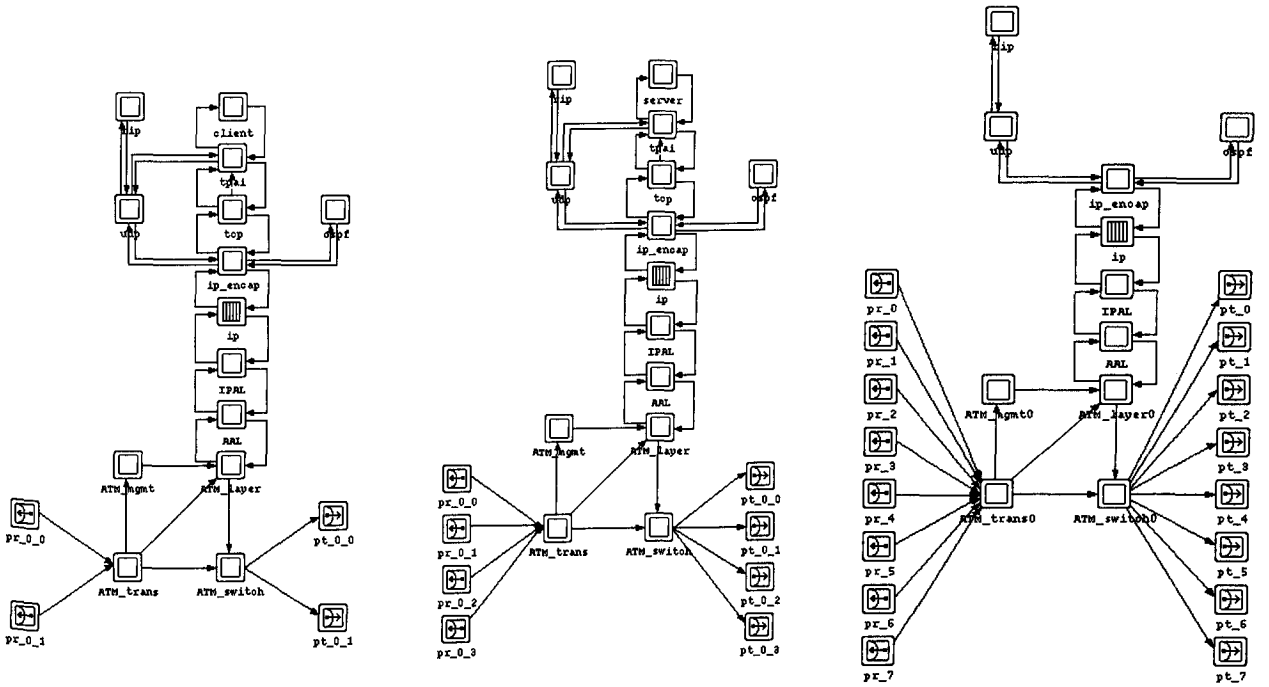
Figure 1

The ISS communicates with the WSGT through the GEO satellites. The topology is such that the WSGT communicates solely with GEO-1. The GEO-2 and GEO-3 satellites can communicate directly with GEO-1, but not with each other. The ISS communicates with the closest GEO satellite. All communication in the network between the ISS, the GEO satellites, and the WSGT is at OC-3 (155 Mbps).

NODE ARCHITECTURES

The node architectures of the terrestrial client, GEO router and the LEO server are depicted in Figure 2.

The WSGT and the LEO satellite communicate using a client-server paradigm of interaction. The LEO server application waits passively for contact, while the client initiates communication actively. The node architecture of the terrestrial site consists of the application-level FTP client using TCP/IP over ATM. The ATM layer uses the ABR (Available Bit Rate) service category. The OSPF (Open Shortest Path First) protocol is used for routing. The node architecture of the LEO is complementary to the terrestrial client, and has the application-level server. The three GEO satellites function as routers and their node architectures are comprised of IP over ATM, with OSPF being again used for routing.



Terrestrial client LEO server GEO router

Figure 2

TCP/IP OVER ATM

The TCP layer implements connection-oriented, reliable, byte stream transport using the potentially unreliable datagram service provided by the IP layer [4]. TCP is used to establish and terminate connections using three-way handshake protocols. Sliding window based flow control is used to prevent the transmitter from overwhelming the receiver with data. Reliability on an end-to-end basis is achieved by using acknowledgements. The retransmission time-out (RTO) is dynamically varied using Jacobson's algorithm. To avoid the retransmission ambiguity problem, Karn's algorithm is used. For congestion avoidance and control, the slow-start algorithm is used. The silly window syndrome is avoided using Nagle's algorithm.

The IP layer is a connection-less network protocol which enables the integration of heterogeneous networks. It provides an unreliable datagram service. Routing across multiple networks is the responsibility of the IP layer. The IP layer allows data to be interpreted consistently as they traverse the network.

The ATM is a streamlined protocol with minimal error and flow control features, which reduces the number of overhead bits for each cell and therefore the overhead involved in the processing of each cell [5]. This combined with the fixed-size cells of ATM enables it to operate at high data rates. The ATM layer provides connection-oriented, in-sequence, unreliable, and guaranteed quality-of-service cell transport. The ABR service category of ATM is intended for bursty traffic sources whose bandwidth range is known approximately. An application using ABR specifies the minimum cell rate (MCR) required and the peak cell rate (PCR) at which it will transmit cells. The network then allocates resources to ensure that all ABR applications receive at least their MCR capacity. ABR is the only service category in which the network provides explicit feedback to the sources, asking them to reduce the transmission rate in the presence of congestion and thus enabling the fair allocation of resources.

RESULTS

To investigate the performance of TCP/IP over ATM using ABR in a constellation of GEO satellites, we simulated file transfers using FTP. The requests for transfers are generated by the client at WSGT using a Poisson distribution with a mean of 5 requests per hour. Each request results in one TCP session, which transfers the file from the server on the ISS to the client. The average size of the files is modeled using

a normal distribution. Results are presented for a range of means: 60 KB, 300 KB and 1500 KB. The simulation results are for one-half day (43,200 seconds) of operation of the satellite network for the indicated file sizes. Since the orbital period of the ISS is 91.66 minutes, these simulations will involve at least 7.86 orbits.

In our simulations, we monitored the number of FTP requests submitted to the transport layer by the application layer of the client, and the corresponding number of FTP responses received by the application layer of the client. As both plots are nearly identical, Figure 3, we conclude that although the ISS is circumnavigating the earth in its orbit, the satellite network is functioning so as to allow continuous communication from the ISS to the terrestrial client even if there is no line-of-sight communication.

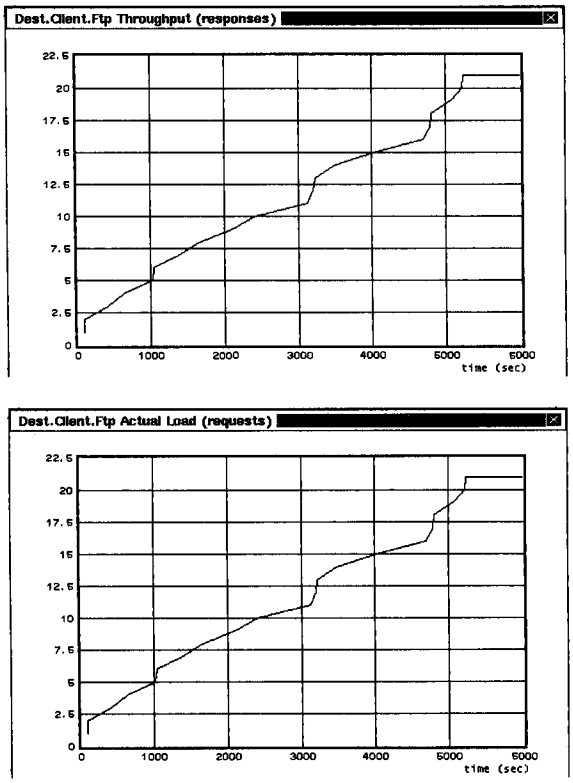


Figure 3

The conformity of the plots indicates that for every request sent by the client, a response follows shortly thereafter. This simulation is for 6,000 seconds, which is adequate time to simulate a single LEO orbit of 5,499.6 seconds. Additionally, we examined the end-to-end (ETE) delay, which is the time from the transmission of a request from the FTP application in the client to the time a response packet is received by the client. For this test we used a high mean file transfer rate of 10 files per hour and a small mean file size of 5 KB. The ETE delays were of the order of 200 ms, approximately the round-trip time delay in transmitting a message and receiving a response from a geosynchronous satellite. Figure 4 is a plot of the ETE delays in the scenario just described, for one-half day of operation (43,200 seconds).

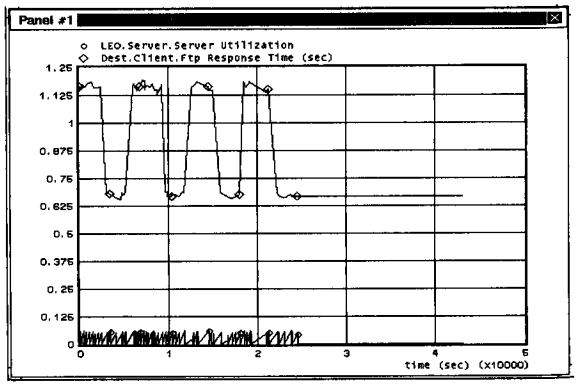


Figure 4

The fact that the server was utilized only for about 25,000 seconds was determined by the number of file transfers, which is a random variable, and the file sizes, which is also a random variable. The plot of end-to-end delay is oscillatory in accordance with the fact that the elliptical orbit of the ISS is also oscillatory. Points on the curve which are minima correspond to those times in the simulation when the ISS is closest to the GEO-1 satellite which communicates directly with the client. After the server stops transferring files at approximately 25,000 seconds, for the remainder of the simulation, the end-to-end delay is determined by the rate at which the client processes the accumulated files in its queue and hence the essentially constant nature of the delay. Any deviation from a constant end-to-end delay is not due to the randomly chosen file transfer sizes, which are centered about the mean; but it is due to the propagation delay through space.

The throughput represents the average number of bits successfully received or transmitted by the receiver or transmitter channel, as the case may be, per unit time in bits per second. At various points in time, the throughput is essentially the running average from the start of the simulation up to that point. The throughput for the communication path consisting of the LEO server, GEO-3 satellite, GEO-1 satellite and the terrestrial client at WSGT is shown in Figure 5. In Figure 5(a), the throughput shown is for communication between the transmitter on the LEO

server and the receiver on the terrestrial client for 60 KB files. This throughput is determined by the frequency of the requests for file transfers, the average size of each file and the data rate of the channel. As expected, the throughputs of the LEO transmitter and client receiver are almost identical, and the client receiver throughput is offset from the LEO transmitter throughput due to the propagation delay. Also, the client receiver throughput is slightly more than the transmitter throughput because the client receives duplicate packets from the GEO-2 satellite. In Figure 5(a), the sharp increases in throughput correspond to those periods when the LEO is in direct line-of-sight communication with GEO-3 and the server on the LEO is transferring a file to the terrestrial client. The throughput in the forward direction, i.e., from the client transmitter to the LEO receiver via GEO-1 satellite and GEO-3 satellite is shown in Figure 5(b). The LEO receiver has a higher throughput since it receives duplicate packets from the GEO-2 satellite.

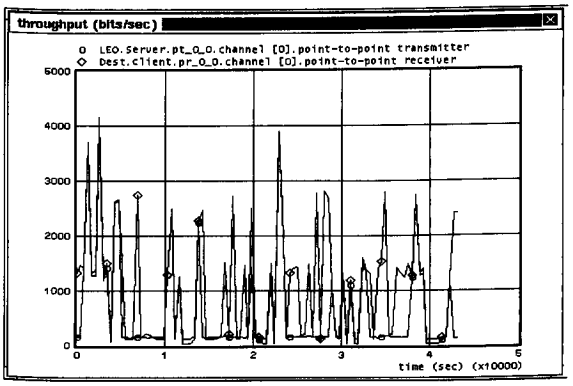


Figure 5(a)

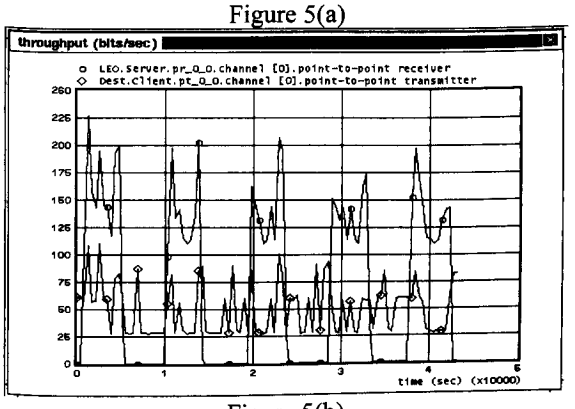


Figure 5(b)

Figure 5 – 60 KB Files

A similar set of results for 300 KB files is shown in Figure 6.

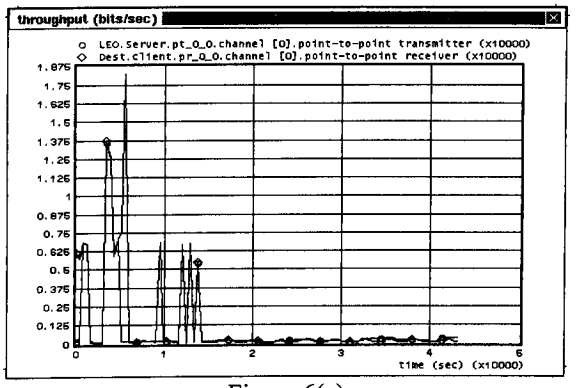


Figure 6(a)

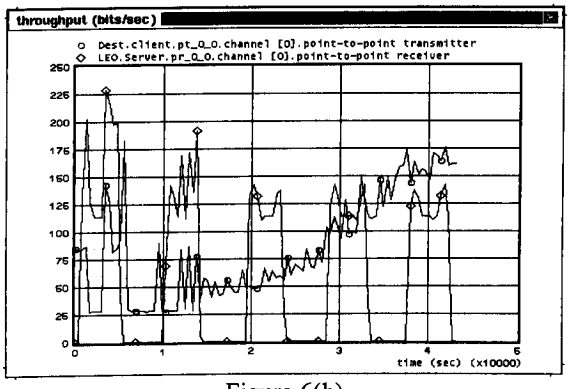


Figure 6(b)

Figure 6 – 300 KB Files

The FTP response time, which is the end-to-end delay measured from the initiation of a request from the FTP application in the client to the completion of the file transfer, is shown in Figure 7 for 60 KB file transfers. There are two reasons for the periodic variation of this statistic. First, the round-trip propagation delay from the terrestrial client to the LEO server depends on the orbit of the LEO satellite and its position relative the satellites in the GEO constellation. This delay has a periodic behavior. Second, the end-to-end delay is dependent upon the queueing and processing delays at the server. This is a function of the file size and the frequency of the file transfers. The FTP response time for 60 KB files varies from 3 seconds to 6 seconds. Since the throughput for 60 KB files is low in comparison to the data rate of the channel (Figure 5), the large end-to-end delay is due to the slow server, i.e., the processing and queueing delays in the server.

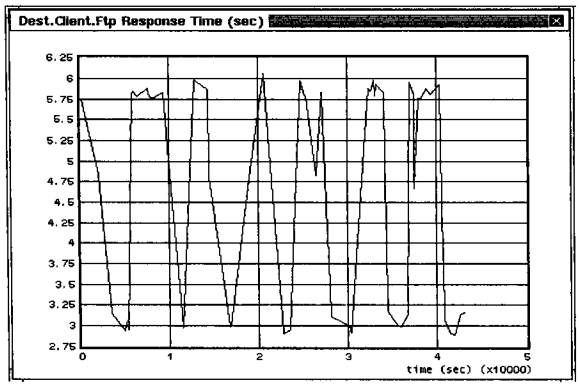


Figure 7

Figures 8 and 9 show the FTP response time for 300 KB and 1.5 MB files, respectively. As the file transfer size increases, the queueing and processing delays in the server lead to excessive end-to-end delays.

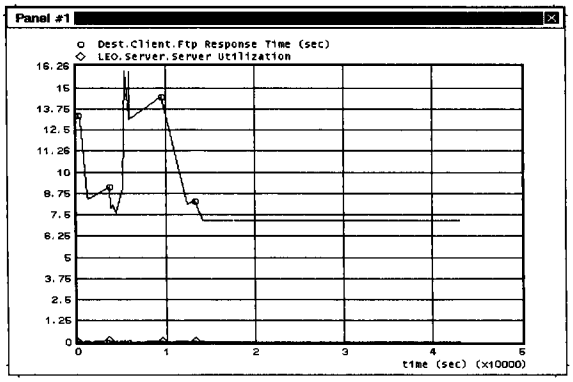


Figure 8

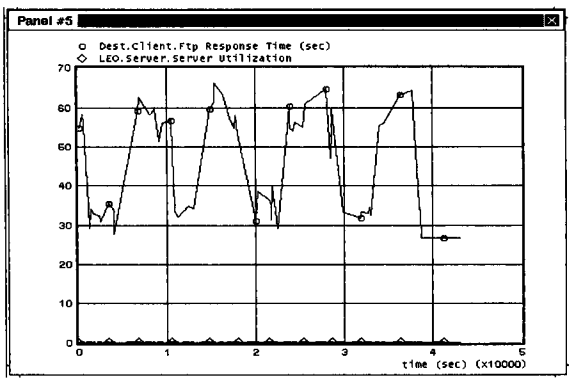


Figure 9

CONCLUSIONS

In this paper a network comprised of a terrestrial site, a constellation of three GEO satellites and a LEO satellite with orbital characteristics of the International Space Station was modeled and simulated. The communication in the network is based on TCP/IP over ATM with the ABR service category providing the QoS. The OSPF protocol was used for routing. We simulated FTP file transfers, with the terrestrial site serving as the client and the LEO satellite being the server. A comprehensive set of performance characteristics – throughput, end-to-end delay and server utilization – for a range of FTP file sizes were presented. When the file sizes increase, the end-to-end delays are quite large; this is due to the processing delay in the server. Since the orbital characteristics of the US Space Shuttle are similar to that of the International Space Station, we expect similar performance characteristics.

ACKNOWLEDGEMENT

The authors thank Dr. Kul B. Bhasin and the Satellite Networks and Architectures Branch of the NASA Glenn Research Center for the support of this research.

REFERENCES

[1] V.K. Konangi, T.M. Wallett, and K.B. Bhasin, "Modeling and Analysis of a Satellite-Terrestrial Hybrid Network," Proceedings of Sixth International Conference on Telecommunication Systems, Modeling and Analysis, 1998, pp. 427-432.

[2] http://www530.gsfc.nasa.gov/tdrss/oview.html

[3] C-K. Toh and V.O.K. Li, "Satellite ATM Network Architectures: An Overview," IEEE Network, September/October 1998, pp. 61-71.

[4] W. R. Stevens, TCP/IP Illustrated, Vol. 1: The Protocols, Reading, MA: Addison-Wesley, 1994.

[5] W. Stallings, High-Speed Networks: TCP/IP and ATM Design Principles, Upper Saddle River, NJ: Prentice Hall, 1998.

Integration of High Volume Data Collection Systems with Mobile Satellite Networks

Marina Ruggieri

University of Roma "Tor Vergata" - Dpt. of Electronics Engineering
Via di Tor Vergata, 110 - 00133 Roma - Italy
Email: ruggieri@uniroma2.it

ABSTRACT

The possible integration between a scientific satellite for data collection from remote areas and mobile integrated networks is analysed. It is found that, according to the universal communications aim, this integration could be usefully exploited from both sides. In fact, the experimental system could become a content provider in the frame of the UMTS (Universal Mobile Telecommunications System) architecture and, on the other end, existing mobile networks could be usefully exploited in the frame of the scientific mission operations. The potential synergy between the two contexts is investigated in the paper.

INTRODUCTION

The Italian Space Agency (ASI) has recently issued a "Call of Ideas" for gathering proposals about small-satellite-based scientific missions [1]. Only a few of those proposals has been selected for the phase A study, that has been concluded on December 1998 [2]. Among them, the telecommunications mission DAVID (DAta and VIdeo Distribution), proposed by F.Vatalaro, M. Ruggieri and A.Paraboni on an initial idea of F.Vatalaro, is presently in the process of being selected to be developed in the next years and likely launched in 2003 [3]+[7].

This missions aims at the deployment of two communications scientific experiments. One of them, proposed by the *Politecnico di Milano*, aims at testing the feasibility of a satellite system able to allocate dynamically the on-board power resources, accounting for meteorological data [3]+[7].

The other experiment (***D**ata **C**ollection **E**xperiment at **W**-band , **DCEW**), proposed by the *Universita' di Roma Tor Vergata*, aims at testing the feasibility of a satellite system for the exchange of large amount of data from a number of content providers to fixed users.

The system envisages a link in W-band and a Ka-band Inter-Satellite-Link towards the ESA *ARTEMIS* satellite, exploiting its gateway station in Redu to reach the finals users through INTERNET [3]+[7].

In the frame of the phase A study an important area of investigation has concerned the identification of experimental users in the *DCEW*, exploiting DAVID to collect their scientific data and forwarding them to specified final users. A very promising class of potential experimental sites are the scientific bases in the Antarctic region. Those sites, which belong to various countries with claims or simply scientific activities in the Antarctic area, usually store on tapes the scientific data gathered in the frame of the summer and winter campaigns. The tapes are then shipped to the original countries. In this frame, DAVID represents an interesting opportunity to the Antarctic sites to try a different way of transferring their scientific data at home.

The performed analysis show that the amount of scientific data produced daily, for instance, in the frame of the Antarctic Italian bases are fully compatible with the envisaged payload features (e.g. in terms of on-board storage capability) [8]. The highlighted capability of the experimental mission opens the way to various co-operation scenarios between DAVID-based systems and mobile integrated networks.

In the paper the above scenarios are explored, highlighting and quantifying the impact of this co-operation of the performance and complexity of the DAVID mission.

INTEGRATION SCENARIO: CASE 1

The first scenario of co-operation between DAVID and mobile networks moves from the basic service architecture envisaged for the UMTS, where the terrestrial and satellite mobile components will be highly integrated (Fig.1 [9]).

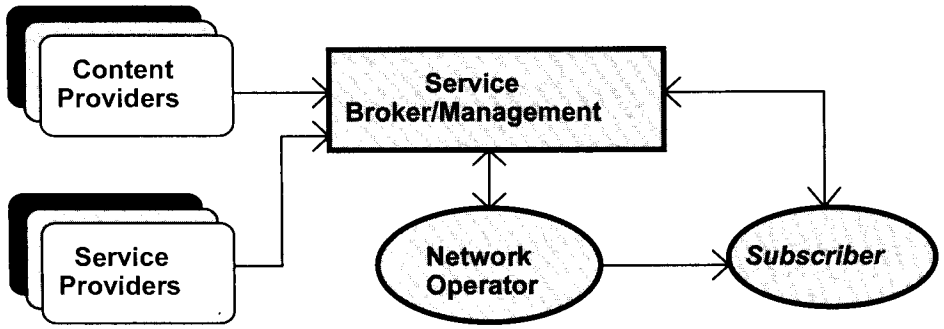


Fig.1 - UMTS basic architecture and interfaces [9].

A further piece of the mosaic is given by the highlighted capability of the DAVID system to collect, forward and locate on INTERNET huge amount of scientific data, produced in extremely remote areas of the Earth. In the previous section, the Antarctic stations have been pointed out as a very suitable application of the DAVID experimental service. Nonetheless, sites where Earth observation data are gathered by means of dedicated satellites or sky observatories located in remote mountains, as in the southern American countries, or maritime remote sites, such as oil tanker platforms, are further examples of possible remote sites where scientific data can be gathered and collected through DAVID.

The architecture of the *DCEW* is depicted in Fig.2, where the LEO/GEO Satellite Network includes the DAVID satellite, which is envisaged in low Earth orbit, the geostationary ARTEMIS satellite and the inter-satellite link between the two. The architecture envisages also links - not shown in the figure - in the opposite direction (2 Mbit/s), which can be in particular exploited to remotely control the experiments conducted in the remote sites.

The scientific data source represents the remote station where the data to be collected are gathered; the mirror provider offer an interface between the data transported through ARTEMIS to Redu and INTERNET.

The possible interaction between the architecture depicted in Fig.1 and the structure of the *DCEW* - and of a related future operational system - is the following: the architecture in Fig.2 is able to offer a volume of data not otherwise available, due to the remote location of the sites where the data are collected. In addition, the nature of this data is very peculiar, because they are scientific and related to particular experiments conducted in remote regions of the Earth.

Therefore, the *DCEW* could be seen as a content provider, that through proper interfaces - could be exploited by other networks and, in particular, by the UMTS architecture.

The above concept is sketched in Fig.3, where two of the parameters that characterise the DAVID-based content provider are highlighted:

- R_c is the data collection rate in the W-band up-link of the DAVID system;
- R_a is the available data rate at the ARTEMIS gateway.

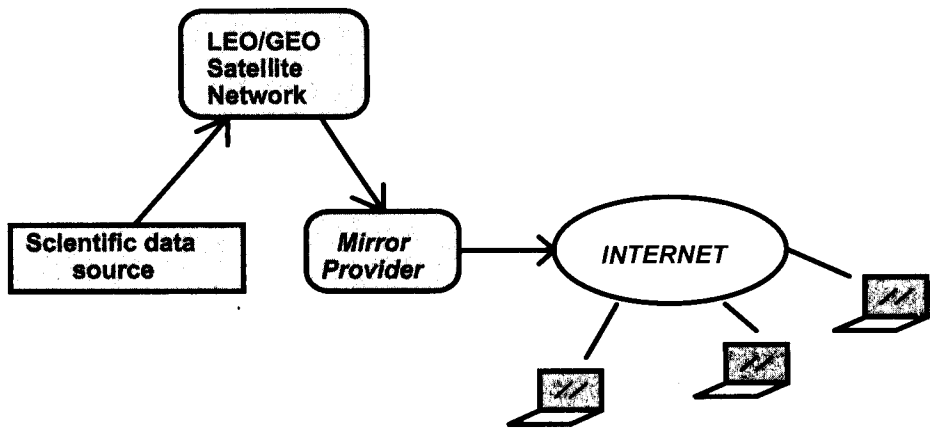


Fig.2 - Network architecture of the ***DCEW***.

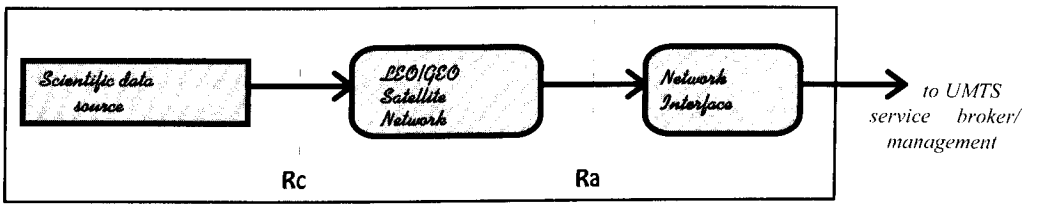


Fig.3 - Possible integration scenario with UMTS

In Fig.4 further parameters of the DAVID system, that can be used in its characterisation as content provider to the UMTS, are highlighted. In particular, they consists of:

- R_i : data rate of the inter-satellite link;
- V_d: data volume collectable per day through the LEO component.

In the envisaged architecture for the *DCEW*, $R_i=R_a$ due to the transparent operation of the ARTEMIS satellite.

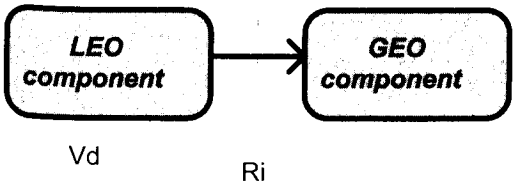


Fig.4 - Detail of the satellite components interface.

The data volume V_d (Gbyte) collectable per day depends on the time visibility window T_v (sec) per LEO satellite passage, the data collection rate R_c (Mbit/s) in the W-band up-link and the number N_p of satellite passages per day over the site location of the W-band terminal.

The time visibility windows is in the order of a few minutes, while the number of useful satellite passages per day (i.e. where the terminals sees the satellite with a suitable elevation that allows positive link margins) is one or two, depending of the site location.

In Fig.5 and Fig.6 V_d is displayed as a function of T_v and R_c for N_p =1 and N_p =2, respectively. The data rate of 120 Mbit/s is the baseline value identified in the phase A study. A storage of the collected data is envisaged on-board the LEO satellite and a data rate $R_i=R_a$ =8 Mbit/s has been identified in the phase A investigations.

A further analysis has been performed, relating the output performance of the DAVID-based UMTS content provider to the technological capabilities of the DAVID system.

In particular, in Fig.7 the achievable data collection rate in the W-band up link is evaluated as a function of the margin M that can be obtained - with respect to the baseline case with R_c=120 Mbit/s - through tighter requirement to the on-board and on-ground front-ends.

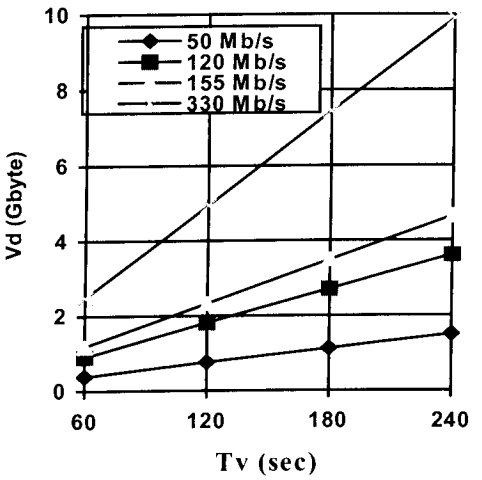


Fig.5 - Daily collectable data volume (N_p=1)

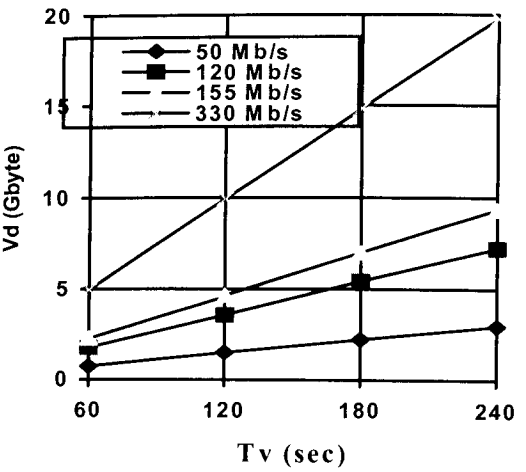


Fig.6 - Daily collectable data volume (N_p=2)

In Fig.8 and Fig.9 the data collection rate is more directly related to the W-band performance of the up-link.

In particular, R_c is related to the output power P (mW) of the on-ground terminal and the noise figure NF (dB) of the on-board receiver, respectively.

In Fig.8, the three curves refer to the assumption that the whole margin M or part of it (M/2 or 3M/4, M in linear units) can be obtained through tighter specification of the on-ground transmitter. The same approach is applied in Fig.9, but referring to the noise performance of the on-board receiver.

In the baseline case of R_c=120 Mbit/s, the on-ground power is 200 mW and the on-board noise figure is 8 dB.

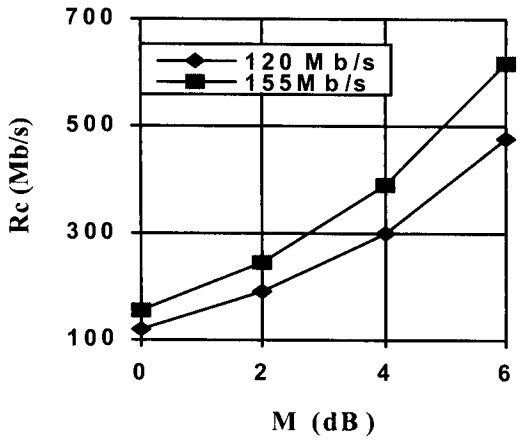


Fig.7 - Achievable data collection rate vs. technological capabilities

INTEGRATION SCENARIO: CASE 2

In the previously described integration scenario, the DAVID-system is seen as a potential provider of an integrated terrestrial-satellite mobile network. In this second scenario of possible co-operation between DAVID and mobile networks, the *DCEW* could exploit existing mobile satellite systems in order to support its ground operations.

In fact, the ground segment operations (satellite control, operations control centre, payload control centre) generally envisages an exchange of some of the data through terrestrial networks.

Nonetheless, the location of the W-band terminal in the Antarctic area or in other remote sites of the Earth, where terrestrial networks are not available, indicates that the operations information (such as, for instance, those necessary to correctly point the terminal antenna prior to the next useful satellite passage) can be transferred in two ways:

- through the return link of the *DCEW;*
- through other satellite systems covering the remote region of interest.

Considering the latter option and, for instance, the location of the W-band terminal in the Antarctic sites, a satellite system serving the poles could be exploited by the DAVID ground segment to the above mentioned purposes.

As some of the mobile satellite networks are envisaged to cover the poles (hence, in particular the Antarctic sites) providing low rate connections at reasonable cost, this could represent a further opportunity of integration at operations level between the *DCEW* of the DAVID mission and a mobile satellite network.

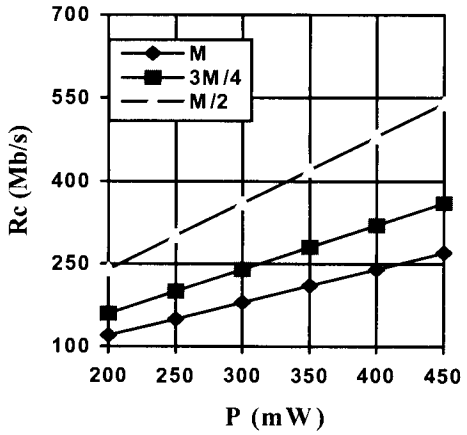


Fig.8 - Achievable data collection rate vs. power capabilities

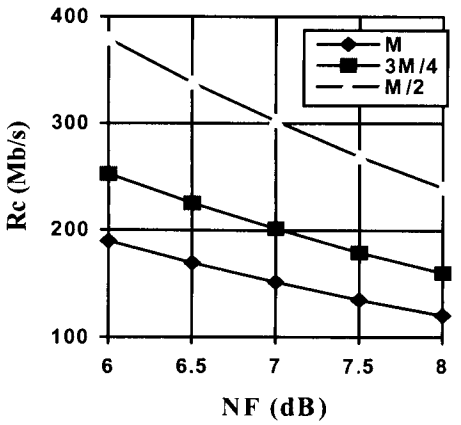


Fig.9 - Achievable data collection rate vs. noise capabilities.

CONCLUSIONS

The paper has highlighted two possible co-operation scenarios between a communications experiment of the scientific mission DAVID and existing or envisaged mobile satellite and integrated networks.

Both scenarios open a number of points that need investigations and further work.

In particular, the scenario where DAVID could offer a content provision to UMTS needs thorough investigations on the network interface issues.

The other scenario, where DAVID could exploit mobile satellite networks for ground operations needs investigations about the operations modes, the reliability of them and a possible degree of integration at terminal level.

Nonetheless, the highlighted scenarios indicates that through a proper integration level, the remote sites - where valuable data are gathered - could become part of the global communications mosaic.

ACKNOWLEDGMENT

The author wish to thank F.Vatalaro and A.Paraboni; C.Bonifazi, ASI; all the DAVID phase A team.

In particular: S.Greco, Alenia Aerospazio; E.Saggese, Space Engineering.

Further thanks to: M. Zucchelli, ENEA/PNRA; A.Dickinson and J. Sandberg, ESA-ESTEC/*ARTEMIS* program; S. D'Elia and E. Forcada, ESA-ESRIN, H.Pfeiffenberger, AWI; M.Thorley, BAS; M.Okada, NIPR/JARE; L. Anselmo, CNUCE.

REFERENCES

[1] Small Mission Call for Ideas, ASI, June 1997.

[2] S.Di Pippo, The ASI "Science Small Missions" Program, in Proceed. 4th International Symposium on Small Satellites Systems and Services Antibes-Juan Les Pins, September 1998, Session 3.

[3] A.Paraboni, M.Ruggieri, F.Vatalaro: Proposal of a Telecommunications Mission with a Low Earth Orbit Small Satellite, in Response to the ASI Small Mission Call for Ideas, June 1997.

[4] The DAVID Mission: Cost Report, ASI Contract n. ARS-98-6, September 1998.

[5] The DAVID Mission: Final Report, part I and part II, ASI Contract n. ARS-98-6, October 1998.

[6] M.Ruggieri, F.Vatalaro, A.Paraboni, C.Bonifazi: DAVID: A Small Satellite Mission for Data Distribution, in Proceed. 4th International Symposium on Small Satellites Systems and Services, Antibes-Juan Les Pins, September 1998, Session 3.

[7] M.Ruggieri, A.Paraboni, S.Greco, E.Saggese, C.Bonifazi: The Use of Ka-band in the DAVID Small Mission", in Proceed. 4th Ka-band Utilisation Conference, pp. 357-363, Venice, November 1998.

[8] M.Ruggieri, C.Bonifazi, S.Falzini, E.Saggese, F.Corbelli: Scientific Data Collection from Remote Areas through the DAVID Satellite 94 GHz Experiment, Proceed. DASIA '99, Lisbon, May 1999.

[9] The Path Towards UMTS Technology for the Information Society, UMTS Forum Report, n.2, 1988.

MSAT Dispatch Radio Service: "2-way" Trunked Radio Service throughout North America

Allister Pedersen
TMI Communications
1601 Telesat Court
P.O. Box 9826, Ottawa ON Canada K1G 5M2
Email:apedersen@tmi.ca

ABSTRACT

TMI Communication's MSAT-1 Network™ provides a comprehensive range of telecommunications services to mobile, transportable and fixed MSAT (Mobile SATellite) Communicators™ throughout North America, Mexico, the Caribbean, Central America and the northern parts of South America. Independent circuit-switched and packet-switched networks provide the basis for the end-user service offerings. Circuit-switched offerings include phone service, data (2400/4800 bps), fax, STU III (secure voice, data, fax) and group-oriented dispatch radio service. This paper provides a general outline of the MSAT Network and services with a focus on MSAT Dispatch Radio service.

The paper will describe the MSAT Dispatch Radio service which provides a push-to-talk, half-duplex group oriented voice service that replicates many features found in traditional terrestrial trunked-radio networks. The significant difference of this satcom implementation of dispatch radio is that a dispatcher can easily and simultaneously communicate with an entire fleet throughout the MSAT-1 coverage area.

The paper describes various features of the MSAT Dispatch Radio service including priority 1 interrupt, monitor codes, single beam vs. multi-beam coverage, talk group configurations, private mode, hangtime variables, as well as dial-in and dial-out dispatch capability.

The paper also outlines call setup/teardown and signaling channel usage that allows an individual mobile user to simultaneously monitor for incoming phone calls and dispatch radio conferences. The paper concludes with a discussion of the benefits of dispatch radio and typical end user organizations using MSAT Dispatch Radio.

INTRODUCTION

Coverage

TMI Communications own and operate the MSAT-1 geostationary satellite network which provides a comprehensive range of telecommunications services throughout a coverage area which embraces Canada, U.S.A., Mexico, Central America and part of Colombia and Venezuela. The MSAT-1 satellite located at 106.5° W longitude also provides coverage 200 n.m. off the east and west coasts of North America. Coverage is provided through 6 spot beams; four covering North America and one each covering the Caribbean and Alaska/Hawaii as shown below along with an overlay of elevation angles.

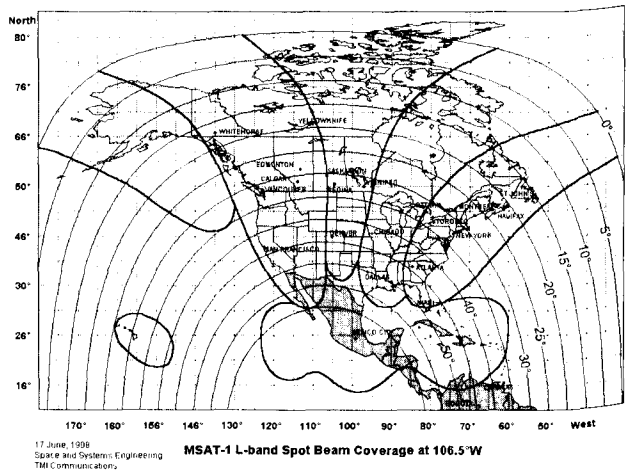


Figure 1-MSAT-1 Coverage

Line-of-sight radiocommunications networks, both terrestrial and satellite, are subject to terrain blockage and, depending on the radio frequency bands utilized, vegetation shadowing. Prior to commencing commercial service TMI Communications carried out coverage surveys in areas where terrain and vegetation shadowing were considered to pose serious limitations to coverage. It was theorized that MSAT-1 coverage would be poor in B.C., Canada's westernmost province, which is dominated by mountain ranges and is heavily forested. A coverage survey was conducted by driving 3,500 km of B.C. highways. Coverage was far better than expected; high quality signal was available along 93% of the total route driven from Vancouver to Prince George to Prince Rupert and through the Rocky Mountains to Dawson Creek and to Fort Nelson which is less than 150 km from the B.C./Yukon border. In fact MSAT offers excellent mobile coverage all the way to the Arctic Ocean as determined in February of 1996 when an MSAT-equipped vehicle was driven from Whitehorse to Tuktoyaktuk on the shores of the Arctic Ocean. In very general terms the land

mobile coverage surveys involved the logging of MSAT signal levels and GPS location on a laptop computer while driving selected highway segments. Using a standard MSAT mobile terminal and a spectrum analyzer, a total of 401 signal levels were measured and stored every second.

Services

Through a North American network of service providers and distributors, TMI Communications provides circuit-switched services, packet-switched services, asset tracking/monitoring service and a differential GPS service in B.C. in a partnership with the B.C. provincial government.

The circuit-switched network provides telephone service, data (2400 and 4800 bps), fax, STU III (secure voice, data, fax) and the dispatch radio service outlined below. The packet-switched network provides basic asynch and X.25 services as well as RTS (Reliable Transaction Service) for short ($\leq$ 64 byte) inquiry-response transactions and UDS (Unacknowledged Data Broadcast Service) providing 1 kilobyte packet broadcasts.

The newest TMI service is an asset tracking/monitoring service that provides for the monitoring of remote assets such as railcars, trailers, utilities and heavy equipment. The service provides information on the GPS location of the asset, speed and the status of sensors such as temperature, pressure and trailer door openings. The service provides for regularly scheduled reporting, event-triggered reports and polling. Data is available to end-users on a web site or is downloaded directly.

MSAT-1 satellite capacity is provided to the B.C. government who manage the provision of a real-time differential GPS service that offers 1-10 m accuracy in and around B.C. using active control stations located at Williams Lake and Terrace. The MSAT-1 DGPS receiver is a battery-operated handheld receiver connected to one of two antenna options; a patch antenna that is only 11.4 cm in diameter or a quadrifilar antenna that optimizes performance for lower elevation angles (10°-30°)

MSAT Communicators

Several manufacturers provide a comprehensive range of MSAT Communicators suitable for land mobile, aeronautical (fixed and rotary wing), marine, fixed and portable applications. A 2.4 kg notebook-size portable unit offers the full range of circuit-switched services. A hybrid unit offers both circuit-switched and packet-switched communications capability.

DISPATCH RADIO SERVICE

Service Description

MSAT dispatch radio is a half-duplex push-to-talk group-oriented telecommunications service that replicates much of the functionality found in traditional 2-way trunked radio systems that are often used for dispatch purposes. In addition to providing a one-to-many group-oriented communications conferencing there is the capability for private mode one-to-one operation.

Service Area(s)

Service area can be defined on the basis of the 6 beams provided by MSAT-1. End-users can request service for 1 beam or as many as 6 beams.

Call Setup and Signaling

To place a call (i.e., establish a conference) a dispatch radio service user such as the dispatcher (Figure 2) selects the desired talkgroup and keys the mike button on a PTT (push-to-talk) microphone. Signaling units identifying the type of call and identity of the talkgroup are transmitted to MSAT-1 and relayed to the TMI Communications ground segment in Ottawa. The ground segment assigns a communication channel for the call and transmits this information which switches the relevant mobiles in the selected talkgroup to the communications channel identified. Average call setup time is 2.3s. Visual and audio alerts indicate when the conference is established and the user can start speaking. The other members of the talkgroup do not have to take any action. The speaker's voice is heard automatically subject to the monitor codes assigned to the talkgroup in question on each Communicator. The speaker ID (4-digit number) of the party establishing the conference and talkgroup ID (2 digit number) are indicated on the Communicator handsets.

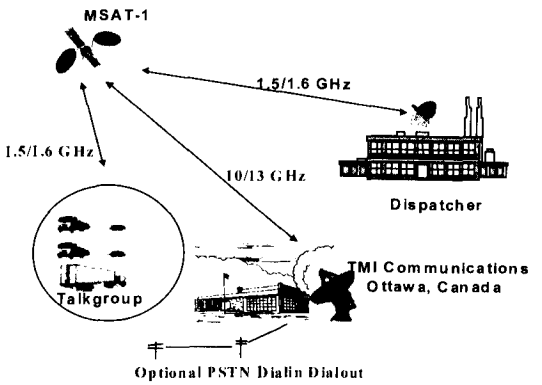


Figure 2-Dispatch Radio Network

After the originator of the conference finishes speaking and releases the PTT button, the other members of the conference are advised by audio and visual alerts that the net is "Vacant" or "Available". If no one keys their PTT within a preset time, called the hangtimer, the conference is torn down and the Communicators return to the signaling channel.

Information on active talkgroups is transmitted on the signaling channel every 10 seconds. If another talkgroup conference is initiated while a Communicator is engaged in the first conference the Communicator will move to the second conference within 10 seconds after the first conference concludes provided the second conference is still active.

There is a single signaling channel for all circuit-switched calls and therefore end users subscribing to services other than dispatch radio, such as phone data and fax, are in effect monitoring for all potential incoming call types to which they subscribe. Parties already engaged in a telephone call or another dispatch radio conference are on a communications channel and are not monitoring the signaling channel and can not be advised of the conference being established.

Talkgroups

MSAT Communicators can be equipped with 15 talkgroups as well as a 16th talkgroup that provides the private mode one-on-one service. A talkgroup is simply a definition of which Communicators participate in a conference when a user PTTs.

Priority 1 Interrupt

Unlike other TMI Communications circuit-switched services which are all full-duplex, dispatch radio is half-duplex so that the speaker's audio is not directed back to the speaker. Priority 1 Interrupt is a feature that allows another member of an active talkgroup to become the speaker after pressing a key(s) on his handset.

DISPATCH RADIO FEATURES

Dispatch Radio satisfies a wide range of end user requirements. Some organization's requirements are satisfied with one talkgroup operating in an area covered by a single beam. Other organizations require some or all of the following advanced features.

Monitor Codes-Priority Settings for Talkgroups

Monitor codes allow 1 of 3 priority codes to be attached to each talkgroup installed on each Communicator. The priority level, and other circumstances, determine whether the Communicator responds to the establishment of a conference.

If a Communicator has a low priority monitor code associated with a talkgroup that is activated the Communicator will only respond if that talkgroup is currently selected. Communicators respond to medium priority talkgroups, regardless of which talkgroup has been selected, except in the case where an end user has locked on to a talkgroup. Communicators always respond to high priority talkgroups even if an enduser has locked on to another talkgroup.

It is important to emphasize however that if a Communicator is already engaged in a conference, or another circuit-switched call type (phone, fax, data), it will not respond to even a high priority dispatch radio conference because it is on a communications channel and unable to receive information on the establishment of other calls.

MSAT Communicator Dispatch

Figure 2 above portrays an MSAT Communicator established at a dispatch centre. The advantage of MSAT Communicator dispatch is that it is completely independent of any networks such as the PSTN which are subject to damage or overloading during natural disasters. Another advantage is outlined under the section entitled "Y2K compliance".

Dial in and Dial out Dispatch

Figure 2 indicates the optional dial in and dial out dispatch options available via an interface at TMI with the PSTN. Dial in dispatch allows a dispatcher to dispatch using a PSTN line instead of an MSAT Communicator at the dispatch site. To establish a conference the PSTN caller would dial the phone number associated with a specific talkgroup and then enter a PIN (for security). The conference is established and the PSTN caller's voice is heard by all members of the talkgroup.

Dial out dispatch operates in a similar manner. A single dial out PSTN phone number can be associated with each talkgroup. A member of a talkgroup presses several keys to activate dial out dispatch after which the PSTN phone number rings. A typical application of dial out dispatch would be to have a cellphone or pager number as the dial out dispatch number which would be used to contact a dispatcher away from the central dispatch site during quiet hours such as evenings and weekends.

Net Bar

The net bar feature allows a Communicator user to disable dispatch radio (also known as net radio) to allow a Communicator to remain on the signaling channel for an important phone call and not be pulled off onto a communications channel by a dispatch radio conference.

MANUFACTURER IMPLEMENTATIONS OF DISPATCH RADIO

Mitsubishi and Westinghouse manufacture Communicators for TMI Communications and both manufacturers support dispatch radio services although with different implementations.

Mitsubishi implemented dispatch radio by developing a PTT microphone specifically for MSAT dispatch radio. The PTT mike has function-specific keys for activating priority 1 interrupt, locking on to a talkgroup and barring net radio calls. Single key strokes allow the end user to make either the phone handset or the PTT mike the active interface.

Westinghouse offers a number of different dispatch radio implementations. The original Westinghouse telephone handset can be used for dispatch radio without the requirement for additional hardware. The "1" or "3" key is used to PTT (establish a conference). Westinghouse also offer the option of a PTT button on the side of the telephone handset or the use of a separate Shure PTT microphone.

ACTIVATION/REVISION OF DISPATCH RADIO

TMI Communications introduced dispatch radio service approximately 2 years after MSAT phone service was introduced. The addition of dispatch radio service to existing customer Communicators or changes to dispatch radio service such as adding additional talkgroups is a very straightforward and effective process.

Adding or revising dispatch radio configurations is done over-the-air with a download to the Communicator from the network operations centre that takes approximately 20 seconds for the first talkgroup to be added and 10 seconds for each additional talkgroup.

FUTURE DISPATCH RADIO DEVELOPMENTS

Operating a network that is highly software oriented allows for ongoing development and the introduction of enhanced capabilities and features.

Full Service Dispatch

Full Service Dispatch (FSD), targeted for introduction in Q2 1999, is an interface that allows a subscribing organization to integrate MSAT Dispatch Radio into a dispatch console. Each FSD interface consists of an FSD voice communications interface and a Dispatch Data Communications (DDC) interface. At least one interface is required for each Talk Group that the subscriber wants their dispatch platform to communicate with simultaneously. The subscriber's dispatch platform can use the DDC protocol to dynamically change the Talk Group that an FSD interface is associated with and to receive information on activity within the Talk Group. The FSD interface is designed for leased line connections, though dial-up connections are also possible.

The FSD can send DDC messages to do the following:

- Set up a conference;
- Set up a private mode call; or
- Change the hangtimer for a talk group.

The FSD receives DDC messages containing the following real-time conference parameters :

- Start time of the conference;
- Unique number to identify conference;
- Which channel unit pool is being used;
- Talk group identifier;
- Hangtime;
- Whether the conference was initiated by a Priority 1 interrupt;
- Speaker ID of initiator;
- Number of L-band beams included in the conference;
- The current speaker ID (provided each time the speaker changes); and
- The date and time of the end of each conference

Broadcast Mode

Although it has not been implemented at this time, there is the capability within the dispatch radio service to offer a broadcast mode. This would be a talkgroup defined for broadcast information only. Typical examples include a broadcast talkgroup for road/weather information or marine forecasts and notices to mariners. The end user would select or "tune" to this talkgroup at certain scheduled times to acquire the report of interest which could be a portion of a taped broadcast for weather information for all of Canada for example. The cost of the airtime would be paid for by the organization providing the information. Access to such a broadcast talkgroup could be on a cost recovery "subscription basis".

DISPATCH RADIO ECONOMICS

Using 4 communications channel pairs (one pair per beam) MSAT-1 offers dispatch radio coverage throughout North America resulting in a significant economic advantage over terrestrial networks that would require repeater sites and frequency pairs several orders of magnitude greater for comparable coverage.

Organizations have the opportunity to select and pay for the coverage needed based on everything from single beam coverage to 6-beam coverage suitable for operations throughout North America, Central America, the Caribbean and northern portions of South America.

APPLICATIONS AND BENEFITS OF MSAT DISPATCH RADIO SERVICE

During the first 12 months of commercial service various categories of MSAT Dispatch Radio Service users have realized very quickly the benefits of Dispatch Radio service.

Provincial/Municipal User

The first MSAT-1 Dispatch Radio customer signed up initially for phone service with the proviso that dispatch radio service would be available within 6 months. Although the customer already had their own terrestrial trunked radio network in place the economics of upgrading the existing terrestrial network and ongoing maintenance versus opting for MSAT dispatch radio resulted in a decision favourable to implementation of MSAT.

Dispatch radio was easily downloaded over-the-air to the existing Communicators equipped with MSAT phone service. Customer vehicles include ambulances, snowplows/graders, maintenance vehicles and provincial police vehicles. The flexibility in defining talkgroups allows the customer to segregate individual activities such as policing on a dedicated talkgroup but also defining a single talkgroup that embraces all units during local emergencies or joint operations.

Marine User

MSAT Dispatch Radio was evaluated for use as a daily audio conferencing capability instead of continuing with a cellular telephone conferencing capability. Approximately 30 vessels were involved in the daily conference requiring the establishment of 30 long duration cellular calls and the cost for use of a conference bridge. The conference could be carried using MSAT dispatch radio using a single call in a single beam at a significantly reduced cost.

Commercial Users

Some of the commercial MSAT dispatch radio customers include logging truck operations, long haul trucking companies, a produce company and a helicopter fleet.

Y2K COMPLIANCE

Interest in Y2K (Year 2000) issues has generated considerable interest in the use of the MSAT-1 Dispatch Radio as a critical element in contingency plans for Y2K telecommunications. TMI's network offers an emergency backup solution for organizations that require reliable communications in the event of a terrestrial network failure due to the Year 2000 problem. TMI has completed a Y2K Readiness program for Dispatch Radio.

TMI's Y2K readiness program activities consisted of the following:

- Establishment of a clear definition of the Year 2000 Readiness
- Establishment of categories for all computer and computer related systems
- Created an inventory of all possible systems which may be affected by the Year 2000
- Performed an assessment of each system as to it's Year 2000 Readiness (in-house and vendor supplied)
- Prepared a schedule for correction of all in-house systems not Year 2000 Ready
- Upgraded in-house systems for Year 2000 Readiness
- Had vendors upgrade vendor-supplied systems with Year 2000 Ready versions
- Extensive testing of all systems
- Contingency plans for all essential TMI infrastructure
- Completion of Year 2000 readiness

November 1,1999 to March 1, 2000 has been identified as the critical time for access if a terrestrial network failure occurred. TMI is responding to a requirement for guaranteed availability by offering a channel reservation service for those 4 months. Note that a lead time of up to six months may be required to order the necessary equipment.

SUMMARY

MSAT Dispatch Radio provides advanced cost-effective wide area 2-way dispatch radio services. Flexibility is offered in regard to coverage required, number of talkgroups and the option of having a network completely independent of terrestrial networks. Advanced features include priority 1 interrupt, monitor codes and the option of dial in and dial out interconnection with the PSTN. MSAT Dispatch Radio has been successfully tested for a range of Y2K scenarios. MSAT provides wide area coverage of all of North America, Mexico, the Caribbean, Central America and the northern portion of South America. The end user pricing of dispatch radio is significantly better than comparable wide area terrestrial alternatives.

ACKNOWLEDGMENT

The author would like to acknowledge the review of and comments on this paper received from colleagues at TMI Communications; Donna Lyson, Manager of Applications and Customer Engineering and Jeff Cotroneo, Dispatch Radio Product Manager.

REFERENCES

John W. Jones, MSAT Broadcast Voice Services; in Proceedings of IMSC '95 pp. 401-403

MSAT Data and Image Transmission Trials for Airborne Scientific Applications

J.E. Jordan
Institute for Aerospace Research
National Research Council
1500 Montreal Road
Ottawa, Ontario
Canada K1A 0R6
Jim.Jordan@nrc.ca

ABSTRACT

Flight trials of data and image transmission using an aeronautical MSAT terminal were carried out using the NRC Convair 580 aircraft during 1998. The system was used in two major atmospheric experiments as well as in a Synthetic Aperture Radar (SAR) imaging trial. The aeronautical terminal was interfaced to a personal computer, on board the aircraft, which initiated data transmission calls to a variety of data services including an experiment dial-up server located in our laboratory with a dedicated telephone line and modem. Both terminal and TCP/IP PPP (Internet protocol) software were used depending upon the type of service required. Internet access was used extensively to up-link both data and images for the atmospheric experiments and to down-link information for the radar trials. This included information from/to an ftp server running on the dial-up access server as well as from remote Web sites located on the Internet providing up-to-date meteorological information. Performance of the system in the transmission of data and images is described.

INTRODUCTION

Background

The Flight Research Laboratory of the NRC's Institute for Aerospace Research operates a Convair 580 aircraft, which is shown in Figure 1. This aircraft is used in a number of airborne research projects including atmospheric studies, SAR radar development, geophysical mapping and navigation trials. Many of these projects have involved operation of the aircraft in remote locations such as offshore and in the Arctic. In these circumstances, up-to-date weather information is a clear requirement for safe operation of the aircraft, but often communications beyond the line of sight of established airports and aviation facilities is difficult or not always possible. As well, during the in-situ measurement of the atmosphere in meteorological studies, the availability of up-to-date meteorological information and imagery either from adjacent sites or using remote sensors is beneficial to the experiment.

Figure 1: NRC Convair 580 Aircraft

To meet this requirement, our group has explored a number of options. Initially, an APT weather satellite imaging system was developed for the aircraft to directly receive broadcasts of imagery from polar-orbiting weather satellites such as the NOAA and Meteor series, passing over the vicinity of the aircraft. This is particularly useful at high latitudes due to the frequent passes of these satellites and the relative lack of established facilities in the polar regions. The alternative to direct reception of satellite imagery is to relay it from a ground receiving facility over a suitable communications system.

With the advent of the MSAT mobile satellite system (MSS) and associated aeronautical terminals providing voice, data and fax transmission, the possibility of reliable wide-area communications to an aircraft at an affordable cost was opened up. Initial tests of data communications were undertaken in the fall of 1997 using a ground terminal to confirm feasibility and to investigate Internet access using TCP/IP protocols (at the initial 2400 baud data-rate). The decision was made to procure an aeronautical terminal, which was installed on the Convair 580 in late 1997. The system was operational in the first week of 1998, coinciding with the inauguration of the 4800 baud circuit-switched data service. This satcom system was then used in the Canadian Freezing Drizzle Experiment, which had been in progress since December of 1997.

Purpose

The purpose of this paper is to describe the operation and performance of the MSAT satellite communications system installed on the NRC Convair 580 in trials conducted during three scientific research experiments flown during 1998.

Scope

The first section of the paper includes a description of the hardware, software and networking aspects of the system. Subsequent sections highlight experiences during each of three experiments: the Third Canadian Freezing Drizzle Experiment (CFDE-III), the First ISCCP (International Satellite Cloud Climatology Project) Regional Experiment III / Arctic Clouds Experiment (FIRE-III/ACE) and the MUST-98 Radar Imaging Trial. The next section gives details of the performance of the system including measurements of raw data and image transmission speed and progress towards more efficient image transmission using compression techniques. A final section of the paper concludes with an overview of our experiences.

DESCRIPTION OF THE SYSTEM

Hardware

The satellite communications system used in this experiment was an aeronautical satellite phone, which operates using the MSAT geo-synchronous mobile communications satellite serving North and Central America. The terminal unit used was a CALQuest CQ-100, which provides voice, data and fax capabilities using a mechanically oriented antenna unit mounted on the top of the fuselage. This terminal operates over a digital voice channel from the aircraft to the satellite and down to an earth station, which is connected to the conventional terrestrial public switched telephone network (PSTN), allowing calls to be dialed to (or from) any public telephone number. The data service is circuit switched, operating at 4800 baud over the same channel, with a Hayes modem-compatible interface. An industrial grade rack-mounted 486-66 personal computer with a 1024 x 786 x 256 color display was interfaced to the transmitter/receiver unit using the standard serial COM interface port. A block diagram of the system is shown in Figure 2.

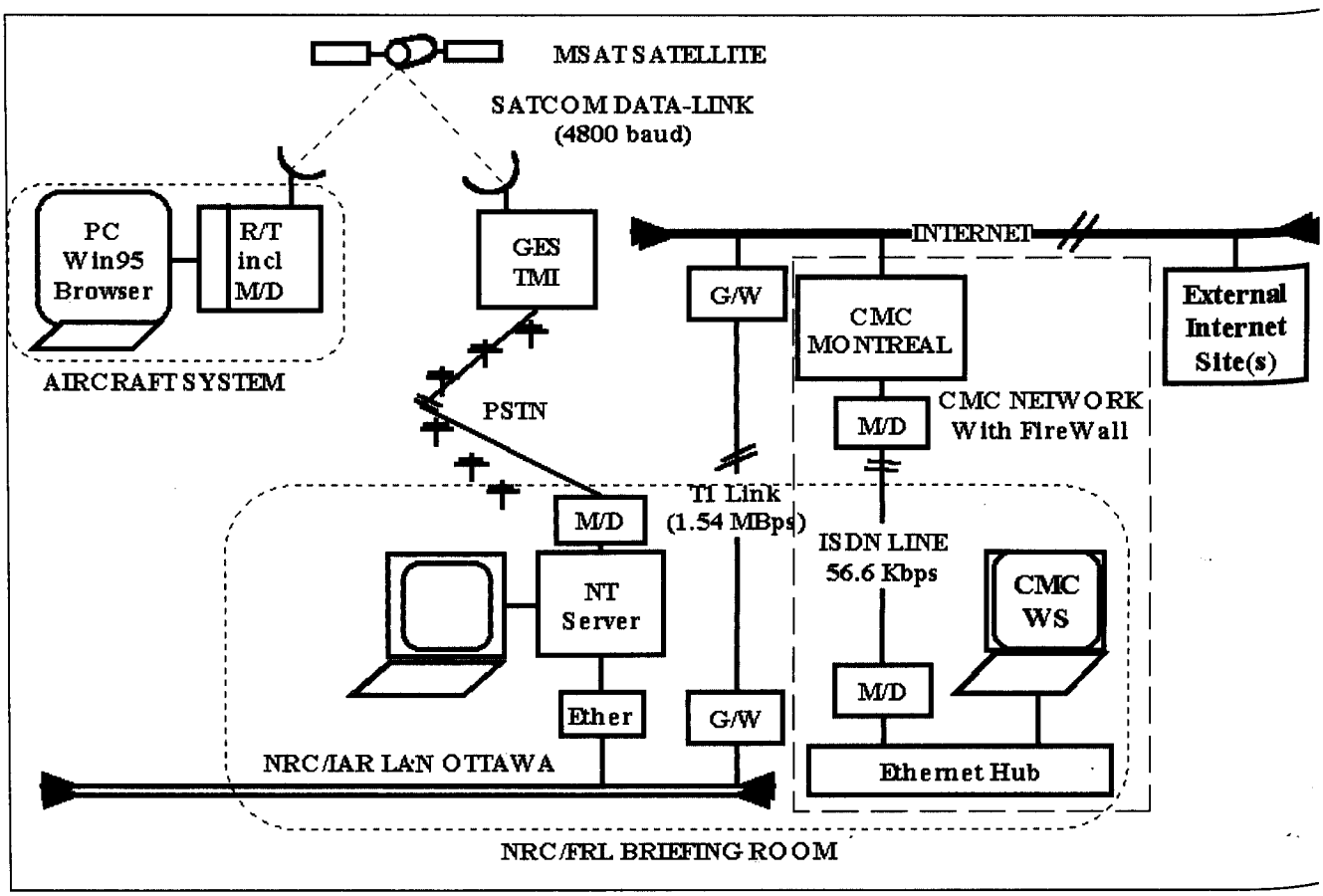


Figure 2: Block Diagram of CFDE-III System

Software

For these experiments, the satcom data-link was configured using telephone modem software as might be used on a standard computer-to-computer telephone link. This included bulletin-board-compatible software and Internet-compatible software using TCP/IP-PPP (Transport control protocol, Inter-networking protocol, Point-to-point protocol). For DUATS (Dial-up User Acess Terminal Service) aviation weather information, vendor-supplied software using an undocumented modem protocol was used to dial up and connect to a Canadian-based service provider called QTABS. For the rest of the work, operating system vendor supplied TCP/IP PPP stacks were used on the aircraft computer running Win95 and the ground-based Internet gateway/remote server running WinNT Server/RAS using standard dial-up networking techniques. Both standard ftp (file transfer protocol) and http (hypertext transfer protocol) Web browser application software were used on the aircraft for accessing information on desired hosts on the Internet, including an ftp server dedicated to experiment usage located on the dial-up server.

Networking

In the case of Internet access using TCP/IP protocols, the aircraft computer was effectively connected to the Internet, through the NRC dial-up NT server, which was configured differently for each experiment. The data-link to the ground-based computer connected to the Internet was established by dialing up the dedicated telephone number of the server modem port, establishing a link through the satellite, through the earth station and through the PSTN to this host. This is illustrated in the baseline network diagram of Figure 2. In the case of CFDE-III and MUST-98, the NT server was connected to our building LAN (local area network), which was connected to the Internet through a gateway, a T-1 line (1.544 Mbps) and router.

SCIENTIFIC EXPERIMENTS

CFDE-III

The first of the three experiments, conducted during 1998, was the Canadian Freezing Drizzle Experiment. CFDE-III was the third experiment in a series involving the measurement and characterization of freezing drizzle during winter conditions. This experiment was based in Ottawa, involving flights in Ontario and Quebec over the Great Lakes and St. Lawrence River valley region from December 1997 to late February 1998, in contrast to previous experiments in St. John's Newfoundland and Ottawa, Ontario. The data-link itself wasn't installed until the first week of January 1998. The objective of the experiment was to measure atmospheric properties associate with freezing drizzle to provide a better understanding of its micro-physical and dynamic properties, which can be used to improve forecasting models. Freezing drizzle involves precipitation droplets between 30 and 400 microns, in contrast to larger freezing rain or smaller cloud condensation particles, and has been implicated in recent aircraft incidents.

In the case of CFDE-III, an experiment weather forecasting and briefing room was established in our laboratory with meteorological workstations for accessing weather information from the Canadian Meteorological Center (CMC) weather models and assimilation data/observations from the Atmospheric Environment Service (AES) national data network. These workstations were connected to CMC using a 56.6 Kbps ISDN line and were located within the network firewall. The dial-up server, in contrast, was connected to the NRC network with full access to the Internet. Meteorological data or images could be provided for upload to the aircraft by manually transferring required files from the AES workstations to the dedicated local ftp server. The aircraft computer could access public information available from any site on the Internet without restriction using these protocols.

The satcom system performed quite well during this first operational trial during an airborne research experiment. Connections were solid with no interruptions experienced during almost all of the flights over this two month period. Occasional slowdowns were evident during browsing of large images from the Internet, which may be attributable to link errors and subsequent retransmissions under the TCP/IP protocol. In particular, the system was used during the major ice storm, which hit Eastern Canada during the first week of January. No failures were evident with the satellite, earth station or PSTN connections though some problems were experienced with network connections at NRC in Ottawa and CMC in Montreal, which were likely a due to electrical power failures. Fortunately, our laboratory only experienced short power outages and we were able to circumvent network problems by using alternate Internet Service Providers during these situations. Voice mode was used occasionally to provide guidance on research flights as well as to contact aircraft maintenance personnel regarding in flight concerns.

FIRE-III

The second experiment conducted in 1998 was the First ISCCP (International Satellite Cloud Climatology Project) Regional Experiment III / Arctic Clouds Experiment (FIRE-III/ACE). This experiment was based in Inuvik, NWT during the month of April 1998, with flights over the Beaufort Sea North of the Northwest Territories and Alaska. Highlights of the experiment included four flights over the SHEBA (Surface Heat Budget of the Arctic Ocean Experiment) site, operating concurrently with FIRE-III, from the CCG Des Groseilliers icebreaker frozen into the ice approximately 500 miles Northwest of Barrow, Alaska near 76N, 165.5W, as well as many flights North of Inuvik over the Beaufort Sea.

During the FIRE-III experiment, a weather forecasting and operations briefing room was established in Inuvik, NWT. In this case, both the meteorological workstations and the dial-up NT server were connected to a common Ethernet hub and a temporary 56.6 Kbps ISDN telephone link to the closest network access point in Yellowknife. The Ethernet Hub replaces the NRC LAN as shown in Figure 2. This allowed aircraft access to sites within the AES domain as well as on the Internet. Personal computers were connected separately through a modem to the local Internet Service Provider operating over a cable television network.

MUST-98

The MUST-98 (Multi-SAR) Trial was conducted off the east coast of Nova Scotia during November 1998 using a SAR radar system being developed and tested by the Canadian military on the Convair 580 aircraft. In this case, SAR radar images were transmitted to the ground over the satcom system as a trial of the system. Images were transferred both to the ftp server running on the dial-up NT server located at the FRL laboratory, as well as to a temporary experiment ftp server located in Halifax to test the ability of the system to deliver imagery in near-real time to a field-based facility. As well, text files were transferred up to the aircraft to confirm the ability to transfer information in the reverse direction. The voice system was also used to coordinate operations with research personnel located on the ground.

PERFORMANCE OF THE SYSTEM

Data Transmission

One objective of these trials was to characterize the data-link performance, particularly for TCP/IP protocol use. It is well known that TCP/IP protocols can experience reduced performance over geo-stationary satellite data-links [1,2]. One aspect of this is related to window buffer sizing in relation to the delay-bandwidth product of the link. Also, the long path delay affects the speed with which the protocol can initiate transfers. This also exacerbates the problem of retransmission of packets due to link errors, which are interpreted by the protocol as traffic congestion leading to avoidance procedures that reduce data transfer speeds.

It is estimated that the delay-bandwidth product for this data-link is in the order of 720 Byte-seconds. This is based on a round-trip time (RTT) of approximately 1500 milliseconds, measured using Ping, and a raw data-rate of 480 Bytes/second for a 4800 baud serial link, configured with 1 start, 1 stop, 8 data and no parity bits or 10 bits per data byte transferred. Since the receive-window buffer greatly exceeds the data in transit over the propagation period, no slowing of the link would be expected due to buffer size. Link measurements confirm this.

The characteristics of the TCP/IP data-link are listed in Table 1, based on default Windows 95 and NT network

Table 1: TCP/IP Parameters	
Link Data-Rate	4800 baud
Link Configuration	1 stop, 8 data, no parity
Max. Transfer Unit (MTU)	1500 Bytes
Receive Window	8192 Bytes
Round Trip Time (RTT)	~ 1500 msec

values confirmed by system values reported. The TCP and IP headers account for ~ 40 Bytes, the PPP headers for 10 Bytes and the PPP octet insertion loss is estimated at 198 Bytes per frame, when certain single bytes are replaced by two bytes, assuming randomly distributed bytes. It is calculated that there is an overhead of 20% for the serial bit framing (8N1) and an additional 13.3% overhead due to the PPP/IP/TCP protocols for this assumption. This reduces the raw serial port data-rate of 480 Bytes/second to 400.68 Bytes/second, which is in agreement with the effective transfer rate observed during long ftp transfers of binary files, without TCP/IP compression. Figure 3 shows measurements of the transfer times in seconds for data files in Kilobytes. The straight line plotted represents ~400 Bytes/second, which corresponds to the diamond-shaped data points

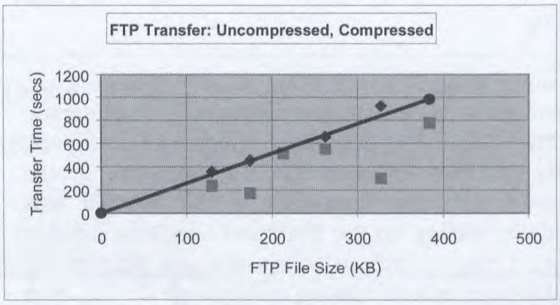


Figure 3: FTP Transfer Times

(◇ - No Compression, ■ - Compression)

representing ftp file transfers using no TCP/IP compression. The square data points below the line represent ftp file transfer times using TCP/IP compression where the data-rate is effectively increased. Web/http file transfers are shown in Figure 4 with TCP/IP compression enabled. This would indicate that ftp transfers are faster

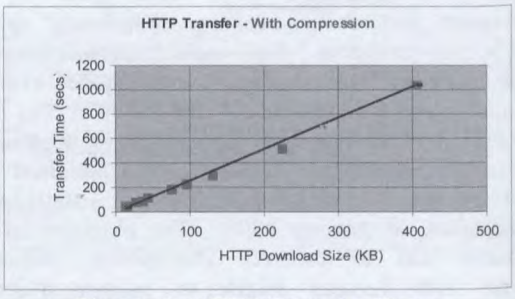


Figure 4: HTTP Transfer Times

(■ - Compression)

than http transfers, when TCP/IP compression is used, as might be expected due to the additional protocol overhead.

Windows TCP/IP uses compression of data packets and headers by default. Improved data transmission was observed in many cases, particularly for ascii file transfers but also to a lesser extent for binary files. It is known that TCP/IP compression is based on the LZW compression algorithm; however, efforts to predict link speed-up using a batch file compression utility (Zip) were unsuccessful.

No quantitative measurements of link errors and associated retransmissions or link start-up times were made. Occasional slowdowns in the link were noted particularly for large http transfers In these situations, the link transfer would stall for several seconds before continuing with rest of the transfer. It was assumed that this was due to a link error, possibly due to the tail of the aircraft shading the antenna on the reciprocal azimuth heading of the satellite.

Image Transmission

Figures 3 and 4 describe the performance of the link for the transmission of image files and browses as well as data. Images of from 1/3 to 1/2 MB in size were transferred in approximately 10-15 minutes, which was considered the practical limit for image transmission for operational use. For transmission of larger images, it is possible to use lossy image compression techniques. This is commonly done for JPEG images in Web pages, which are encoded using the discrete cosine transform (DCT) to reduce the size of the file, transmitted over the network and decoded at the client site using the inverse DCT transform prior to display. The resulting image appears to the user to be almost identical to the original image, relying on the ability of the human visual system to "fill in" missing details. The limit on the size reduction, as expressed by the compression ratio (CR), depends upon the image, but is generally limited to ratios of 10:1 or less since larger ratios introduce a blocking effect.

Recently, other lossy compression methods have been used including wavelet and fractal compression. These algorithms operate using similar principles but have been shown to provide improved performance at higher levels of compression. In particular, blocking effects have been found to be less objectionable. Wavelet compression was tested using two image compression implementations: 1) the SPIHT algorithm [3] and 2) a commercial image compressor and browser plug-in "demo" based on the HARC-C technology (available from Compression Engines [4]). In some cases, a reduction of image size was provided by cropping the original image down to a smaller "region of interest", corresponding to the area of operation for that day's flight. Image compression was then used to further reduce the file size. An example of one of these tests is shown in Table 2. A modest improvement of about 2:1 is shown over JPEG for both procedures. Larger compression ratios were found to introduce smoothing of the rich texture in this image, which was considered an important feature for their interpretation. This is somewhat contrary to current research results using simple images without much texture. The test example image shown in Figure 4 is from the infrared channel of the GOES meteorological geo-stationary satellite, showing the Pacific Ocean on April 22/98 at 2200 UTC. Figure 4(a) shows the original image and Figure 4(b) shows the image after standard JPEG compression. Figure 4(c) shows the image after wavelet compression at an effective CR of 5.7:1 and Figure 4(d) shows the same image after wavelet compression at an effective CR of 11.5:1 (both using the CE package). These compression ratios are actually deflated by a factor of 3 since the viewing software converts the resulting images from 8 bits to 24 bits per pixel (bpp), despite their mainly 8 bit gray scale content. In particular, note the smoothing effect on the rich texture of the image is shown in Figure 4(d), [within the limits imposed by image printing technology].

Table 2: GOES IR 04/22/88 2200Z	**Size**
Original Image (gray scale)	900 KB
Image in GIF Format	402 KB
Image in JPEG Format	141 KB
Image Compressed – SPIHT Codetree (8 bpp)	78 KB
Image Compressed – Comp. Engines (24 bpp)	71 KB

CONCLUSIONS

The use of an MSAT aeronautical terminal in scientific airborne research is illustrated in this paper. Access to the Internet was found to be a useful aspect of this trial, allowing up-to-date data and images to be linked to researchers in near real time. Despite the limitation imposed by the 4800 baud data-rate, the system was found to be useful for some applications.

ACKNOWLEDGEMENTS

The author acknowledges the collaboration and funding provided by the Atmospheric Environment Service (AES) and the Department of National Defence (DND).

REFERENCES

[1] M. Allman, C. Hayes, H.Kruse and S. Ostermann, TCP Performance over Satellite Links, Proc. 5th International Conf. On Telecommunications Systems, March, 1997.

[2] H.Kruse, M.Allman, J.Griner and D.Tran, HTTP Page Transfer Rates Over Geo-Stationary Satellite Links, Proc. 6th International Conf. On Telecommunications Systems, Mar. 1998.

[3] A. Said and W.A. Pearlman, A New Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees, IEEE Transactions on Circuits and Systems for Video Technology, vol. 6, pp. 243-250, June 1996.

[4] http://www.cengines.com

(a) Original GIF - 402 KB

(b) JPEG - 141 KB

(c) Wavelet Compression - 71 KB : eff. CR = 5.7:1

(d) Wavelet Compression - 35 KB : eff. CR = 11.5:1

Figure 4: GOES IR Weather Satellite Image – Apr 22/98 2200 UTC

Satellites and Transportation: Emergency Management

E. Sterling Kinkler, Jr.
Southwest Research Institute
P.O. Box 28510, San Antonio, TX, USA 78228
SKinkler@swri.org

ABSTRACT

Billions of dollars are being funneled into deployments of Intelligent Transportation Systems (ITS) in many parts of the world. Many different configurations with widely varying goals are taking shape, but in one way or another, they are all data acquisition and control systems, albeit widely distributed over entire cities or regions. They are also highly communications intensive developments with each using hundreds of miles of copper or fiber systems and some dedicated microwave channels. There is, now, much emphasis on migrating the benefits of these technologies to the vast areas of most countries that are not metroplexes.

The upcoming global Internet Service Provider (ISP) networks, which are envisioned by many to be the foundation for commercial consumer oriented satellite services, can also reflect an enabling communications infrastructure that makes ITS in rural areas feasible. Routine data collection and video surveillance requirements for vast stretches of interstate highway systems may arise in the future. Certainly, focused bandwidth-on-demand applications for communications links that are mobile and deployable at crash scenes and construction sites, for example, will be used. The former example reflects the huge data communications demand that these systems represent. The latter illustrates the potential for temporary, mobile, emergency response links that can save lives, property, and dollars.

This paper is about ITS and selected relevant projects. It is about efforts to position both the provider and user communities to understand the plans and needs of one another and to search for common ground and influence plans so that the need for services and the need for markets overlap.

INTRODUCTION

Modern communications are enabling advances in education, health, transportation, and most other socioeconomic systems in the United States and other countries. We have seen many examples of the enhancements to life experiences that these technologies can make in areas of dense population where market forces support deployment of wired and mostly terrestrial wireless connectivity programs. In the near future, satellite-based communications will enable the provision of services that we now find in major cities to significant population sectors that live in small towns and villages in rural America and beyond. These technologies will provide almost unimaginable leaps in access to infrastructures and all that this implies for many less-developed countries all over the world.

Wireless telephony, terrestrial and focused in high-density market areas today but soon to be ubiquitous in its availability, has shown an attraction to "staying in touch" that is reflected in this exploding market among consumers, businesses, and government agencies. Also around the corner are satellite-based ISP systems that will be capable of providing higher capacity computer connections between earth-bound consumer and business terminals at almost any place on the face of the earth. Nomadic tribes in grass-roofed huts in Africa will literally be "wired" and able to surf the internet and email their friends and relatives on the other side of the world to say hello, or trade stocks on their choice of world exchanges on a moment's impulse.

We will all feel the impacts of these new services in the daily conduct of our fun and work. The utility of these connections will spur new markets, generate new technologies, and make the world a much smaller place to live. There are benefits to look forward to that most of us cannot imagine and may not recognize, even when it is happening. For instance, some major cities with major traffic problems are testing programs with many technologies to monitor freeways for accidents in order to respond more quickly, thereby saving lives and reducing resultant congestion and delays. The almost ubiquitous presence of the cell phone in most of these cities has outrun the most sophisticated traffic monitoring technologies!

By the time our sensors and cameras have detected a problem, motorists with cell phones have reported the incident to authorities minutes ago. Unfortunately, most motorists are unable to describe the location on a freeway to authorities, and this presents a problem that has spurred a legislated solution. Soon cellular telephone service providers will be able to tell a 911 operator the location where an emergency call was initiated. In major

metropolitan areas, however, investigations have shown that we can detect motor vehicle accidents AND their approximate location by simply making intelligent use of cell phone statistics. Detecting the explosion of cell phone calls that accompany an accident and which cell towers (locations) were involved in the avalanche of calls can often provide needed information.

ITS is communications. Routine data collection, control output measures, and video surveillance are required. Implementing these functions without ready access to the bandwidth required has not been feasible. Implementing them with wireless connections requires many channels of ISP-type capacity or more and is still not feasible with terrestrial wireless systems. If even a fraction of the miles of rural interstate highways in the USA were "connected" as such, one can imagine the significance of this data communications demand on the marketing plans for such services. In addition, mobile, rapid deployment of these types of connections to emergency scenes is also needed.

Without these advances, our crumbling and over-stressed infrastructures will continue to deteriorate. With these advances, we look forward to saving time, dollars, property, and lives.

TRANSPORTATION AND COMMUNICATIONS?

In the early 1990s, the US administration began focusing attention on an outdated and over-extended ground transportation infrastructure. The Intelligent Vehicle Highway System (IVHS) initiative was founded, and that initiative has grown into the current Intelligent Transportation System (ITS) industry. Last year, the US Congress passed the largest spending bill ever, the "Transportation Equity Act for the 21st Century" or TEA 21 bill. This act provides over $198 billion over six years to the transportation community with much of it targeted toward deployment of ITS technologies around the country. The bill also provides special emphasis to migration of the benefits of ITS technologies, which have been demonstrated in many metropolitan areas, to the rural areas of America, hence, the term "Equity" in the title of the bill.

Basically, the aim of the ITS movement is to better manage our ground transportation infrastructure to increase its capacity through the use of technology. The costs of right-of-way and new construction have spiraled. The already over-stressed freeways and interstates in the country will become even more congested in the future. The number of vehicle-miles driven in the US is expected to double over the next 25 years. ITS, then, is about using technology to better manage these transportation systems to add capacity in ways other than building more miles of wider freeways.

ITS developments and deployments deal with much more than roads and traffic sensors, however. Public transportation, private automobiles, commercial trucks, and government vehicles all use our roadways. Trains, subways, airlines, ships, and all modes of transportation and cargo shipping intersect with the highway system, some more than others. ITS has become a forum for better management of the operations of each of these infrastructures and the intersection of or *intermodal* aspects of transportation. It is easy to see that the economic vitality of any country is highly dependent on the capacity and reliability of its ground transportation infrastructure.

Why have we focused on cities? Indeed, the most ominous sign of under-capacity transportation systems is the terrible traffic congestion that we experience in any major city today. Traffic delays due to congestion total up to billions of dollars per year in lost productivity, wasted fuel, unnecessary pollution, and more. Several things cause congestion. Too many vehicles on a road designed for less traffic will surely cause slowdowns. Accidents that block lanes can create parking lots lasting for an hour or more. Curious onlookers slowing down in opposing lanes to see what is happening at an accident site can jam up even the other side of a freeway system. Construction and maintenance operations on roadways are also big contributors to congestion. Often, accidents or stalled vehicles happen *because* of congestion, and the effects of this combination are exponential.

There is an even better reason why we have focused on cities so far in ITS. *ITS is communications!* Modern traffic management systems use dedicated buried fiber optic systems that are distributed about metropolitan areas following freeways and major roads. These fibers carry tremendous amounts of sensor data and video signals from roadside cameras into a central facility that uses this data to detect incidents and congestion and provide output to drivers through signs and directions. Drivers are alerted in advance of trouble ahead and can switch to clear lanes or even alternative routes. Many such systems use leased telephone lines, microwave, coaxial cables, and even twisted pair copper wires to collect and distribute such information between the roadways and the Advanced Traffic Management System (ATMS). Local (and even distant for travelers) traffic conditions are broadcast on radio, television, internet, and any other means of distribution that can reach drivers in time to alter routes and plans. In San Antonio, in-vehicle route guidance computers can receive this information over commercial radio broadcasts to include real-time traffic condition information in the algorithms that compute the least-time route from a current location to a desired destination.

RURAL ITS

Even smaller cities and towns can become congested; however, the ITS emphasis in areas other than congested metropolitan areas is on efficiency and safety. Efficiently, and therefore thoroughly, interacting with commercial traffic can yield huge gains. Streamlining the enforcement and tracking of permits, hazardous materials marking,

weight limits, and other regulatory requirements can mitigate the huge impacts absorbed by truckers traveling from state to state as they line up at long cues for inspection near state lines. Border crossings between the US and Canada and Mexico are snowballing in numbers due to passage of NAFTA agreements, inundating border crossing facilities and forcing automation, which depends on communications. Toll collections are occurring at 50 miles per hour or more when both the tollbooth and the vehicle are suitably equipped to communicate. Again, billions of dollars in lost time, productivity, and wasted fuel are at stake.

EMERGENCY MANAGEMENT.

Roadways carry vehicles. Rural roads, interstate highways, and freeways carry vehicles at higher speeds and sometimes also at very high traffic densities. Motor vehicle crashes can and do happen - often. Freeways and interstate highways also intersect with the public in many other ways. People cross them in vehicles, on bicycles, on foot, and even on horses and mules. Domestic and wild animals such as cows, deer, and moose love to inspect these ribbons of asphalt that make walking so different. Roads cross roads. Trains cross roads. Natural disasters change road usage patterns such as the sudden use of all freeway lanes to evacuate low-lying areas in advance of a serious hurricane. Sometimes we do this routinely to add capacity where needed such as in High Occupancy Vehicle (HOV) lanes that run toward a city in the morning and away from a city in the evening. Natural and manmade disasters often damage or change roadways in uncontrolled events such as earthquakes, train derailments, and 18-wheelers falling off of elevated roadways onto traffic below. The list goes on and on.

Almost all transportation agencies and organizations include traveler safety at the very top of their list of priorities and mission statements. This includes management of the design and signing of roadways to provide easy and straightforward operation, avoiding confusion and potential hazards. Most do an excellent job of this; however, we continue to suffer staggering statistics in the loss of human lives and the cost of property loss and lost productivity. Recent reports by the National Highway and Traffic Safety Administration (NHTSA) put these figures at over 42,000 people killed each year in the United States alone. Also, in the US, over 5.2 million people are injured in motor vehicle crashes (MVC) per year, and the economic impact of the medical care, lost time and productivity, and property loss from these MVCs total over $150 billion dollars per year. And remember, the number of vehicle-miles driven in the US (and therefore the exposure to MVCs) will double in the next 25 years. We must do all that we can to avoid these incidents. Much attention is being focused on this, especially in future on-board systems for vehicles and interactive systems that communicate between the roadway and the vehicle. We also try to mitigate the effects of MVCs with better crashworthy designs and such innovations as seatbelts and airbags. The reality is, however, that all these things, like the best road design and signing practices, will only make a dent in the overwhelming statistics generated by transportation-related incidents.

It is also worthy to note that only a small fraction of these MVCs occur in rural areas, with the crowded city streets contributing far more to the total. More than 60% of fatality MVCs reported, however, occur on rural interstates or roads. Persons with serious medical emergencies in the rural regions of the US are 4-to-7 times (sometimes even greater) more likely to perish from their injuries or illnesses than their counterparts in metropolitan areas. This imbalance is primarily due to longer response and transport times and a generally poorer level of readiness and training among rural volunteer fire and EMS systems. This is where much of ITS and TEA 21's focus on rural America is targeted: the safety of the population that lives or travels through these regions. The efficiency discussed above is important. There is much to improve; efficiency contributes to safety, but rapid detection and appropriate response to traffic related incidents is where the statistics focus our attention.

ONE APPROACH

The detection and rapid response to major accidents or other emergency incidents where life, limb, and property are placed in jeopardy is a huge problem in most of the world. The problems are compounded in rural areas where detection and response can be agonizingly slow. Terrestrial wireless systems such as the cellular telephone infrastructure have helped a great deal in notification of authorities when an emergency incident occurs, but only where such coverage is available, and notification delays are only a small part of the problem! As we noted above, today's cellular coverage is generally found near cities and towns and in interconnecting ribbons about major highways and interstates. It is not found in the vast majority of the land areas of any country, where significant sectors of the population live, work, and play.

It seems that the way we deliver information and services to persons outside major metropolitan areas is about to undergo a revolution. Nowhere are these new capabilities needed more than in the management of serious emergency incidents and disasters which affect the lives and safety of so many in our population.

ATMS facilities in our major cities are being set up as regional emergency management centers. Police, fire, traffic management, EMS, and many other agencies involved with incident response are coming together within these facilities to integrate and coordinate with each other to provide rapid, coordinated responses to emergencies. Truck incidents often require a hazardous material response in order to safeguard the public from danger or loss. Damaged vehicles on the roadway often lead to secondary collisions, even before a first responder has arrived and marked the problem with flashing lights.

Construction sites often create reductions in roadway capacity, create traffic slowdowns, and precipitate MVCs. Train derailments can damage highway structures and create hazardous material emergencies, making associated highways dangerous for vehicular traffic. Floods and earthquakes can damage highways, making them unsafe for use or at least limiting their capacity. Again, the list goes on and on.

Rapid detection of these incidents by monitoring of road data and surveillance, as currently happens in many cities, may never really reach the wide-open spaces. It is feasible, however, to implement such systems, along with detection automation, along major interstates and highways that carry most of the traffic and produce most of the incidents - with appropriate communications. On-board systems are capable of detecting a collision and placing a call to alert authorities and even provide GPS-derived location and crash severity information. Such systems are being tested in limited deployments now. But they only work if sufficient communications bandwidth is readily available to them at the time of need. With enabling communications, emergency management personnel can be notified and provided with location, types, severity, scope, and nature of incidents through many different infrastructure and on-board or even on-person implementations. They can dispatch the right equipment from the closest location, monitor the scene for changing requirements, advise personnel on the scene who may lack information or training, and just as importantly, not dispatch or recall un-needed equipment and resources to keep them available for other emergencies.

INTERMEDIATE STEPS

A program is unfolding in San Antonio, Texas that is aimed at addressing many of these issues and positioning the transportation and emergency management communities to make use of near future satellite-based systems and services. The LifeLink™ project, which was conceived and developed at Southwest Research Institute (SwRI) and ultimately funded by the Texas Department of Transportation aided by federal ITS funding, is now up and running. Ten ambulances in the city have been equipped with computers and displays, video cameras, and wireless radios to connect them with the distributed communications systems that are part of the city's TransGuide ATMS. From there, the signals are routed to connect the ambulances with doctors in the emergency departments of the city's level one trauma centers. The result is a TCP/IP-based 2-way live video-conference capability including a vital data telemetry system that allows doctors, paramedics, and patients to view each other and participate in all that is happening in the world of "Telemedicine" while at emergency scenes and during transport to a hospital.

An effort is underway now, again primarily funded by transportation programs, to develop, integrate, and test a satellite-based extension of this type of system to be used not only for medical emergency services but for all aspects of coordinated emergency incident management as discussed herein.

THE TERRESTRIAL WIRELESS SYSTEM

The LifeLink system is a distributed mobile LAN designed to link ambulances on or near San Antonio's freeway system with trauma care providers in the city.

Three major axioms were provided by the LifeLink team of consulting emergency and trauma physicians during the development of the system. These axioms provided design goals that are reflected in the implementation of the system. The axioms are:

- Provide sufficient performance for emergency triage and assessment.
- Do not increase scene time nor distract the paramedic from his job.
- Bring the conference to the doctor, not the doctor to the conference.

Development of the LifeLink system pursuant to these three general goals led the SwRI team to confront several formidable technical issues. The technical solutions implemented in this system reflect a number of firsts in telemedicine:

1. A high capacity hybrid terrestrial/wireless mobile communications technique that provides the required connectivity for telemedicine applications between moving vehicles and hospitals.
2. An engineered solution adapting telemedicine to the emergency medical environment. Automation of the call placement and management, largely transparent to the users was a requirement. Automatic management of communications, computer systems, power systems, and other technologies are built into the LifeLink system.
3. Networking of video and other required data to provide important options in medical control of emergency situations. The intuitive and easy-to-use capability to switch medical control among appropriate departments or physicians online and/or to share data from the field for consultation purposes, again online and immediate, reflect a revolutionary step in telemedicine applications. This capability was also judged significantly important to the use of telemedicine in the emergency environment.

The link utilizes the facilities and roadside fiber-optic network of the TransGuide ATMS, as shown in Figure 1. A video teleconference is initiated at the discretion of the ambulance crew. Once established, the link can be handed off from hospital node to hospital node, but only one hospital node can be in communication with an ambulance at a time. Consulting hospital nodes may view the video and listen to the conversation but are not able to participate

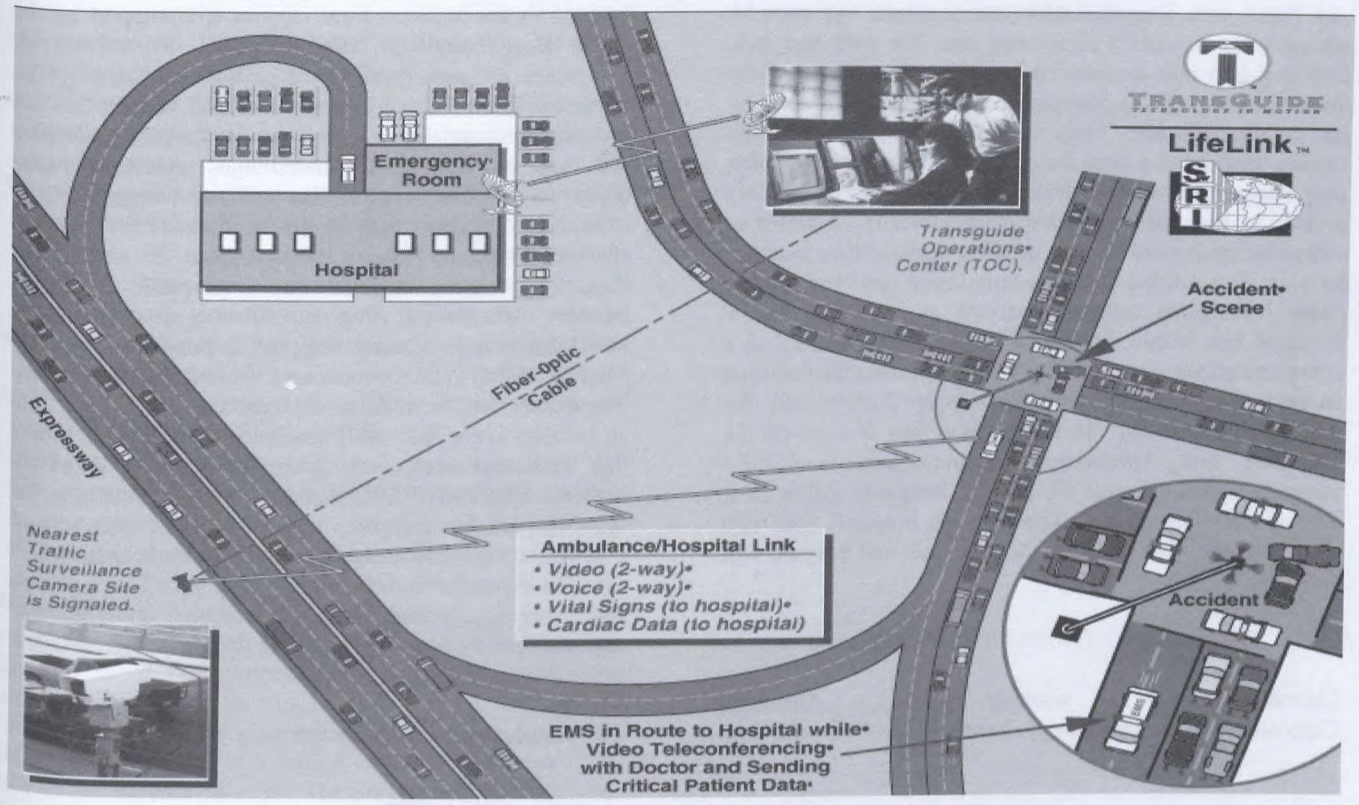


Figure 1. Graphic Illustration of LifeLink Terrestrial/Wireless Communications System

in the conference except through normal telephone contacts between hospital nodes. The LifeLink system supports multiple nodes within a hospital and nodes distributed among multiple hospitals.

Each ambulance carries a computer that is configured for a LAN-based video teleconferencing application. The view at each hospital node of the teleconference is a full screen view of the video sent by the ambulance node with a small image in the corner of the screen containing the local image that is being sent. A status bar is placed at the bottom of the screen indicating the current ambulance unit number, the current hospital node and telephone number, and other support information.

The digital video communications system at the heart of the LifeLink system provides a very important feature in a wireless environment. Occasional transient losses of communications are inevitable and are due to noise, interference, multipath interference, loss of line of sight, and combinations of these phenomena. The LifeLink system handles these issues by controlling video performance to take maximum advantage of the available digital throughput at any time. Reduced throughput is reflected by reduced frame rates in the displayed video, and most communications breaks go unnoticed. When communication is lost completely, such as when operating out of range, etc., then the last good frame of video is displayed as a still picture, providing the best available image until communication is restored.

The system uses a wireless link between an ambulance and a nearby TransGuide camera location. The TransGuide cameras are typically located at intervals of one to one-and-a-half miles apart along the city's freeway system. An unlicensed 2.5 GHz spread spectrum radio is located in each ambulance and on each camera pole. The line-of-sight radio link operates to connect the ambulance to the nearest

suitable camera location and to seek a new connection when the existing connection begins to fade, thus providing continuous interconnect as the ambulance moves along the freeway system.

The half-duplex digital radio using Ethernet-like protocols is connected to a full duplex converter within each associated TransGuide fiber hub. A long-haul fiber transceiver then interconnects the duplex converter to each of the Tx and Rx fibers, which terminate within the TransGuide building. These fibers are currently used in the TransGuide system, and wave-division multiplexer techniques are used to operate the LifeLink system on the existing fiber network but at a different light wavelength. This technique basically uses some of the unused bandwidth available in the TransGuide fiber backbone.

Within the TransGuide building, a switched hub is located and fitted with long-haul fiber transceivers. A network management computer is located near the switched hub. The switched hub operates to interconnect any one of the hospital nodes with an ambulance in the field as directed by the ambulance crew. Only one ambulance can use any camera location at a time; however, multiple hospital nodes may interconnect with multiple ambulances simultaneously as long as the ambulances are geographically separated or otherwise positioned so that separate camera locations can be used. Typically, a LifeLink ambulance can "see" two or three TransGuide camera locations at any time. The switched hub within the TransGuide facility connects to a communications system capable of supporting the required connection(s) between the TransGuide facility and the respective hospital(s). In some cases, this is a leased T1 telephone line. Ultimately, all communications links between TransGuide and the member hospitals will be over TransGuide fiber to the nearest (to the hospital) fiber hub and dedicated fiber between the fiber hub and hospital will complete the link.

EXTENDING THE SCOPE AND REACH

Currently, SwRI is working with the Advanced Communications Technology Satellite (ACTS) office at NASA's Glenn Research Center to extend the reach of such systems to the remotest rural regions with support by the Texas Department of Transportation. Several private companies are also contributing partners to the program, primarily from the satellite terminal and satellite service provider communities. The program is designed to develop and test an integrated, autonomous, practical, rapid-deployable mobile terminal that will use communications capacity to be provided in the near future by planned satellite systems. Further, the program is targeted at integrating such a terminal into a practical emergency incident management video surveillance, data, and audio link integrated in a rapid response Incident Management Module (IMM) to be mounted on a vehicle or a small trailer. The system will function to implement a remote link from an incident scene into the TransGuide facility, and thereby link into, and interoperate with, the LifeLink terrestrial wireless mobile TCP/IP network. While maintaining the link provided for emergency medical services when serious injuries are involved, the resulting network deployment could be expanded to include many emergency management and response authorities, yielding unprecedented opportunities for integrated, coordinated emergency incident response capabilities.

The program is currently progressing in the first two of six

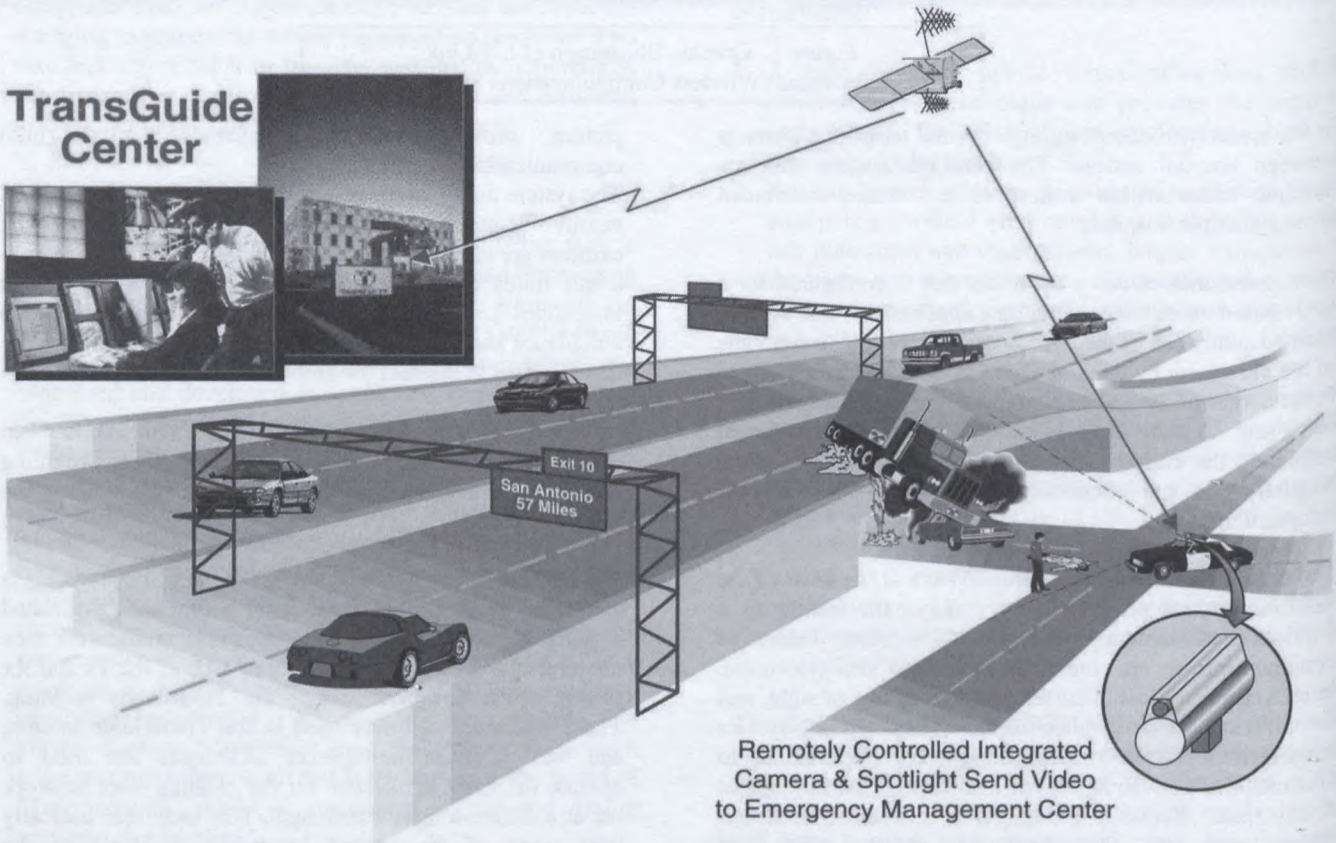


Figure 2. LifeLink™ "First Responder" Application
Mobile Wireless or Future Satellite Communications

phases:

1. Integrate and test a suitable satellite link into the LifeLink configuration at static locations.
2. Integrate and test as in phase 1 but include mobile wireless (terrestrial) components moving about in the configuration.
3. Develop and implement a prototype commercial terminal for ultimate use in emergencies from fixed rural sites linked to the metropolitan LifeLink system for remote incident surveillance, management, and coordinated response.
4. Migrate the terminal design to include maintaining a link during transport.
5. Merge the terrestrial and satellite based system to provide seamless inter-operability.
6. Migrate the developments and lessons learned to commercial terminal providers and commercial satellite service providers to make the links generally available to the transportation and emergency services community.

Recent interest by the commercial airlines and maritime industries may result in additional phases or sub-phases in the near future. Preparations for the testing of the phase 1 and 2 configurations are underway as of this writing. A mid-phase 3 milestone involving testing of an integrated autonomous terminal at a staged incident is scheduled for late in the summer of 1999. The program as now planned will span a 2-to-3 year timeframe, ending with the fielding of operational systems.

In the foreseeable future, the ISP satellite systems will provide near ubiquitous global coverage. These mobile terminals may ultimately be small and inexpensive enough to be deployed on every sheriff's car or other first responder vehicle as shown in Figure 2. Response and dispatch decisions will be made on a coordinated basis. More experience and better training and equipment will be available at the scene, some by remote link. Secondary incidents will be minimized, hazards to the general public will be dealt with swiftly, delays will be avoided, resources not needed will be available for other emergencies, and lives and dollars will be saved.

This program represents an outreach by the transportation and emergency services communities to highlight the significant needs and requirements that are faced in bringing the advantages of ITS initiatives to rural locations. Communications infrastructure is a requirement for this migration. The ubiquitous satellite-based ISP market envisioned by most commercial Ka-and Ku-band LEO and GEO (and hybrid) systems may play a very important role in reaching these areas with connectivity that reflects *enabling technology*. The author calls upon satellite service providers, terminal equipment providers, and all others working toward these implementations to reflect on these issues and join this program team in pursuit of such needed solutions. Our universal goal is to research the common ground that already exists between plans and needs and to influence designs and business plans, if need be, to accommodate and provide required services and interfaces for this significant and important market as early as possible.

Dichroic Aeronautical Antenna System For DBS Video Reception

Peter C. Strickland
EMS Technologies Canada Ltd.
1725 Woodward Drive, Ottawa, Ontario, K2C 0P9, Canada
Email: strickland@calcorp.com

ABSTRACT

EMS Technologies Canada Ltd. has developed aeronautical antenna and transceiver products for services such as Aero-H, Aero-I, MSAT and most recently Ku-band direct broadcast video. The DTA (Direct To Aircraft) antenna is a two-axis steered dichroic reflector with a diameter of 11.5" that fits within the tail-fin cap on business jets and larger aircraft. The antenna system includes a precision beam steering controller, low noise amplifiers, a two-channel down-converter and radome. A system G/T of 9dB/K is achieved on the ground, and a higher G/T is achieved at elevation.

INTRODUCTION

Direct broadcast television services such as Hughes Direct TV are typically designed to operate with earth station apertures of 18"-36" in diameter. It would be impractical to put an antenna of even 18" in diameter on an aircraft, either in the tail or within a top-mount radome due to drag, cost and safety considerations. It is consequently necessary to use a smaller antenna with exceptionally low noise temperature in order to achieve the required system G/T. In this case a reflector diameter of 11.5" is used, and the low noise amplifiers are integrated directly into a compact scalar feed horn. In order to minimize losses the probes, which excite the feed waveguide, are printed directly onto the LNA printed circuit board.

In addition to the DTA antenna the tail fin may also house other antennas such as the EMS pitch over roll steered AERO-H antenna and a VOR antenna. In order to ensure that the Ku-band antenna does not interfere with the radiation patterns of these other elements a dichroic reflector design has been developed. The dichroic reflector is transparent at most frequencies that are not harmonics of the DTA band. At frequencies in L-band and below the reflector has negligible transmission loss.

Radome loss in the DTA system is of considerable importance due to the need for maximizing the gain to noise temperature ratio. Radome loss is mostly reflective in nature however some reflected components can be directed towards the earth or engines resulting in increased noise temperature. In addition the reflected components reduce gain, and refraction and diffraction effects result in beam pointing losses. In order to maximize G/T a line of foam cored radomes has been developed having thin Quartz/polyester skins. The quartz fibre results in a rigid radome able to withstand bird strikes while maintaining low loss at Ku-band.

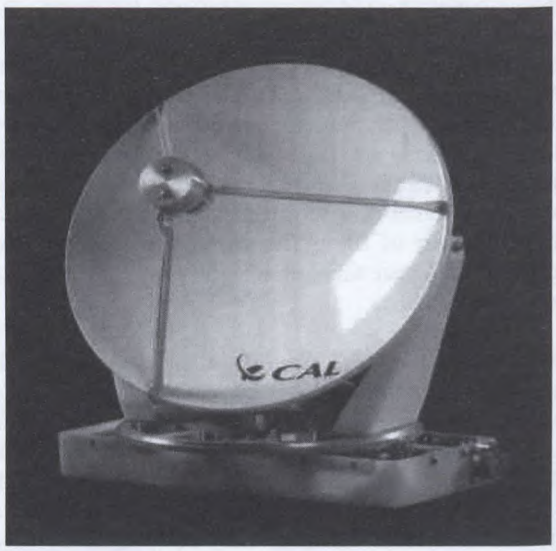


Figure 1 DTA Antenna System

ANTENNA PROFILE

The DTA antenna profile has been selected to minimize feed blockage and to ensure that the feed does not increase the swept volume of the antenna. In order to minimize the blockage introduced by the feed, the feed must be as small as possible and this corresponds to a relatively wide feed beamwidth. In this case the beamwidth is 112 degrees at the –10dB points. Analysis of the antenna using a physical optics code indicates that a –10dB feed taper to the edge of the reflector provides a near optimal trade-off between spill-over loss and illumination efficiency, providing the highest gain achievable using a simple compact feed. The corresponding F/D ratio is 0.44 and this is short enough to ensure that the end of the feed does not extend beyond the volume swept by the reflector.

The reflector profile is shown in Figure 2 along with the feed position and a representation of the feed blockage.

A physical optics analysis of the reflector structure yields the directivity pattern shown in Figure 3. In this analysis the surface of the reflector is treated as a continuous perfect conductor although the actual reflector surface is dichroic. Over the operating bandwidth of the DTA systems for which this antenna is to be used the dichroic surface does have a reflection coefficient magnitude very close to unity and the simple solid-reflector model applies quite well.

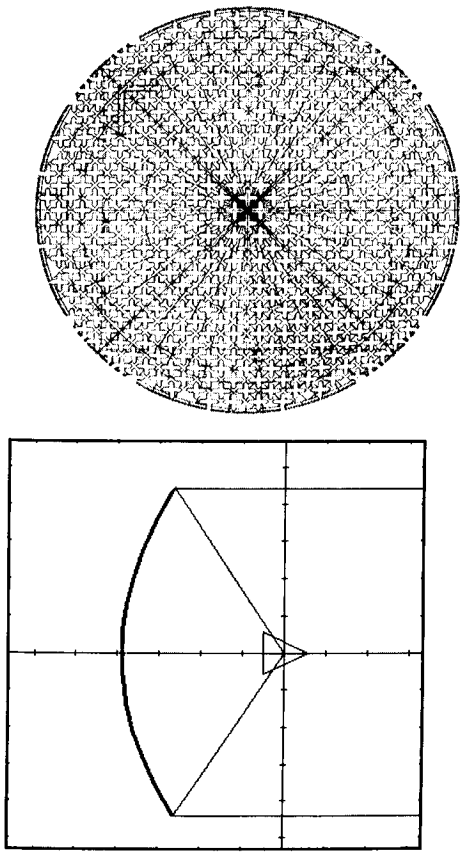


Figure 2 Reflector Profile

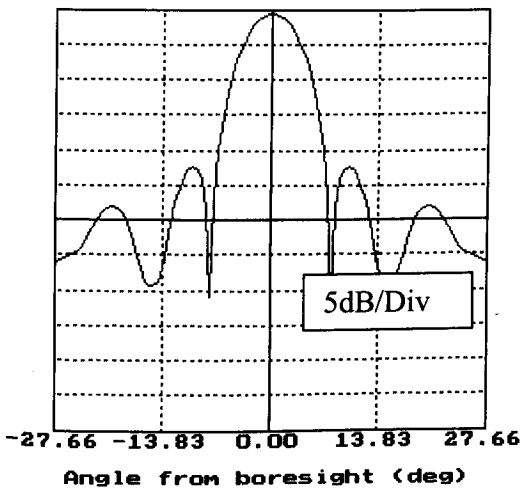


Figure 3 Secondary Pattern

DICHROIC SURFACE

A wide variety of dichroic structures have been described in the technical journals over the years, and these can generally be categorized as band-pass or band-reject designs although one could conceive of low-pass and high-pass configurations as well. In the case of the DTA antenna the goal is to have a reflector which is transparent at L-band and reflective at Ku band. At L-band the structure must be transparent at all reflector orientations relative to the L-band source while at Ku-band it must have a near unity reflection coefficient magnitude for incidence angles of +/-30 degrees from the reflector axis. Also, the reflector must be insensitive to incident wave polarization in either band.

The band-pass reflector configuration is typically implemented as a solid reflector with slots cut into the surface which are resonant at the desired band-pass frequency; L-band in this case. In order to pass all polarization components well, crossed slots or Y shaped slots may be used. The principal advantage of the band-pass configuration is that the surface reflection characteristic is very broadband. Unfortunately however the band-pass characteristic is quite narrow and the centre frequency of the band-pass is very sensitive to the incident angle making it unsuitable for use in the DTA application.

The band-reject reflector configuration is usually implemented with metallic resonators printed onto the reflector surface. A variety of resonator designs are possible as illustrated in Figure 4 below and each results in a unique band-reject characteristic and requires considerable optimization for broadband performance. In this case a commercial Moment Method analysis code (IE3D) has been used to optimize the surface in the presence of the composite dish. The composite dish has two fibre-reinforced skins with an inner core of low loss dielectric foam and the IE3D code includes each of these dielectric layers in its analysis. It is trivial to design a band-reject surface with a high reflection coefficient at a particular incidence angle and over a narrow range of frequencies, however it is quite difficult to obtain a near unity reflection coefficient magnitude over a 17% bandwidth and +/-30 degree range of incidence angles. In general the frequency of maximum reflection coefficient is quite sensitive to incidence angle and optimization is required to minimize this sensitivity. It has been found through extensive analysis that the interlaced cross structure can provide the highest minimum reflection coefficient over the frequency band and range of incidence angles required in the DTA design. The surface typically has a transmission loss of over 20dB within the operating band and this corresponds to a reflection coefficient magnitude of 0.995. The typical Ku-band response of the selected surface is plotted in Figure 5 below. In L-band the surface has a measured transmission loss of less than 0.2dB.

Truncation of the dichroic surface at the edge of the reflector is a concern since a truncation of the resonators shifts their resonant frequency and can result in a reduction of the radar cross section of these elements and a corresponding reduction in the overall aperture efficiency. In the case of the cross elements one side of the cross will continue to contribute to the cross section if the other is truncated making the cross structure much less sensitive to truncation than the Y structure. In the EMS design significant cross truncation has been avoided by wrapping the crosses around the reflector edge. In production this is achieved through the use of a three dimensional laser photolithography process. The composite dish surface is plated with a high purity ED copper, this is then covered with a photo-resist and the resist is selectively exposed with a multi-axis-scanning laser. The unexposed resist is removed along with the copper below producing the desired dichroic surface.

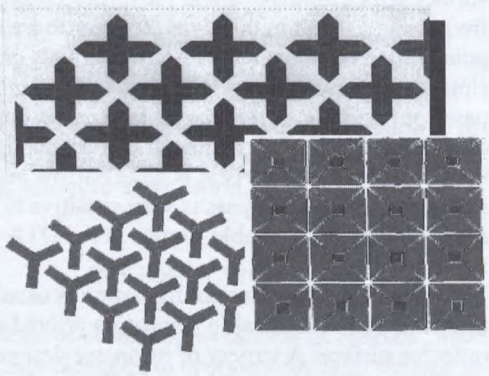


Figure 4 Dichroic Structures

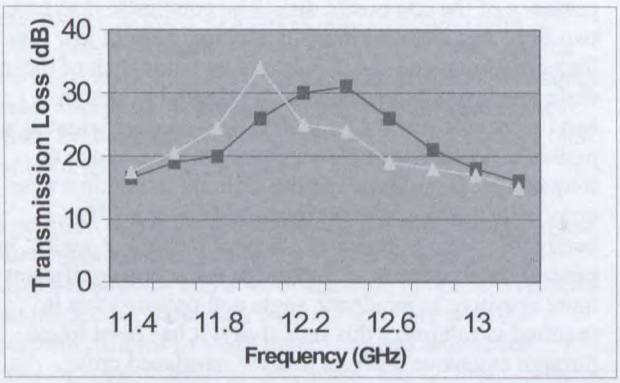


Figure 5 Transmission Loss
0deg and 45deg Incidence

FEED DESIGN

The maximum diameter of the reflector used in the DTA system is strictly limited to a maximum of 11.5" by aerodynamic and structural considerations with respect to the aircraft tail design. The broadcast services for which this antenna will typically be used are designed to work with earth stations having apertures of greater than 18" in diameter and LNB's with noise figures on the order of 1dB. The key then to receiving these signals using the smaller airborne antenna is to minimize the LNA noise temperature and the system losses. In order to achieve a system noise temperature lower than that of the typical terrestrial terminal a design was selected which integrates a PHEMT LNA directly into the feed with the launching probes being etched directly onto the PCB. A clam shell configuration was selected where the PCB containing the LNA and probes is clamped between the feed waveguide and back short.

Typically multiple corrugations are used around the feed aperture in order to equalize the E and H plane beamwidths however in this case a single optimized groove has been used with excellent results. Similarly the feed has been matched with a single circular waveguide transformer in order to minimize the waveguide length and thus the swept volume and losses. The feed radiation pattern is given in Figure 6 below and the feed mechanical structure is illustrated in Figure 7 along with the printed probes.

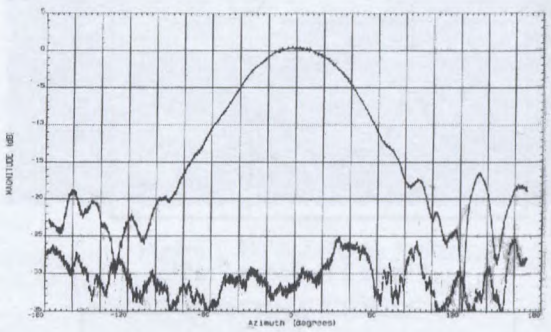


Figure 6 Feed Pattern
Co-polarized and Cross-Polarized

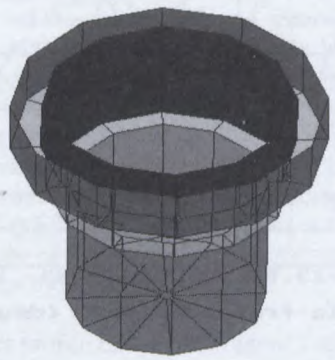


Figure 7a Feed Design

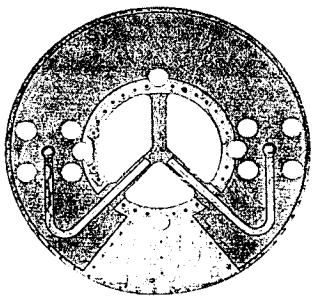


Figure 7b Feed Design
Printed Probes

G/T PERFORMANCE

LNA Noise Figure:	0.6dB
LNA Gain:	20dB
Feed Hot Loss:	0.1dB
Feed Mismatch Loss:	0.15dB
Radome Reflection Loss:	0.35dB Typ.
Radome Dissipative Loss:	0.05dB Typ.
Reflector (dichroic) hot loss	0.1dB
Dichroic Band Edge Loss (cold loss)	0.2dB
Brightness Temp: (Beam well above horizon, Includes allocation for sidelobes)	25K
Post LNA Noise Figure:	3dB max including cables and joint.

T_{REC} $= (F_{LNA} -1)290 + (F_{PLNA} - 1)290/G_{LNA}$
$=(1.148-1)290 + (1.995-1)2.9$
$=42.96 + 2.89$
$= 45.9K$

T_{ANT} $= (1-1/L_H)290 + T_{SKY}/L_H$
$= 39.8K$
T_{SYS} $= 85.7K$

Antenna Gain:	29.3dBi
Sum of Losses:	0.95dB

G/T = 9.0dB/K

It should be noted here that G/T is not a particularly good figure of merit for receiving systems since for a given power density the signal to noise ratio for a given G/T is dependent on the square of the antenna gain. A better figure of merit is the effective aperture to noise temperature ratio. For the DTA antenna we have:

$$A_e/T = \lambda^2 G/(4\pi T) = -34.4 \text{ dB-m}^2/\text{K}$$

MECHANICAL DESIGN

The antenna is steered by means of an ultra-light, electromagnetically transparent elevation over azimuth positioner. All support components above the base plate are fabricated as precision fibreglass composites, ensuring that the entire antenna system will have low transmission loss at frequencies below Ku-band. Some basic mechanical and electrical interface specifications are as follows:

Mechanical pointing accuracy: +/- 0.5 deg
Slew rate: 40 deg/sec either axis
Acceleration: 40 deg/sec/sec
Drive System: Belt drive in elevation, gear in az.
Motors: Sine wave driven stepper
Rotary Joint: Two channels
Slip Rings: 7, silver plated rings, fingers are Silver-graphite with Molybdenum disulfide
Weight: 8.8lbs (Antenna, with downconverter)
Interface: Arinc 429
ACU Separation: Up to 150'
Temperature Range: -60C to +70C Operational

REFERENCES

[1] P.C. Strickland, A Low Cost Electronically Steered Phased Array for General Aviation, Proceedings of the International Mobile Satellite Conference, 1990, Ottawa, pp. 169-171

[2] P.C. Strickland, Compact, Low Profile Antennas for MSAT, Mini-M and Std-M Land Mobile Satellite Communications, Proceedings of the International Mobile Satellite Conference, 1995, Ottawa, pp. 340-344

Experimental results with a circular electronically steered antenna for mobile satellite communications

M. Lecours, M. Pelletier, P. Lahaie, T. Breahna, Q. Wang, G.-Y. Delisle,
Dept. of Electrical and Computer Engineering, Université Laval, Québec, Que, Canada G1K 7P4
R. Daviault and M. Lefebvre,
Davicom Technologies, Inc., 2765 de l'Industrie, Trois-Rivières, Que., Canada G8Z 3X9
Email: mlecours@gel.ulaval.ca, rdaviault@davicom.com

ABSTRACT

This paper presents experimental and simulation results for a novel 24 beam electronically steered antenna design with a full circular coverage in the horizontal plane, implemented with low-cost microstrip technology. The antenna is designed to operate in the mobile satellite band from 1525 to 1559 and 1626.5 to 1660.5 MHz and to provide a fixed angular coverage from 0 to 70 degrees in elevation with a gain of 8 dBic at a 35 degree elevation.

The design uses 8 microstrip antennas disposed circularly, slanted from the horizontal plane at an appropriate distance from the center. These antenna elements are activated by groups through a radial switch. One-bit phase shifters are used to reduce the effect of the cross-over at the beam intersections, to compensate for the propagation phase shifts due to circular geometry and to slightly tilt the beam relative to the normal of each group of three, thus permitting to select 24 beam-pointing angles for the entire array.

INTRODUCTION

This paper presents experimental and simulation results for a novel electronically steered antenna design with a full azimuth coverage at medium and low elevation angles for application in MSAT or INMARSAT Mini-M. The antenna array is implemented in low-cost microstrip technology. The antenna beam is fixed in elevation and the beam is scanned in azimuth by switching groups of elements around the array.

The paper discusses the antenna physical configuration and the design choices, and describes the antenna's constitutive elements, namely the radiating microstrip patch elements and the phase shifters, while the switch is analyzed in a companion paper [1]. The paper then presents simulation results as well as our latest experimental measurements.

GENERAL DESCRIPTION

The mobile satellite band used for MSAT and INMARSAT services extends from 1525 to 1559 MHz and from 1626.5 to 1660.5 MHz for the reception and the transmission bands of the earth station. Both services require adequate right hand circular polarization over the covered area. For MSAT an antenna gain between 8 and 13 dBi with a VSWR better than 1.5:1 over the pass-band is specified. For INMARSAT Mini-M, an EIRP of 14 dB +/- 2 dB and a minimum G/T of - 17 dB/K° have to be met.

The novel antenna prototype, presented in this article, uses 8 microstrip patch antennas disposed circularly, slanted downward from the horizontal plane, at a given distance from the center (see figures 1 and 2). These antenna elements are activated in groups through a radial switch [1]. One-bit phase shifters [2] are inserted to reduce the effect of the cross-over at the beam intersections, to compensate for the propagation phase shifts due to circular geometry and to obtain a choice of beam orientations in azimuth for each group of antenna elements, thus permitting to select, in particular, 24 beam-pointing angles for the entire array.

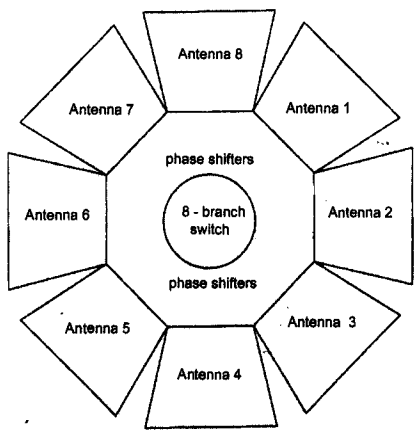


Figure 1. Schematic diagram of the antenna.

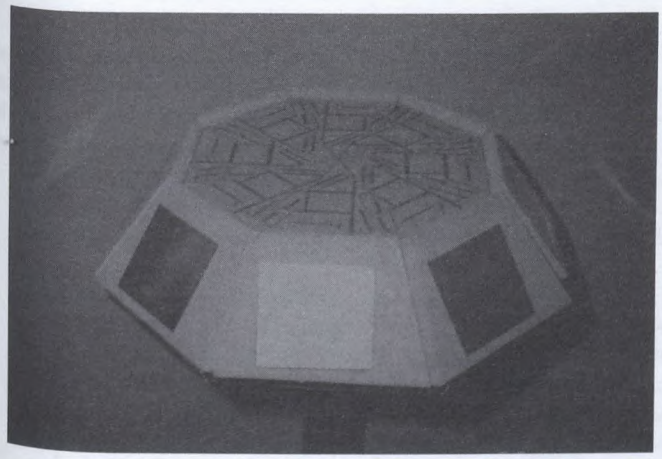


Figure 2. Photograph of antenna prototype.

The antenna, without the radome, is approximately 8.5 cm high and 36 cm wide and is implemented in microstrip technology using an economical substrate.

Surface mount plastic encapsulated PIN diodes are used for the switch and phase shifters, which limits the prototype realized to a transmit power in the order of 15 Watts. Modified designs capable of handling increased power levels are possible. Unlike its nearest market competitor, the mechanically steered antenna, there is no mechanical motion used to make the beam scan and beam scanning can be done almost instantly.

RADIATING ELEMENTS

The characteristics of the radiating elements are central in the design of an antenna array. The bandwidth of a standard microstrip antenna separated from the ground plane by a thin dielectric is inherently limited to some 3 % of its operating frequency, while the bandwidth required in this case corresponds to about 8.5%. One can achieve some broad-banding by using parasitic elements in the neighbourhood of the antenna patch. One particularly effective method is to use two stacked patches. In one popular configuration, one feeds the first patch, which is separated from the ground plane by a thin dielectric, with a second patch stacked over the first patch some distance away. The effective dimensions of each patch determine its resonance frequency and one can achieve a dual band antenna, if the resonant frequencies are far apart, or broad-banding if they are very close from each other. Ours is a somewhat intermediate case, where both patches interact with each other.

In the case presented here, the lower patch is printed on the substrate with a 3.05 dielectric constant, while the upper patch is a brass or aluminium thin plate, separated from the lower patch by a low dielectric constant convenient material. RHCP is achieved by feeding the antenna at two different appropriate points through a simple power dividing network. Figure 3 shows the reflection coefficient for the radiating element, as predicted by ENSEMBLE 4.0. Figure 4 shows the measured reflection coefficient.

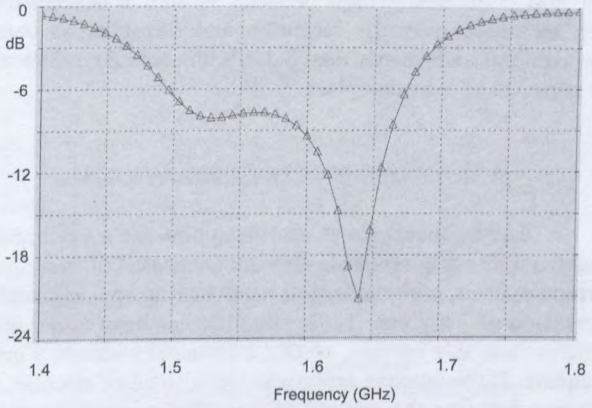


Figure 3. Simulation of the stacked patch antenna reflection coefficient over the band of interest

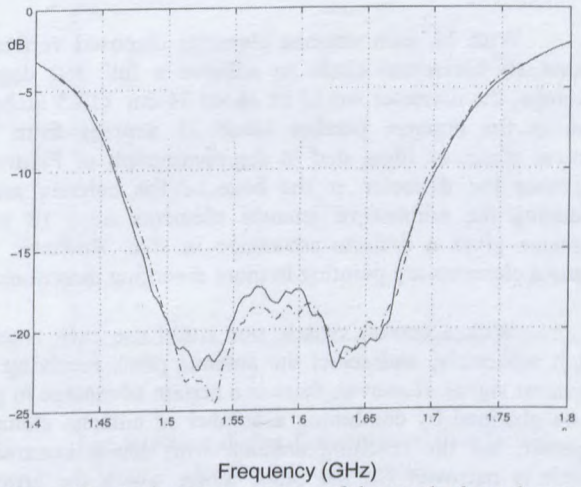


Figure 4. Typical measurements of the stacked patch antenna reflection coefficient over the band of interest.

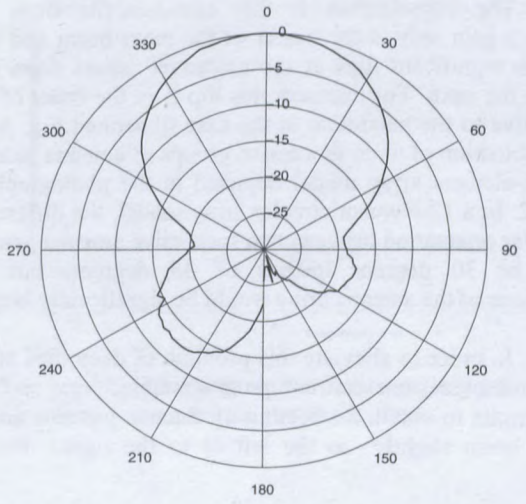


Figure 5. Typical measured radiation pattern in azimuth of a single antenna element over a ground plane.

Figure 5 shows a typical radiation pattern in azimuth of an antenna element. Gains varying between 7.1 and 9.2 dBic and -3 dB beamwidths between 55 and 70 degrees have been measured over the reception and transmission bands. The axial ratio was better than 5 dB, with some degradation in the presence of a ground plane.

ARRAY DESIGN CONSIDERATIONS

Let us then take as starting point for a preliminary design a microstrip radiating element composed of two square stacked patches, with the largest patch having approximately a dimension of 7 x 7 cm. Little coupling has been found when these are put side by side, so that a physical width of 9 cm is adequate. If the antenna array was flat instead of circular, the distance between the centers of the antenna elements would be also 9 cm, leading to a spacing of 0,45 λ between array elements.

With 12 such antenna elements disposed vertically around an horizontal circle to achieve a full 360 degree coverage, the diameter would be about 34 cm (13.5 inches); slanting the antenna patches some 35 degrees from the vertical plane, as illustrated in the photograph of Figure 2, increases the diameter at the base of the antenna array. Reducing the number of antenna elements from 12 to 8 elements gives a definite advantage in size. However, the antenna elements are pointing in more diverging directions.

With a central switch, one could use each antenna patch separately, and select the antenna patch receiving the strongest signal. However, there is a certain advantage in gain to be obtained by combining a number of antenna elements together, but the resulting antenna array has a beamwidth which is narrower (in the plane along which the array is located) than that of a single antenna patch.

The consequence is that one benefits from the maximum gain only at the center of the main beam and that there are significant dips at the crossover points from one beam to the next. For instance, this dip is in the order of 4.5 dB relative to the maximum in the case illustrated Fig. 6 for the combination of three successive groups of antenna patches in the 8-element array model depicted in the photograph of Figure 2. In a 12-element circular array model, the difference in angular orientation between two successive antenna patches would be 30 degrees instead of 45 degrees, but the dimensions of the antenna array would be significantly larger.

In order to alleviate this problem of deep dips at the beam crossovers, one can use programmable phase shifters. This permits to obtain a selection of antenna patterns and to tilt the beam slightly to the left or to the right. For an 8-element antenna array, a good choice is to select phase shifts that will tilt the beam orientation by +/- 15 degrees, thus giving an effective choice of 24 beams. Figure 7 shows a particular example where the dips at the crossovers have been reduced to less than 1 dB.

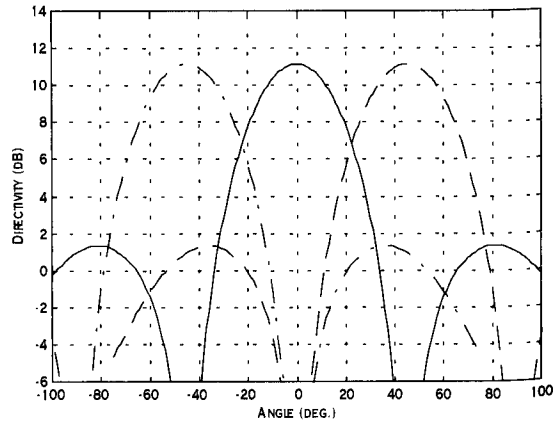


Fig. 6. Radiation diagrams for three groups of antenna patches in an 8-element circular array of the model of Fig. 2.

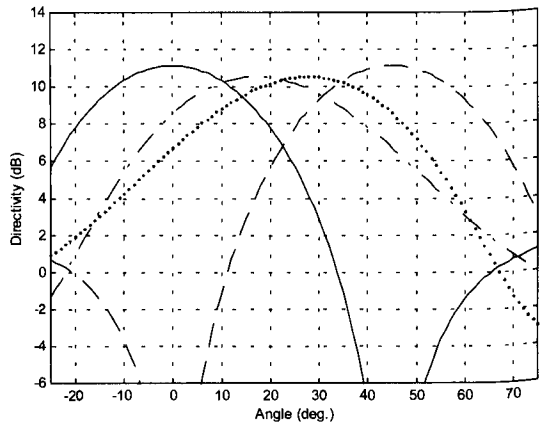


Fig. 7. Modifications to radiation diagrams and obtainment of 24 beams by insertion of a phase shifter.

There is a price to be paid, namely additional losses due to the phase shifters and some degradation in the impedance adaptation. Part of the advantage in directivity obtained in combining the patches is lost through the phase shifter losses. The main advantage of this approach is to provide a full 360 degrees coverage with only small dips at the elevation angles considered. One obtains a fan beam, with the radiation diagram in elevation giving a beamwidth similar to that of the single patch.

PHASE SHIFTERS

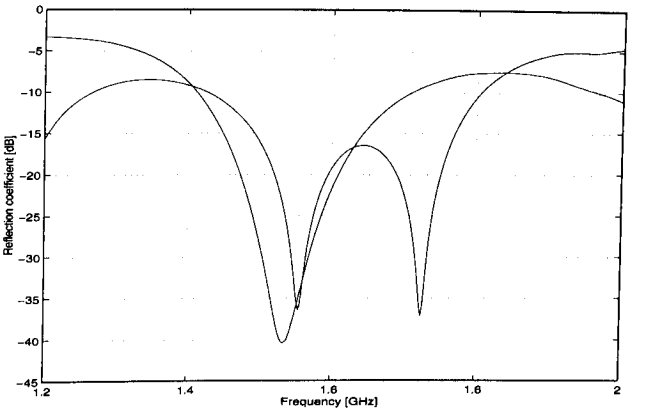
As mentioned above, antenna branches are combined and one-bit phase shifters are used in the antenna array. Each of the phase shifters can take one of two states, 0 or 1, which gives a number of possible configurations. Some of these configurations are particularly useful in order to obtain given beam orientations

For the antenna array prototype, we have made use of reflection phase shifters. This type of phase shifter is based on a 90 degrees hybrid circuit, in our case a microstrip line branch coupler. This device has four ports, two input/output ports and two so-called separating ports. The signals appearing at the separating ports are the input signal divided in half and 90 degree out of phase to each other.

If the same impedance is connected at the two separating ports, the reflected power will recombine and flow through the output port. When the impedances are perfectly matched, no power is reflected an no power flows through the output port. When the impedances are short circuits or open circuits, the whole signal is reflected and flows through the output port. In practice, depending on the particular impedance at the separating ports, different phase differences between the input and output signals can be obtained. In particular, one can put a diode at both separating ports followed by an impedance transformer terminated by an open circuit. When the diodes are closed, the impedance is that of a short circuit, or rather of a very low resistance, wanted as low as possible to minimize the loss. When the diodes are open, the impedance is that of a capacitor, which permits to obtain a given phase difference between the signals at the output ports when the diodes are switched from one state to the other.

Figures 8, 9 and 10 present measurements of the reflection coefficient, transmission coefficient and phase shift in function of frequency for the two diode states. Figure 8 shows that the frequency behaviour of the reflection coefficient is rather complex: in practice, one does not have much flexibility for optimizing the circuit; as shown in the figure, a reflection coefficient better than -17 dB over the band of interest was obtained.

The transmission loss, shown on Figure 9, is in the order of 1 dB. As mentioned above, these losses are to some extent a consequence of the diode forward resistance, which is in this case of the order of 1.5 Ω. The selection of an appropriate diode is a very delicate issue involving not only the obtainment of the desired phase shift, but also the losses to be suffered and the power carrying capacity of the device. The recovery time of the diode in order to maintain an appropriate reverse bias for the maximum rated power is also a consideration.



8. Phase shifter Reflection Coefficient for the two diode states.

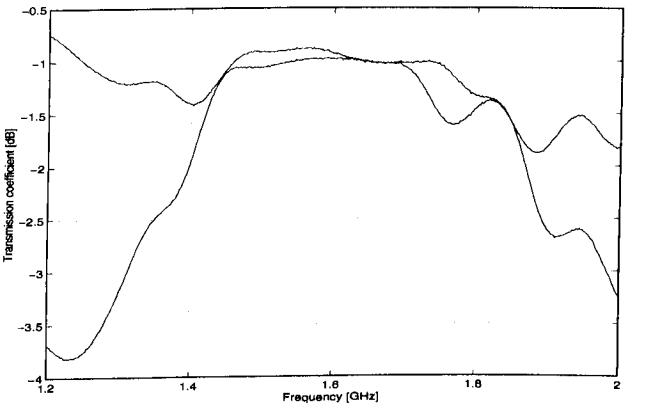


Fig. 9. Phase shifter Transmission Coefficient for the two diode states.

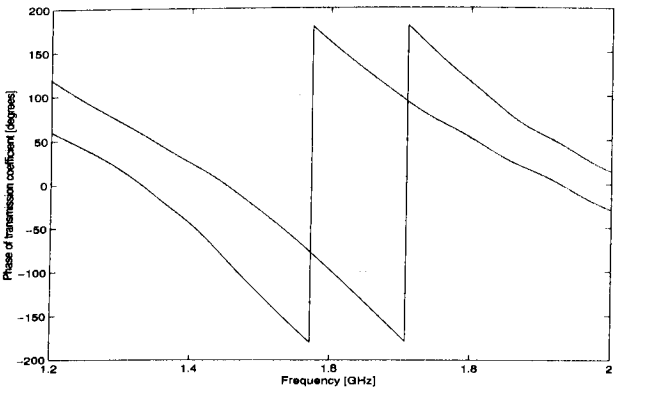


Fig. 10. Phase shifter phase characteristics for the two diode states.

ANTENNA ARRAY MEASUREMENTS

Measurements have been carried out with the 8-element antenna array prototype shown in Figure 2

Figure 11 shows, from left to right, measured antenna array radiation patterns in circular polarization obtained for a group of antenna branches with an appropriate phase shifter configuration. For these measurements, the antenna array was rotating on an horizontal plane with the other antenna fixed at an elevation angle of 35 degrees. These patterns are obviously very close to the design expectations.

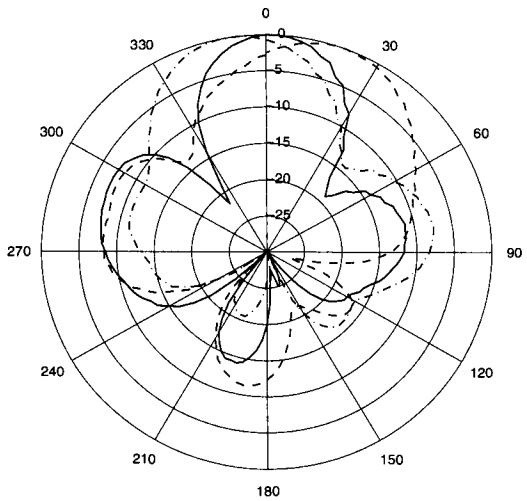


Fig. 11. Measured CP antenna array radiation patterns.

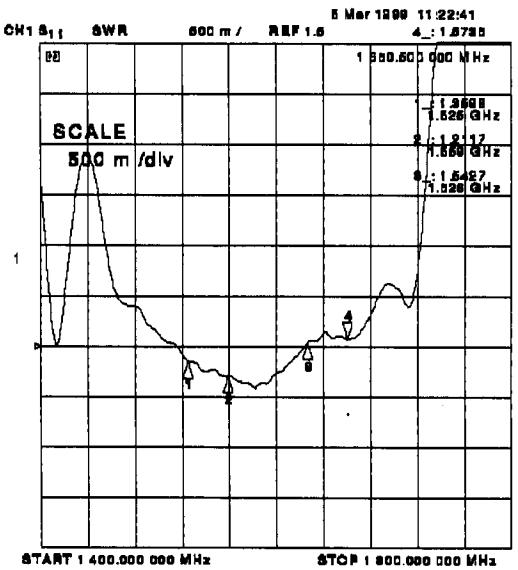


Fig.12. Typical measured VSWR.

Figure 12 shows a typical VSWR measurement over the band of interest for a set of three antennas. The switch is one of the main limiting factors for the bandwidth and the reflection coefficient.

A gain of 8 dBic has been measured at a 35 degree elevation angle. The array has been used in stationary tests for MSAT transmissions and the quality of the voice on the telephone link has been perceived to be comparable with the quality offered by the standard manufacturer's antenna. In its present design with commercial plastic encapsulated PIN diodes, the array is evaluated as capable of handling transmission powers up to approximately 15 Watts.

REFERENCES

1. Q. Wang, M. Lecours, C. Vergnolle, "Optimization of switches for radial satellite antenna array applications", Int. Mobile Satelite Conference IMSC'99, Ottawa, June 1999.
2. P. Lahaie, M. Lecours, "Analysis and measurements of reflection phase shifters at 1.5 and 18.5 GHz", ANTEM'98 Symposium on Antennas and Electromagnetics, Ottawa, August 1998, pp. 163-167.

ACKNOWLEDGEMENT

The authors acknowledge the support of NSERC's Technological Partnership Programme.

Optimization of Switches for Radial Satellite Antenna Array Applications

Qingyuan Wang, Michel Lecours, and Claude Vergnolle
Department of Electrical and Computer Engineering
Laval University, Quebec, Canada G1K 7P4
Tel: (418) 656 2131 ext.4883, Fax: (418) 656-3159
e-mail:qywang@gel.ulaval.ca, mlecours@gel.ulaval.ca, claude.vergnolle@wanadoo.fr

ABSTRACT

In this paper we present some general considerations for the design of a radial switch used in an electronically-steered circular array antenna for satellite communications. A low-pass filter circuit is used in the switch to lower down the insertion loss and reflection. Analytical formulae are derived for the case of ideal transmission lines to show the dependance of the return-loss bandwidth on the choices for the line impedances. Important optimization criteria are drawn.

These design criteria were used to implement economical L-band microstrip switches for use in an MSAT or INMARSAT antenna array, using low cost printed circuits and surface mount plastic-encapsulated PIN diodes. The measured results of one sample switch is presented here in which the overall insertion loss was lower than 0.50 dB, reflection lower than -13.8 dB and the isolation between the feed and a deactivated channel higher than 28.5 dB, for the whole operation band from 1.525 GHz to 1.661 GHz.

1. INTRODUCTION

We are developing a microstrip antenna for MSAT and INMARSAT mobile communications. The 8-channel antenna array[1], which uses a radial microstrip switch, is electronically-steered and covers the operation band from 1.525 GHz to 1.661 GHz. The adjacent channels closest to the direction of the satellite are activated simultaneously, and the other channels deactivated. The system is designed to cover elevation angles up to 70° .

In 1996 [2,3], researchers at University of Queensland reported an electronically-steered microstrip switch for the Australian mobile communication system. Insertion loss was -1.5 dB and the reflection at the boundary of the operation band was -10 dB. The switch worked well except that the insertion loss was a little too high compared with the overall gain of the antenna system, which was about 10 dB. The motivation of the present work is to explore the possibility to lower down the insertion loss while keeping the reflection in the operation band as low as possible, while using low cost printed circuit and commercial PIN diodes.

This paper is composed of 5 sections. In Section 2, the inductance required by a low-pass filter circuit to compensate the shunt capacitance of the diode in each activated channel is presented and the improvement to the operation of the switch is discussed. In Section 3, the Q-factor of an ideal transmission line switch is given and the criteria for wideband and low-loss switch design are established. In Section 4, the physical parameters of a microstrip switch designed using low-pass filter circuits and the above criteria are detailed and the simulated results are presented with the measured ones. In the last section, some conclusions are reached.

2. SELECTION OF PIN DIODES

A dual-anode PIN diode is used to activate a channel or deactivate it by positively or negatively biasing it. Fig. 1 gives the equivalent circuit of the diode. The variable resistor R is the very small forward resistance r_d when the diode is positively biased and it becomes very high (10000Ω in our simulations) when negatively biased. The inductances of the two inductors in the figure are the

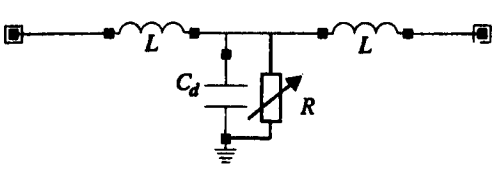


Fig1. Diode equivalent circuit

anode-lead inductances plus inductances we intend to add to the circuit. The inductance of the cathode-lead is compensated by the virtual ground formed by an open circuited quarter-wavelength line segment. When the diode is negatively biased, the channel is activated with some reflection and insertion loss caused by the shunt capacitance C_d. Deterioration arises at higher frequency or when the shunt capacitance of the diode is high. To solve this problem, we can adjust the inductances shown in Fig. 1 to form a low-pass filter circuit. The required inductance is:

$$L = \frac{Z^2}{2} C_d, \qquad (1)$$

where Z is the impedance of the transmission lines connected to both sides of the circuit. For the 1.6 GHz band and for a PIN diode with C_d = 0.75 pF, r_d = 0.6 Ω, Z = 50 Ω, Fig. 2 shows the insertion loss versus frequency for different inductor values L = 0, L = 0.47 nH and L = 0.94 nH. When L = 0 nH, in the operation band from 1.525 GHz to 1.661 GHz, the shunt capacitance will cause in each activated channel an insertion loss of at least -0.16 dB and a reflection coefficient of at least -14.35 dB. Whereas when we choose L = 0.94 nH, the transmission coefficient is better than -1.43×10^{-4} dB, and the reflection is lower than -42.50 dB in the whole band. When the operation frequency is higher, the lead inductance must be adjusted to minimize the reflection coefficient and the improvement will be more apparent. The improvement is also important in the case of a higher diode shunt capacitance. When the capacitance is C_d = 2.2 pF[2], and for Z = 50 Ω and L = 0 nH, the insertion loss will be as high as -1.25 dB and the reflection -6.08 dB in the operation band from 1.525 GHz to 1.661 GHz. If we choose the compensation inductors according to (1), L = 2.75 nH, the insertion loss will be lower than -0.18 dB and the reflection -14.66 dB.

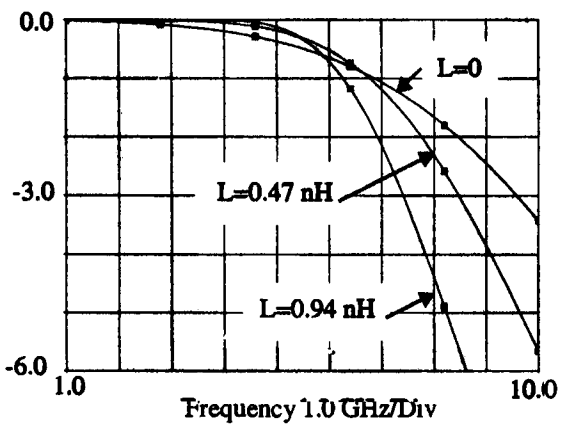


Fig. 2. Insertion loss in dB of a low-pass filter vs. frequency.

3. THE BANDWIDTH AND INSERTION LOSS OF AN IDEAL TRANSMISSION LINE SWITCH

Let us consider the case of an m-channel ideal transmission line switch, among which n channel are activated and the other $m-n$ channels deactivated. Because of the limited place at the center of the switch in which the circuit should be deployed, simplicity was our main principle in choosing the circuit configuration. This led to the switch composed of m radial channels, with each channel as shown in Fig. 3. It is composed of a quarter-wavelength transformer of impedance Z_{02} and another quarter-wavelength transformer of impedance Z_t and is terminated by a load Z_l. The m radial channels join together at the switch junction and are fed by a common perpendicular coaxial line of impedance Z_0. A quarter-wavelength open stub of impedance Z_s is connected in parallel with each channel through a PIN diode. The open $\lambda/4$ stub provides a virtual ground for microwave signals at the cathode of the diode.

When the diode is negatively biased, it provides a high resistance in parallel with a shunt capacitance C_d and isolates this channel from the virtual ground. The channel is then activated. When it is positively biased, it appears as its residue resistance r_d and the virtual ground will cause the power in the channel to be reflected. This channel is then deactivated.

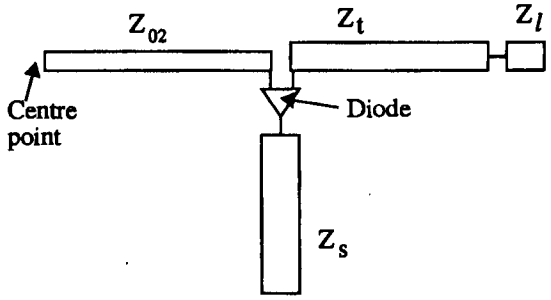


Fig. 3. One channel of the switch.

The transfer matrix of a quarter-wavelength transmission line of center frequency f_0 and arbitrary impedance Z_{00}, with a normalized frequency difference $x = \pi\Delta f/f_0$, is:

$$\begin{bmatrix} -\frac{1}{2}x & jZ_{00} \\ jY_{00} & -\frac{1}{2}x \end{bmatrix} \quad (2)$$

Using such matrices for each of the transmission lines in the switch, we can obtain the input impedance at the feed of the antenna. In the derivation, a positively biased diode is modelled by its residue resistance r_d and a negatively biased one by an open circuit. The effect of the shunt capacitance C_d is deemed negligible once compensated by the self inductance. By comparing this input impedance with that of an oscillating circuit, we can obtain the Q-factor of such a switch:

$$Q = \frac{\pi}{4}\frac{1}{1+\frac{(m-n)\,\bar{r}_d}{n\left(\bar{r}_d+\bar{Z}_e\right)}} \times \left(\frac{\left(\sqrt{\bar{Z}_l}+\sqrt{n}\right)\left[\sqrt{n\bar{Z}_l}-\bar{Z}_e\right]}{\sqrt{n\bar{Z}_l\bar{Z}_e}} + \frac{(m-n)\left[\bar{Z}_s+\sqrt{n\bar{Z}_e}\right]}{n\bar{Z}_e} - \frac{(m-n)\,\bar{r}_d^{\,2}\sqrt{\bar{Z}_e}}{n^{3/2}\left(\bar{r}_d+\bar{Z}_e\right)^2} \right) \quad (3)$$

where

$$Z_e = Z_t^2/Z_l \quad (4)$$

is the equivalent load impedance seen at the anodes of the diode. The impedances with over-bars are the normalized ones with respect to the feeding line impedance Z_0 and Z_{02} is related by the matching condition

$$Z_{02} = \sqrt{nZ_0Z_e} \quad (5)$$

Broad bandwidth is achieved when the Q-factor in (3) is small. We can establish the following criteria for wide band switch design:

1. Low impedance of the open stubs, Z_s. In (3), we recognize that the contribution of the *m-n* deactivated channels to the Q-factor is represented by the second term in the large round bracket. Lowering Z_s, the Q-factor will be decreased.

2. High equivalent load impedance, Z_e. In the same bracket of (3), both the first term, which represents the contribution of the n activated channels to the Q-factor, and the second term will decrease with increased Z_e, suggesting the possibility of a large bandwidth. The third term in the bracket, the contribution of the residue resistances of the diodes when they are positively biased, is usually very small compared to the other two terms.

3. Large n or small *m-n* when m is fixed. This is easy to understand because both cases are tending towards ideally matching a transmission line with n parallel transmission lines.

For each deactivated channel, at the center frequency of the switch, the residue resistance of each positively-biased diode in parallel with Z_e provides a shunt impedance of $\frac{Z_{02}^2(r_d+Z_e)}{r_dZ_e}$ at the feed port of the switch. The total $m-n$ deactivated channel provide a shunt resistance of $\frac{Z_{02}^2(r_d+Z_e)}{r_dZ_e(m-n)}$ to the n channels which are matched to the feed line. The insertion loss due to the m-n deactivate channels at the center frequency is

$$T(dB) = 20\log\left(1-\frac{(m-n)\,\bar{r}_c}{2n\left(\bar{r}_c+\bar{Z}_e\right)}\right). \quad (6)$$

According to this formula, for the insertion loss, we can draw the following criteria:

1. When n, the number of activated channels, is fixed, the insertion loss will increase with *m-n*, the number of deactivated channels.

2. When *m-n*, the number of deactivated channels, is fixed, the insertion loss will decrease with n, the number of activated channels.

3. The insertion loss can be decreased by increasing the equivalent load impedance Z_e at the PIN diodes.

4. MICROSTRIP SWITCH DESIGN UNDER THE GUIDANCE OF THE DESIGN CRITERIA

The above design criteria were used to design microstrip switches with various number of total channels among which some adjacent channels were activated. To decrease the cost of the switch, we used an economical substrate with both sides coated with 1 oz. copper. First we chose Z_s = 43.8 Ω. A lower impedance with a wider microstrip may be used but the radiation will also increase. Second we chose Z_e = 50 Ω, the same as the load impedance Z_l. In our case this choice was good enough for an acceptable bandwidth. For the sample switch we describe here, 3 channels in the total 8 channels are activated. From (5) the impedance of the quarter-wavelength transformers between the center point of the switch and the diodes was 86.6 Ω.. A higher Z_e would require this width to be smaller and this might eventually be an important factor limiting the power-handling capacity of the switch.

A PIN diode with a nominal shunt capacitance of 0.75 pF was used for this test. The inductance of each of the two anode leads and of the cathode lead were measured to be 0.47 nH. In Section 2, we showed that the total inductance of the anode leads should be 0.94 nH. The extra 0.47 nH plus the microstrip transformer Z_{02} behaved as the same microstrip line with reduced length. The extra 0.47 nH inductance added to the anode lead connecting to the microstrip line Z_t was a piece of narrow microstrip. The inductance of the cathode lead and the effective capacitance of the open stub because of the fringing field in each channel were compensated by the open stub. Consequently the actual length of the open stub was shorter than a quarter-wavelength.

The switch was fed at the center by a coaxial cable and then divided into 8 channels. Fig. 4 is the simulated return loss at the feed, the insertion loss between the feed and an activated channel, and the isolation between the feed and a deactivated channel. For the insertion loss, please refer to the numbers on the right side of the figure.

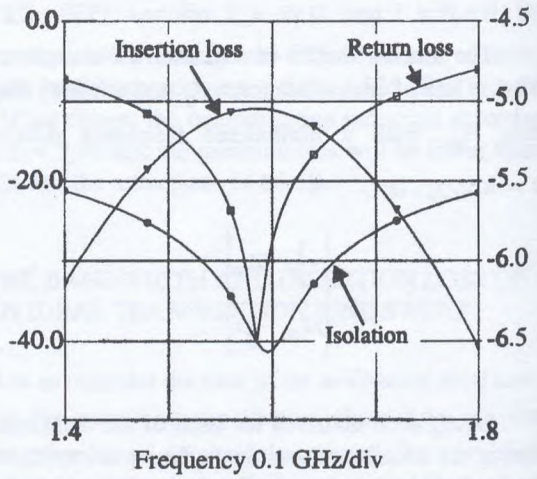


Fig. 4. Simulated return loss, insertion loss and isolation of the switch in dB.

Fig. 5 shows the circuit fabricated according to the simulations with one channel connected to the Port 2 of a HP 8703 Light-Wave Component Analyzer and the other 7 channels terminated by 50Ω matched loads. Fig. 6 shows the return loss, the insertion loss and the isolation measured by the analyzer. From Fig. 6, the return loss at the feeding port was -31.72 dB at 1.580 GHz, the overall insertion loss -0.33 dB and the isolation -42.96 dB at 1.600 GHz. The overall insertion loss was obtained by adding 4.771 dB to the insertion loss. For the whole operation band from 1.525 GHz through 1.661 GHz, the insertion loss was lower than -0.50 dB, the reflection lower than -13.76 dB and the isolation better than -28.45 dB. Similar results were obtained by activating different groups of adjacent channels.

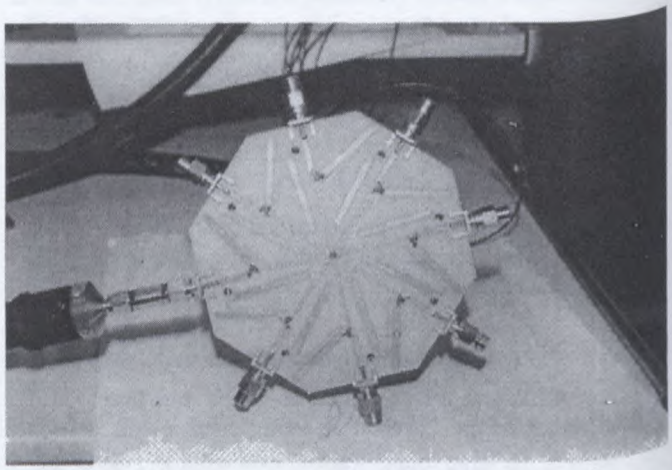


Fig. 5. L-band microstrip switch under test.

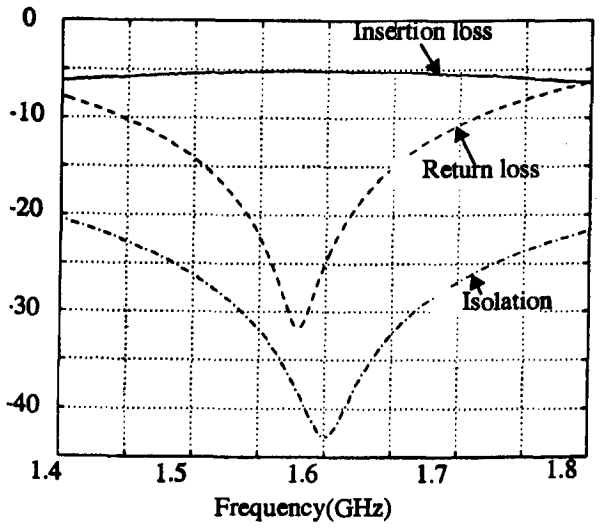


Fig. 6. Measured return loss, insertion loss and isolation of the switch in dB.

5. CONCLUSIONS

In this paper we suggested using a low-pass filter circuit in the switch design for low insertion loss and low reflection in the activated channels. The improvement to the operation of the switch is more apparent for higher operation frequencies and/or higher shunt capacitance of the PIN diodes.

Concerning bandwidth, the Q-factor of a radial switch composed of ideal transmission lines was used to establish criteria for wide bandwidth switch design. The number of deactivated channels should be small and the number of activated channels large to give wide bandwidth. When these numbers are specified, lowering the impedance of the open stubs and increasing the equivalent load impedance seen at each diode, will also increase the bandwidth. In our case the equivalent load impedance Z_e was chosen the same as the load impedance $Z_l = 50\Omega$. In the case of a switch with more deactivated channels, Z_e may be increased to increase the otherwise limited bandwidth.

About the insertion loss of such a switch composed of ideal transmission lines, it was suggested to decrease the number of deactivated channels and increase the number of activated channels, and increase the equivalent load impedance Z_e at the PIN diodes, for lower insertion loss at the center frequencies. Any measure discussed in the last paragraph for increasing the bandwidth will lower down the insertion loss at frequencies on both sides of the center frequency of the switch.

In the test switch presented in this paper, the measured insertion loss was lower than -0.50 dB, reflection lower than -13.76 dB, and the isolation better than -28.45 dB, for the whole operation band from 1.525 GHz to 1.661 GHz.

The insertion loss at the center frequency was measured to be -0.33 dB, which can be divided into three parts: loss caused by the residue resistance of each diode in the deactivated channels, according to (6), was -0.09 dB; loss caused by the microstrips between the feed of the switch and the port of an activated channel was about -0.15 dB; loss of -0.09 dB including radiation loss at the discontinuity between microstrips and mismatch losses at the feed of the switch and at each port of the activated channels.

REFERENCES

[1]M. Lecours, M. Pelletier, P. Lahaie, T. Breahna, Q. Wang, G.-Y. Delisle, "Experimental results with a circular electronically steered antenna for mobile satellite communications", IMSC 99.

[2]M. E. Bialkowski, S. T. Jellett, and R. S. Varnes, "Electronically steered antenna system for the Australian Mobilesat", IEE Proc.-Microw Antennas Propag, Vol.143, No.4, pp. 347-352, 1996.

[3]Ross S. Varnes and Marek E. Bialkowski, "A switched radial power divider/combiner for a mobile satellite antenna application", Microwave Journal, pp.22-36, Nov., 1996.

ACKNOWLEDGEMENT

The authors acknowledge the support of Davicom Technology Inc., of the NSERC Technological Partnership Program, and of the Nortel Global External Research Program.

A Combination Monopole/Quadrifilar Helix Antenna For S-Band Terrestrial/Satellite Applications

Charles D. McCarrick
Seavey Engineering Associates, Inc.
28 Riverside Drive · Pembroke, MA 02359 USA
Email: cmccarrick@seaveyantenna.com

ABSTRACT

The effects on electrical performance of a choke-fed monopole antenna placed coaxially within a quadrifilar helix antenna are presented. A specific application of this antenna configuration is for reception of S-band digital audio broadcasts. The quadrifilar has a beam coverage optimized for satellite reception at elevation angles 25 degrees above horizon. The purpose of the monopole antenna is to receive linearly-polarized signals from terrestrial base stations in the event of signal fading of the primary satellite source, as might occur in urban canyons. The antenna is designed to be compact and intended for external mounting on a subscriber's automobile.

INTRODUCTION

Innovation in antenna design these days has less to do with theory than with implementation. To be of consumer-quality, antennas must be compact, rugged and economical. Diversity can be a highly-valued feature, particularly if it helps the antenna meet the above criteria. To this end, a combination monopole/quadrifilar helix configuration has been investigated that has application as a subscriber antenna for the reception of digital audio satellite broadcasts. Described in this paper is such a device operating over a frequency band of 2326 MHz ±10 MHz. The quadrifilar helix receives the satellite directly, while the monopole is only engaged during situations of inadequate signal strength, such as might occur in an urban canyon. Under such conditions, the reception duty is handed off to the monopole which can receive the broadcast from terrestrial base stations located strategically within the urban area. Studies indicate that vertically polarized signals ensure optimal transmission in multi-path environments of this sort, so the monopole is the natural complement to a vertically polarized base station. The antenna configuration is made compact by placing the monopole coaxially within the quadrifilar helix, and carefully positioning it so as to minimize electrical performance degradation between the two. The two antennas will be described first separately and then integrated together as a single unit.

THE QUADRIFILAR HELIX ANTENNA

Quadrifilar helix antennas have the unique attribute of producing a circularly polarized conical beam while maintaining a relatively slender and compact form factor [1]. These antennas are becoming increasingly popular and have been produced in substantial volumes for such applications as global positioning receivers, handheld satellite phones and radio-com terminals. The quadrifilar helix is relatively insensitive to a conducting element placed within its core, provided that the diameter of the internal element is reasonably less than that of the host quadrifilar. Generally, the diameter ratio between the two antennas should not exceed 1:3 [2], however short sections ($\leq\lambda/4$) of the internal conductor resulting in a diameter ratio of 1:2 are permissible if properly located. This leads to the present topic, in which a $\lambda/4$ choke is placed within a quadrifilar helix of twice diameter for the purpose of creating a coaxial monopole.

Quadrifilar Helix Test Model

Designing the quadrifilar helix is a straightforward procedure, however special care must be taken to ensure that the structure is properly excited for optimum performance. The main complexity of these antennas lies within the feed/matching network design. After selecting the helix parameters that give the desired beam coverage, the mutual input impedance of each winding must be characterized so that an appropriate feeding and matching network can be implemented [3]. The feeding network provides an equal distribution of RF signal power to the four helix windings in quadrature phase rotation so as to excite the proper beam mode. The matching network is a set of four reactive impedance transformers, one for each winding.

The geometry for the quadrifilar helix under consideration is shown in Figure 1. The helices have a winding sense for left-hand circular polarization, and an angular pitch for an elevation coverage of 25 degrees to zenith. The selected quadrifilar dimensions are 2.5-inch pitch, 0.75-inch diameter and 2 turns for a developed length of 5-inches.

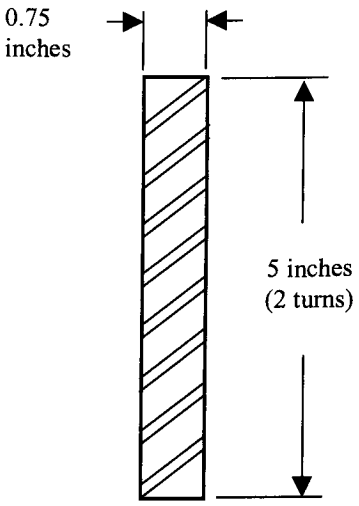


Figure 1 Quadrifilar helix element dimensions

The antenna was fabricated as a printed circuit and etched on a 5-mil substrate of Teflon-fiberglass. The measured input for each winding was 28Ω within the frequency band of interest, so a matching network of four 37Ω quarter-wave lines was incorporated to match the antenna to a 50Ω input. A branchline coupler was etched together with the matching network and helix winding to complete the circuit. The feeding/matching network adds an additional 0.625-inches to the overall antenna height. A typical beam pattern cut in the elevation plane measured for this antenna in the absence of the monopole is shown in Figure 2. The gain between 25 degrees and zenith is at minimum 1 dBic, with a peak gain of 3.8 dBic occurring around 45 degrees. Ellipticity is seen to be less than 2 dB within the coverage region.

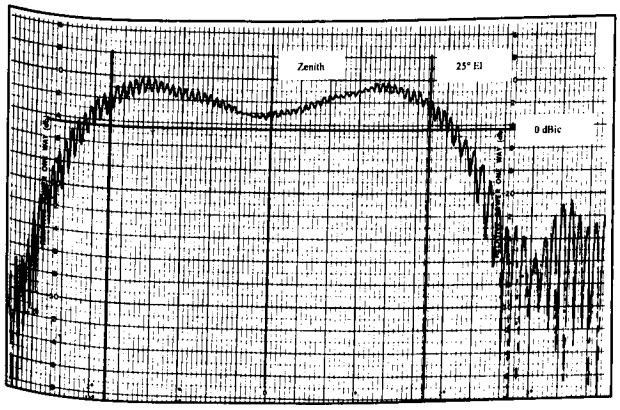


Figure 2 Elevation beam pattern of the quadrifilar helix antenna in the absence of the monopole.

THE MONOPOLE ANTENNA

The monopole antenna used here is actually a form of dipole antenna whose reflective element is a quarter-wave choke. The design consists of a 5½-inch length of UT85 semi-rigid coaxial line terminated with 50Ω SMA connector on one end, and has a section of the outer conductor removed at the other end to expose the center conductor. This forms a sort of rudimentary dipole element. A choke is used to sheath the outer conductor to prevent currents from flowing along its length and producing undesirable radiation.

Choked Monopole Antenna Test Model

The geometry for the choked monopole under consideration is shown in Figure 3. The selected dimensions are 0.93-inches length for the exposed center conductor (monopole), 1.2-inches length and 0.375-inches diameter for the choke. The choke is fabricated from 10-mil wall brass tube, and is soldered to the coaxial line outer conductor as shown.

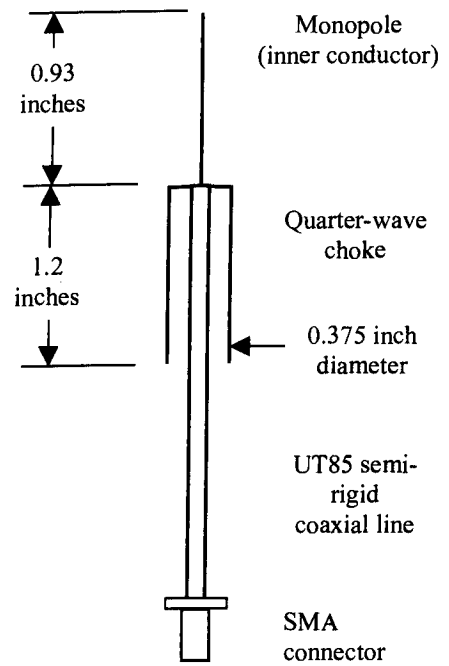


Figure 3 Choked monopole antenna dimensions

The beam radiation pattern for the choked monopole antenna in the absence of the quadrifilar is shown in Figure 4. The gain of the antenna at horizon is approximately 1.5 dBil and 1.9 dBil at 25 degrees above elevation. This region of coverage is deemed sufficient for terrestrial reception in urban areas.

COMBINING ANTENNAS

Placement of the monopole within the quadrifilar helix is critical. If improperly located, coupling between the two antennas will result in severe degradation of the radiation patterns, particularly for the monopole. In addition to being axially concentric, the feedpoint of the monopole (the point along the axis where center conductor and choke coincide) must be displaced vertically below the top of the quadrifilar by a specific distance. This distance was predicted by moment method modeling then verified through experimentation. Properly located, the quadrifilar helix patterns are virtually unaffected while the monopole patterns suffer only marginal degradation, as shown in Figure 5. In the present case, the optimal location was found to be 0.42-inches below the top of the quadrifilar, which means that the monopole protrudes 0.51-inches out the top of the structure. The combined antenna configuration is sketched in Figure 6.

CONCLUSION

A dual antenna configuration has been proposed with the intent of giving both satellite and terrestrial reception capability. This was accomplished by placing a choke-fed monopole within the core of a quadrifilar helix, and finding the location giving optimum performance. Some rules of thumb with regards to diameter ratios between the two antennas were discussed and verified through experimentation to hold true.

ACKNOWLEDGEMENT

The author wishes to thank Paul Medeiros and Mark Loveridge of Seavey Engineering Associates, Inc. for their collaboration and assistance in fabricating the antenna models and obtaining the measured test data.

REFERENCES

[1] C.C. Kilgus, "Shaped-Conical Radiation Pattern Performance of the Backfire Quadrifilar Helix", *IEEE Trans. on Antennas and Propagation*, Vol. AP-23, pp.392-397, May '75.

[2] W.T. Patton, "The Backfire Bifilar Helical Antenna", Technical Report No. 61, Aeronautical Systems Division, Wright-Patterson Air Force Base, Sept. '62.

[3] C.J. Mosher, "... The Impedance Behavior of a Quadrifilar Helical Antenna", Master of Science Research Report, UMASS Amherst, Feb. '97.

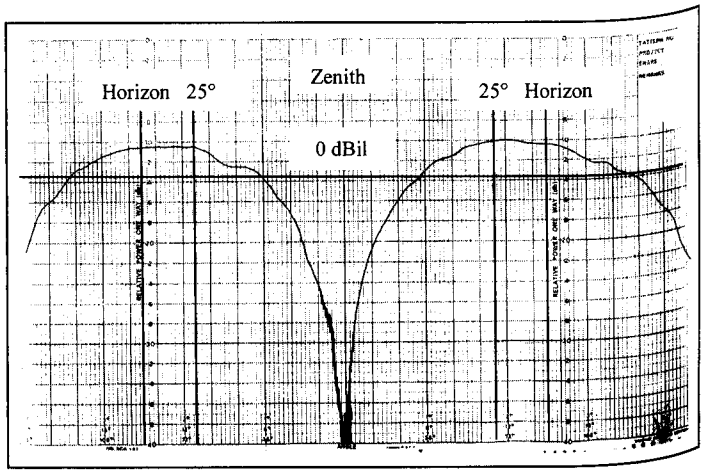


Figure 4 Elevation beam pattern of the monopole in the absence of the quadrifilar helix.

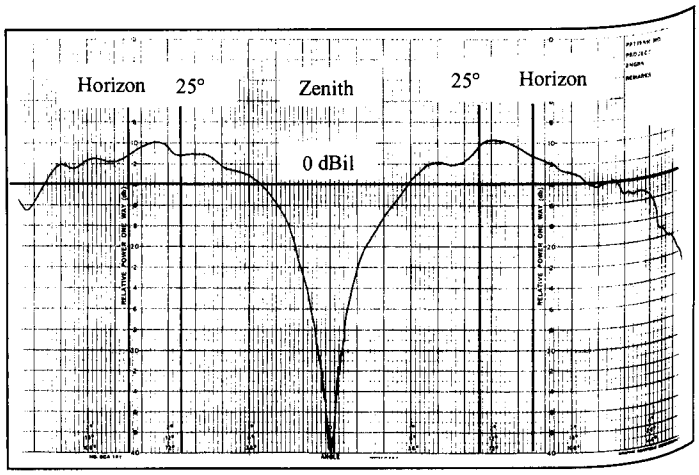


Figure 5 Elevation beam pattern of the monopole in the presence of the quadrifilar helix.

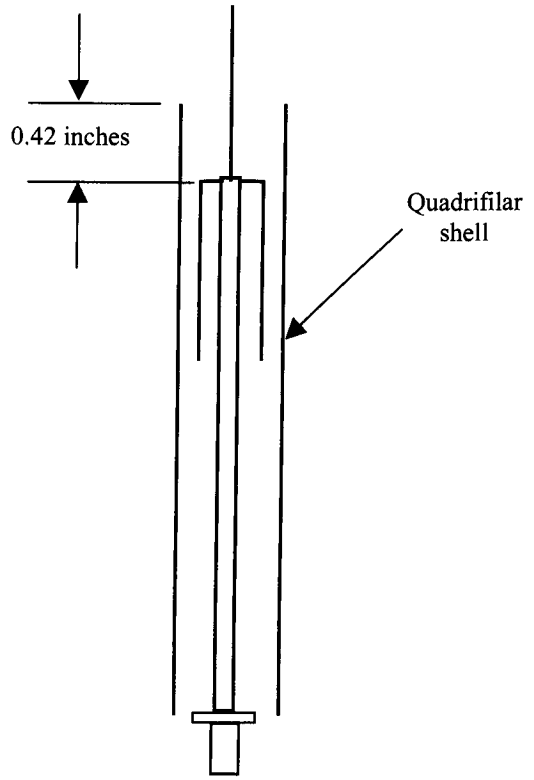


Figure 6 Location of monopole with quadrifilar helix.

A Polarization Agile Antenna for L-Band Mobile Communications

Aldo Petosa, Apisak Ittipiboon, Nicolas Gagnon
Advanced Antenna Technology
Communications Research Centre
3701 Carling Avenue, P.O. Box 11490, Station H
Ottawa, Ontario, K2H 8S2, Canada
e-mail: aldo.petosa@crc.ca

ABSTRACT

This paper presents an antenna configuration which can radiate in any one of four polarization states: clockwise circular, counter-clockwise circular, vertical linear, or horizontal linear. A prototype feed network was developed and integrated with an L-band wide-band microstrip antenna in order to test the design concept. The feed network contains digital phase shifters, whose states could easily be electronically controlled in order to radiate the desired polarization.

The antenna patterns were measured for each of the various settings of the phase-shifters in order to confirm the operation of the polarization agility. All four polarization states were generated. Axial ratios on boresight of 0.5 dB were achieved for both clockwise and counter-clockwise circular polarization, with a 3 dB axial ratio bandwidth of approximately 20%. For the linear-polarized states, cross-polarization levels on boresight greater than 20 dB below the peak co-polarized levels were measured. For every state, 10 dB return loss bandwidths of at least 10% were obtained. This antenna configuration offers polarization agility with little increase in complexity and without requiring additional real-estate or a significant cost increase.

INTRODUCTION

One of the challenges facing mobile communications in urban environments is to overcome signal fading due to depolarization from multi-path scattering. One method for combatting this fading is to use a mobile terminal that incorporates antennas with polarization diversity. Having more than one antenna for the mobile terminal, however, is not the ideal solution due to the added real-estate, complexity, and costs required. Having a single antenna which could radiate with different polarizations would be an attractive solution.

In the past few years, there has been some research carried out on antennas with polarization diversity. One method involved microstrip antennas fabricated on ferrite substrates [1, 2]. The antenna polarization was switched from linear to circular by applying a dc magnetic bias field. Similar results have also been obtained using ferrite resonator antennas [3]. Both these antennas suffer from certain disadvantages such as: narrow bandwidth performance (about 1%); the requirement for an electromagnet which significantly increases the cost and size; and only one orientation of linear polarization can be achieved.

A second approach involved the integration of four varactor diodes with a single probe fed microstrip antenna [4]. By proper biasing of the diodes, the antenna generated vertical linear, horizontal linear, clockwise circular and counter-clockwise circular polarization. This approach is more amenable to practical applications; however, the impedance bandwidth reported was still quite narrow (< 1%). Also, in this configuration, the diodes were inserted into the substrate, adding to the complexity of the antenna, and increasing labour costs in volume production.

A third configuration consisted of a stacked disk antenna with four probes, where the probes were connected using a network of three hybrid circuits [5]. The polarization of this antenna could be altered between CW and CCW circular polarization by appropriate selection of one of two input points of the feed configuration. A similar approach with only two hybrid circuits was also proposed which could provide either VL or HL polarization. Although only two polarization states were achieved, this configuration offered a wide impedance bandwidth of over an octave.

This paper presents a simple four-point feed antenna configuration for generating radiating fields in any one of four polarization states.

ANTENNA CONFIGURATION

The geometry of the polarization agile antenna is shown in Figure 1. It consists of a square microstrip patch antenna, printed on a 0.76 mm thick sheet of fibreglass, and suspended 7 mm above the ground plane with air as the substrate. This height was chosen in order to obtain a 10 dB return loss bandwidth of approximately 10%, required for various mobile communication applications. The antenna is excited with four probes, which are connected to the microstrip feed network, located below the ground plane. The probes were located in positions where there was low mutual coupling between orthogonal probes (i.e. between 1 and 4; 2 and 3). Microstrip stubs were used to tune out the reactive component of the probes and to provide an adequate impedance match.

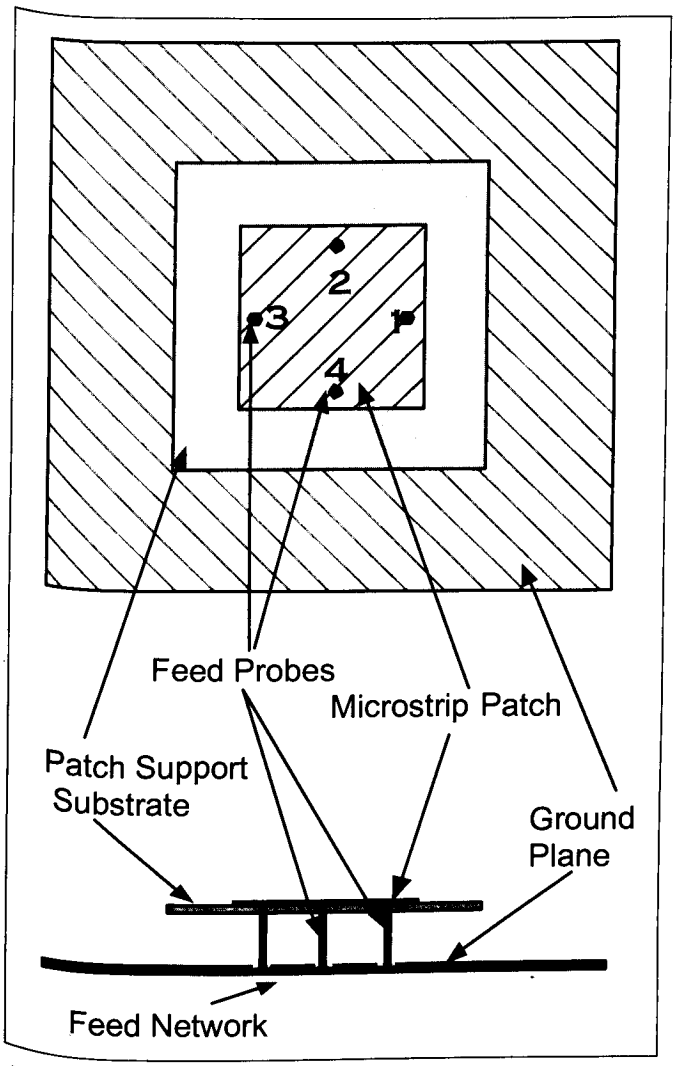


Figure 1. Geometry of polarization agile microstrip antenna.

The four probes are connected together with a combination of Wilkinson power combiners and one-bit digital phase shifters. By properly setting the states of the phase bits, the phase at each of the four ports can be adjusted to produce a radiation pattern with either clockwise circular, counter-clockwise circular, vertical linear, or horizontal linear polarization.

For the prototype antenna, the one-bit phase shifters were implemented by inserting transmission line lengths corresponding to 0° and 180° of phase delay at the design frequency of 1.5 GHz. In practical applications, one-bit digital phase shifters could be used to electronically control the polarization states.

Figure 2 shows the measured return loss of the antenna for the four different polarization states. The impedance bandwidth for the circular polarization states is 24% with a 15 dB return loss. The linear polarization impedance bandwidths are somewhat degraded but still maintains a 10 dB return loss over a 10% frequency range. The impedance bandwidth of the four-port patch is significantly wider than that of the single probe-fed patch due to the Wilkinson power dividers, which absorb out-of-band reflections.

The radiation patterns at 1.5 GHz in the two principal planes for the four polarization states are shown in Figures 3 to 6. A rotating linear source was used to measure the axial ratio of the circular polarized states. A boresight axial ratio of approximately 1.5 dB was measured for both the clockwise and counter-clockwise states. The gain of the antenna when radiating circular polarization was approximately 7 dBic (determined using the conversion from a rotating linear measurement as outlined in [7]). The axial ratio and gain is plotted versus frequency for the clockwise circular polarization state in Figure 7. For the linear polarizations, cross-polarized levels were better than 20 dB below co-polarized peak gains. Linear polarized gains are approximately 5 dBi. The skew in the E-plane pattern of vertical polarized antenna is probably caused by probe radiation.

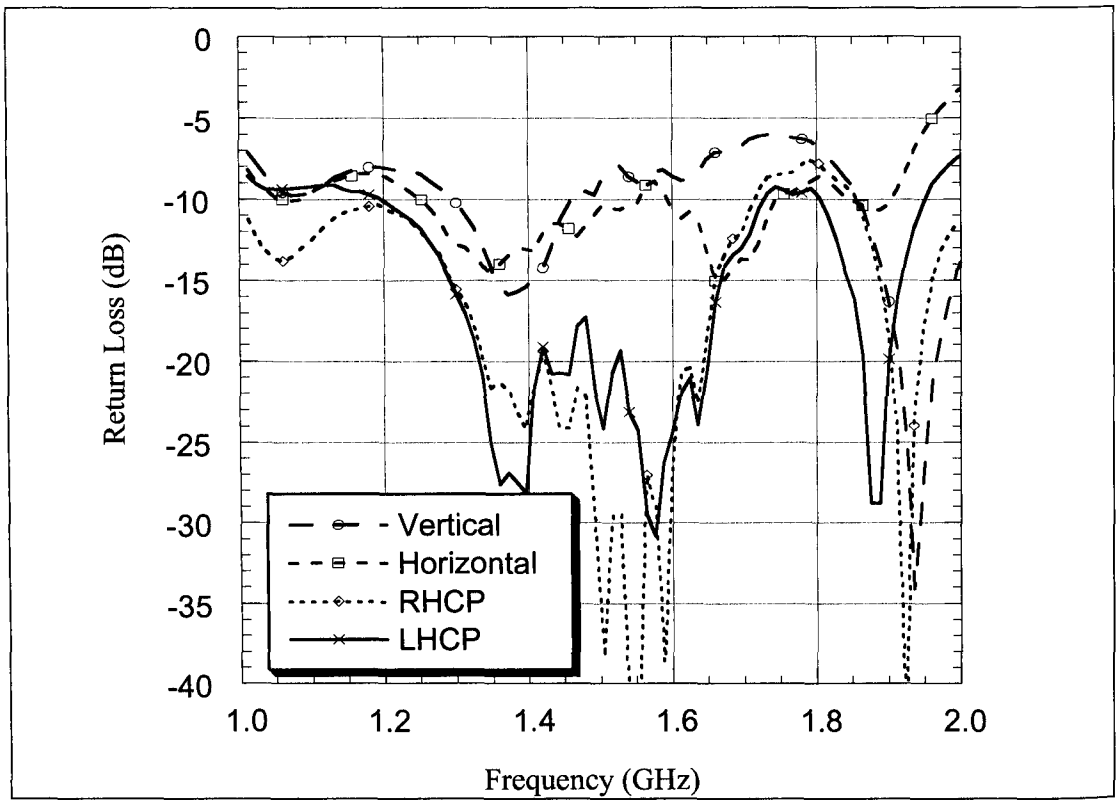


Figure 2. Return Loss for the four polarization states.

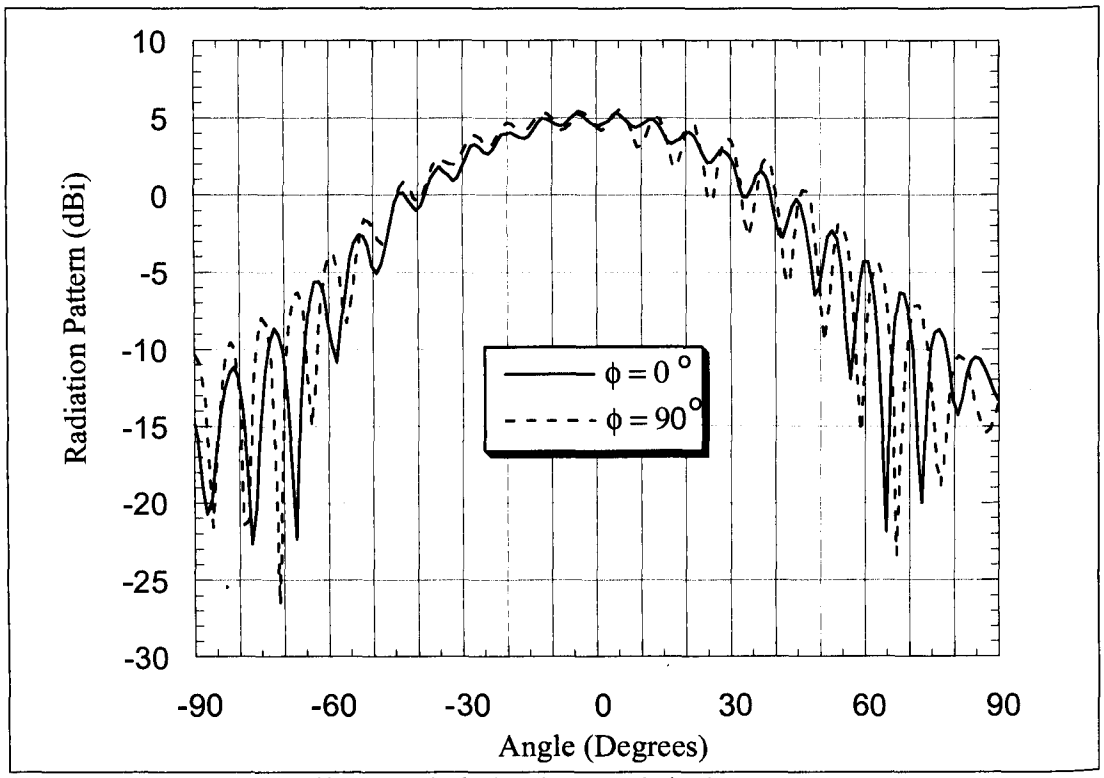


Figure 3. Clockwise circular polarization.

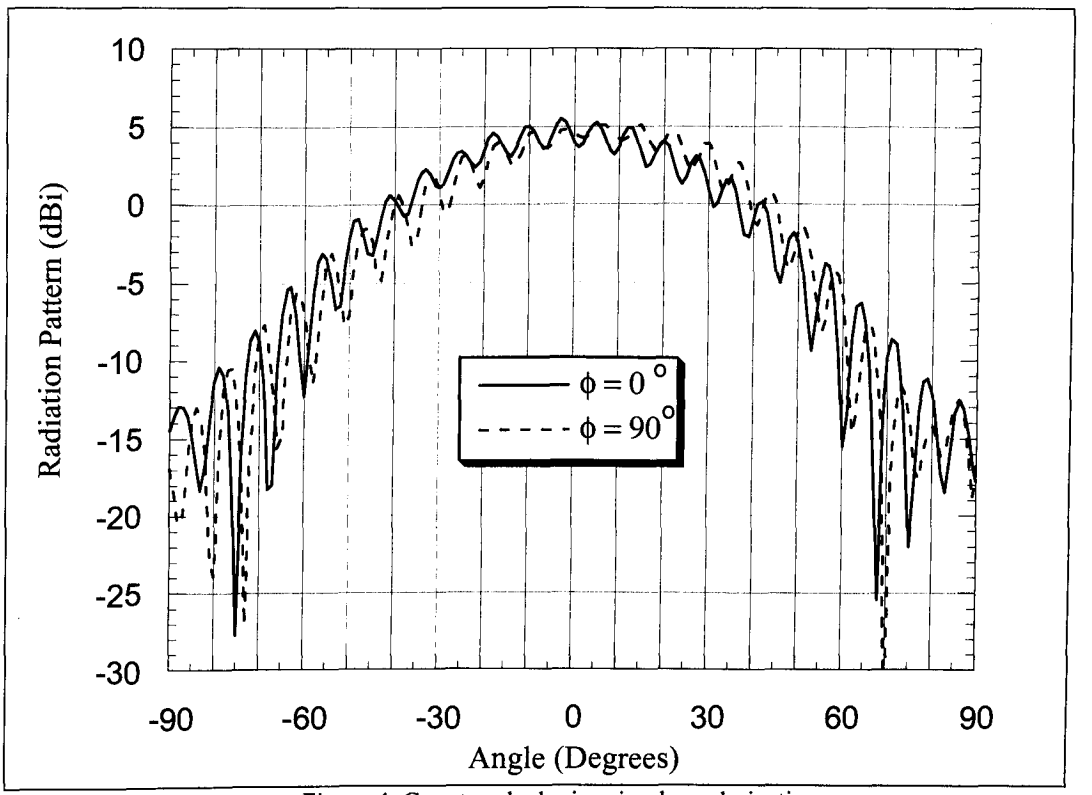


Figure 4. Counter-clockwise circular polarization.

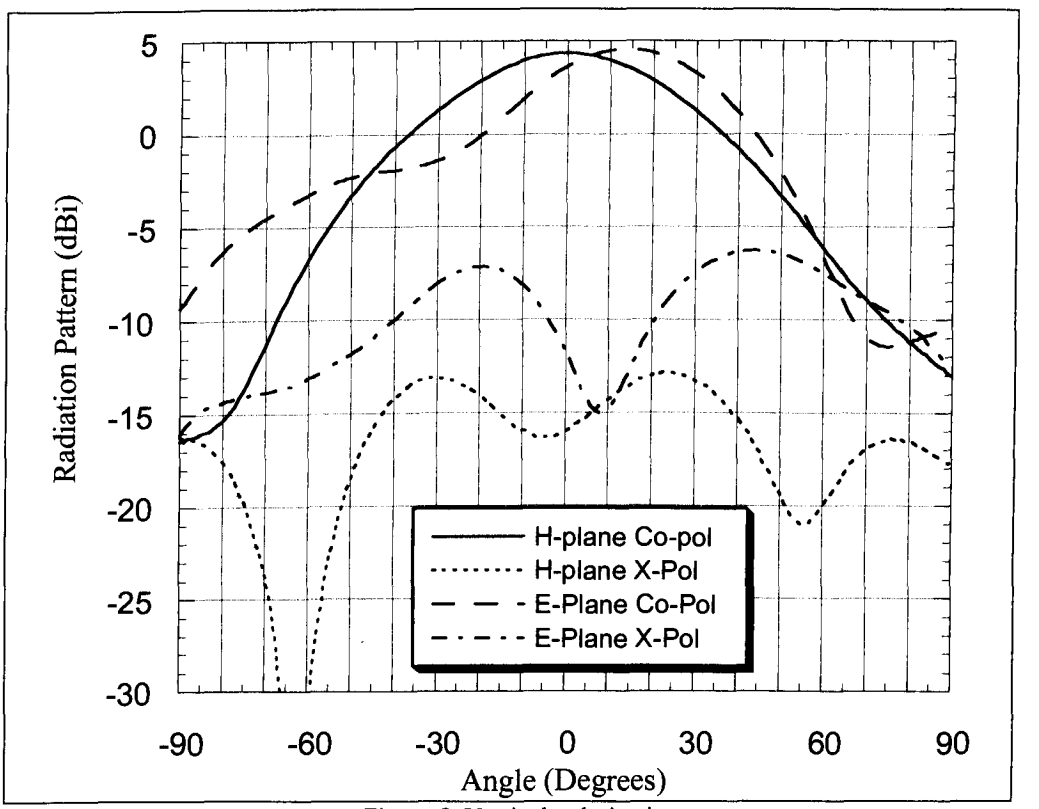


Figure 5. Vertical polarization.

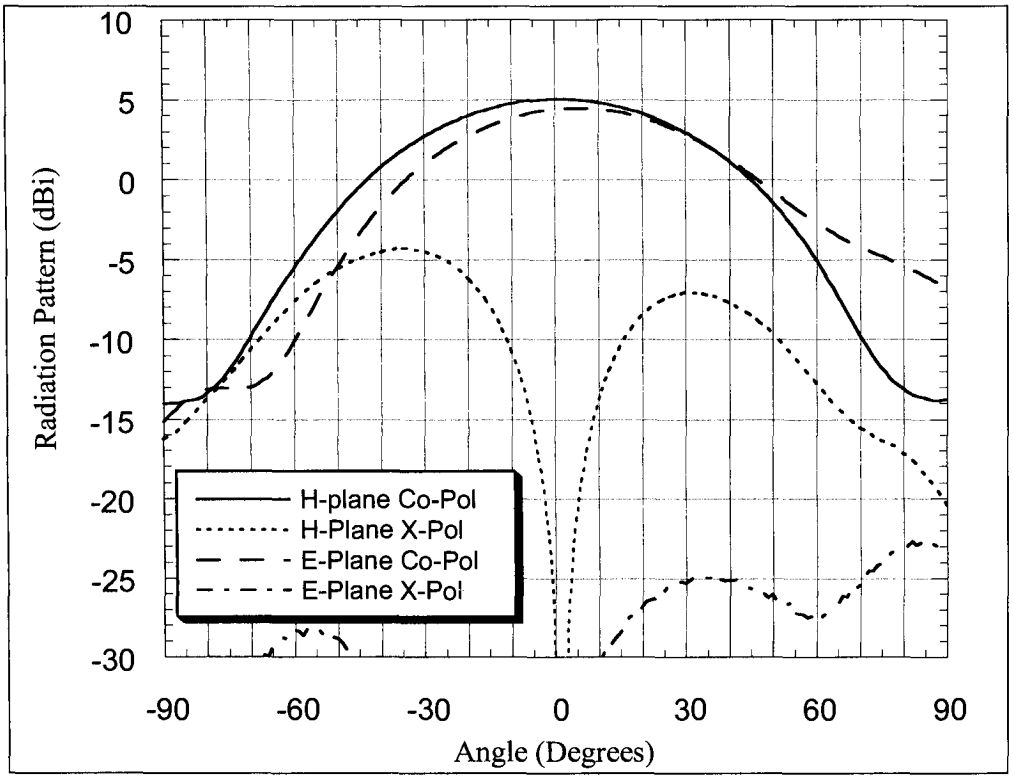


Figure 6. Horizontal polarization.

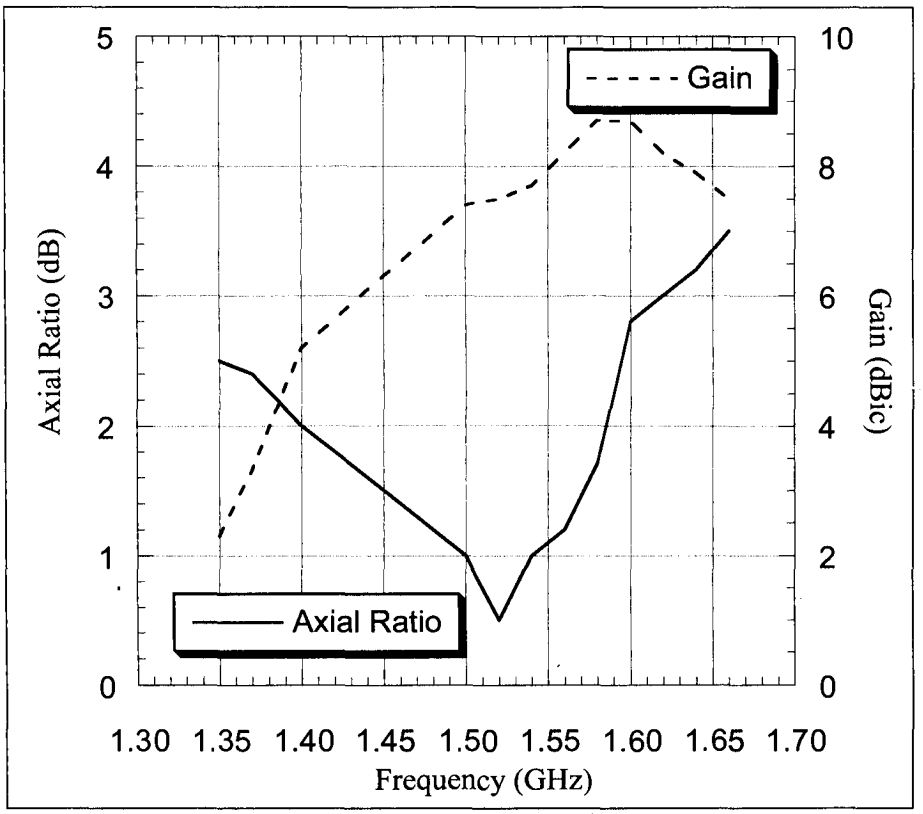


Figure 7. Gain and axial ratio vs. frequency for the CW-CP state.

SUMMARY

This paper has presented a four-point fed microstrip antenna with polarization agility. By incorporating digital phase shifters in the feed network, the phases at each of the four ports can be adjusted to allow the antenna to radiate any one of clockwise circular, counter-clockwise circular, vertical or linear polarization. A prototype antenna designed at L-Band was tested to determine the performance of this configuration. This antenna configuration offers polarization agility with little increase in complexity and without requiring additional real-estate or a significant cost increase. Although this concept was demonstrated using a microstrip patch configuration, it is equally applicable to any antenna having a rotational symmetry of at least $\pi/2$, such as conical horns, pyramidal horns, or crossed dipoles.

REFERENCES

[1] D.M. Pozar, "Radiation and Scattering Characteristics of Microstrip Antennas on Normally Biased Ferrite Substrates," *IEEE Trans. on Antennas and Propagation*, Vol. 40, No. 9, Sept. 1992, pp. 1084 - 1092.

[2] K.K. Tsang and R.J. Langley, "Annular Ring Microstrip Antennas on Biased Ferrite Substrates," *IEE Electonics Letters*, 1994, Vol. 30, pp. 1257-1258.

[3] A. Petosa, R.K. Mongia, A. Ittipiboon, and J.S. Wight, "Switchable LP/CP Ferrite Disk Resonator Antenna," *IEE Electronics Letters*, 1995, Vol. 31, pp.148-149.

[4] P.M. Haskins and J.S. Dahele, "Varactor-diode loaded passive polarization-agile patch antenna," *IEE Electronics Letters*, June 1994, Vol. 30 No. 13, pp. 1074-1075.

[5] A.T.S. Wang, R.C. Chu, K.M. Lee, "Planar, Low-Profile, Wideband, Wide-Scan Phased Array Antenna using Stacked-Disc Radiator," *Antennas and Propagation Symposium*, AP-S 1997, Montreal, Canada, pp. 702 - 705.

[6] C.A. Balanis, *Antenna Theory : Analysis and Design*, Harper Row, 1982, Chap. 2.

[7] D.S. Dunn, E.P. Augustin, "Measuring the Gain of Circularly or Elliptically Polarized Antennas," *IEEE Antennas and Propagation Magazine*, Vol. 36, No. 1, Feb. 1994, pp. 49-51.

Mobile Satellite in Ka-band

Shunichiro Egami
Shizuoka University
3-5-1 Johoku Hamamatsu, 432-8561 Japan
Email:tesegam@eng.shizuoka.ac.jp

ABSTRACT

This paper study mobile satellite systems concept in Ka-band. A wide allocated bandwidth and a large number of frequency reuse based on hundreds to thousands of small spot beams will allow to draw drastically new mobile satellite systems concepts. Requirements for beam size on the surface of the earth for various signals transmission are considered. Based on these requirements, Ka-band geostationary systems with 3.5m and 10m satellite antenna for mobile communications are considered. System parameters examples are shown for these systems. If number of beam is hundreds to thousands, it is not appropriate to provide a fixed power transmitter for each beam because traffic in each beam is not uniform or static. This paper proposes a method to cope with this multiple beam varying traffic, which is applicable in Ka-band. Satellite onboard equipment concept for hundreds to thousands beams are also presented.

INTRODUCTION

Shortage of L- or S-band mobile satellite service (MSS) frequency band and requirement for more higher bit rate multimedia communications will move future required MSS frequency band for more higher frequency band. This paper study advanced multiple beam mobile satellite systems concepts in Ka-band. A broad allocated bandwidth and a large number of frequency reuse based on hundreds to thousands of small spot beam enable to draw a drastically new mobile satellite systems concept. At first, requirements for beam size on the surface of the earth for voice and multimedia signals up link are considered. Signals will be 4.8kbps voice to 1~2Mbps multimedia communications. Based on these requirements, multiple beam system with 3.5m and 10m antenna on the geostationary satellite are considered.

This paper also considered requirements for multiple beam systems with varying beam traffic. If number of beam is on the order of hundreds to thousands, it is difficult to provide a fixed power transmitter for each beam. A large number of beams may include beams on the sea or on the deserted area where traffic is very spontaneous. In order to cope with these spontaneous or varying traffic spots, author proposed the "Multiport Amplifier" concept in 1987[1]. The concept was applied in AMSC/MSAT and Nstar for L- or S-band multiple beam mobile satellites communications successfully [2]. Multiport amplifier (MPA) consists of an input Hybrid, SSPAs and an output Hybrid. The MPA configuration is easily implemented in L or S-band [3]. However, in Ka-band, where coaxial or planer circuit loss are very large, implementing more than 8-port MPA is not realistic.

This paper considered method to implement power sharing in Ka-band multiple beam satellites. As a most appropriate method, a reflector antenna with space-fed equal phase shift active array, its principle was proposed by the author in 1987[4], is presented. The active array consists of hundreds to thousands of similar elements with equal phase shift. Ordinary primary feed horn illuminate the mesh of the active elements and the amplified circular wave illuminates the reflector. Since each beam is radiated from all active array elements, it is not necessary to assume fixed power for each beam. Beam traffic variations are accepted without loss of total transmitting power efficiency. Also, from this property, it is possible to place a spot beam on the ocean or on the deserted area where traffic is very spontaneous.

OMNIDIRECTIONAL TERMINAL

In mobile satellite communications, mobile voice terminal has omnidirectional antenna pattern. Users of mobile terminal need not know direction of the satellite. Fig.1 shows radiation and receiving pattern of the omnidirectional terminal.

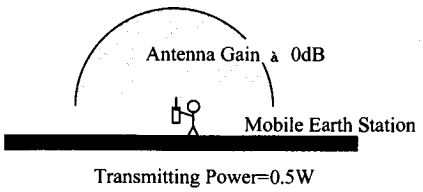


Fig.1 Mobile satellite terminal antenna pattern

This radiation pattern itself determines its isotropic gain. It 's isotropic gain is about 0dB which is not dependent on frequency. This frequency independent omnidirectional earth station antenna pattern provide specific features to the mobile satellite systems concept.

Effective aperture A_{eff} of 0dB isotropic gain antenna is given as follows when wavelength is represented by λ.

$$A_{eff}=\lambda^2/4\pi \qquad (1)$$

This equation shows that if receiving frequency is lower, the effective aperture of the omnidirectional antenna become larger. The larger aperture enables communications by lower satellite power flux density. This principle represents that lower frequency is more suitable for mobile communications although size of the 0dB antenna become larger. Fig.2 shows omnidirectional antenna examples at 1.6GHz and 29.5GHz. Size of omnidirectional antenna is 9-14cm at 1.6GHz and 0.5-1cm at 29.5GHz.

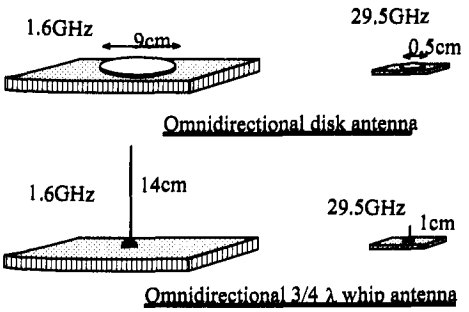


Fig.2 Omnidirectional antenna examples at L- and Ka-band.

REQUIRED SATELLITE ANTENNA DIAMETER

If transmitting power of mobile voice terminal is selected as around 0.5W at any frequencies, eirp of omnidirectional mobile terminal is also frequency independent. Eirp is around -3dBW at any frequencies. In this case, required satellite receiving antenna diameter become also frequency independent. The required satellite antenna diameter D is given as function of the required up link C/N_0 , satellite altitude and satellite receiving noise temperature as follows.

$$D = H*SQRT\{ C/N_{0,up} (32/\eta)(kT_s/eirp_E\} \qquad (2)$$

Where, D : Satellite antenna diameter
H : Satellite altitude
$C/N_{0,up}$: up link C/N_0 including up link margin.
η: Satellite antenna aperture efficiency
k : Boltzman's constant
T_s : Satellite system noise temperature
$Eirp_E$: Mobile terminal up link eirp

Satellite antenna diameter is proportional to satellite altitude. Figure 3 shows required satellite antenna diameter at LEO, ICO and GEO. Transmission signal is assumed to be 4.8kbps voice signal. In this case, required up link C/N_0 is 48 dBHz including 5dB up link margin.

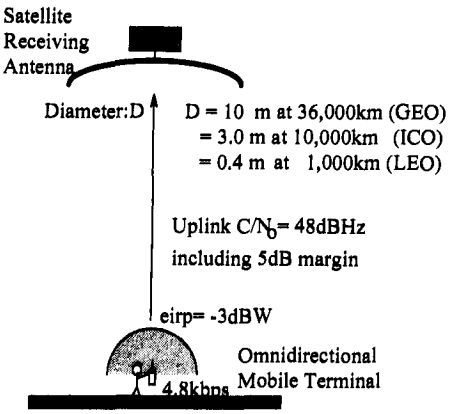


Fig.3 Required satellite antenna diameter to support omnidirectional voice terminal.

REQUIRED BEAM COVERAGE DIAMETER

If mobile terminal up link bit rate and eirp are fixed, required satellite receiving antenna diameter is proportional to distance from the earth to the satellite. It is not dependent on up link frequency. If mobile terminal antenna gain is 0dB and transmitting power is 0.5W(eirp = -3dBW), required satellite receiving antenna diameter is around 10m at GEO and around 0.4m at LEO for 4.8kbps voice communications including 5dB up link margin. The concept is shown in Figure 3. This means if we implement a 10m antenna in Ka-band satellite, we can use 0dB antenna for voice communications up link even in Ka-band. Size of 0dB omnidirectional antenna in Ka-band will be smaller than 1cm. It can be said that by utilizing a large diameter satellite antenna in Ka-band, extremely small size Ka-band terminal can be utilized. However, in these cases, satellite antenna beam coverage become very small and number of beams becomes extremely large.

In multimedia communications at 2Mbps, use of omnidirectional antenna is impossible from link parameters requirement. In this case, use of directional antenna with more than 10dB gain will become necessary. However, in Ka-band, size of this antenna is still extremely small, if large diameter satellite antenna is implemented.

Figure 4 shows required diameter of beam coverage on the earth which support 4.8kbps voice communications up link with 5dB margin at -3dB beam edge. As shown in Equation (2), required satellite antenna diameter is not dependent on frequency. But diameter of the beam coverage on the earth is dependent on frequency.

Author showed in the previous paper (5), that required spot beam coverage diameter R on the earth at nadir is given by

$$R = 0.216\lambda SQRT\{(\eta eirp_{ES}/kT_S)/(C/N_{0,up})\} \qquad (3)$$

where,
R : diameter of the uplink spot beam coverage
λ : uplink wavelength
η : Satellite antenna aperture efficiency

$eirp_{ES}$: earth station transmitting eirp
k : Boltzman's constant
T_S : satellite receiving system noise temperature
$C/N_{0,up}$: required uplink C/N_0 including uplink margin.

From this equation, it can be understood that spot beam coverage diameter is not dependent on the altitude of the satellite but proportional to the uplink frequency. Also, it is inversely proportional to square root of the required up link C/N_0.

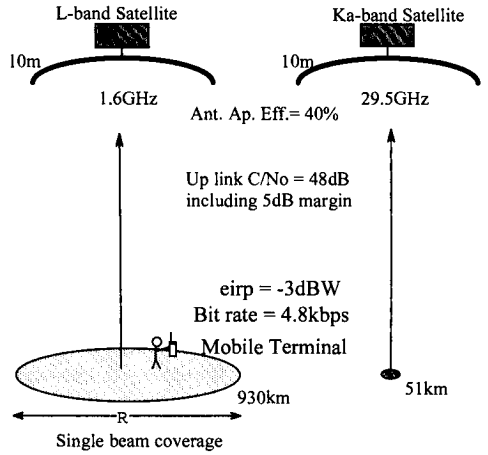


Fig.4 Required beam coverage to support 4.8kbps, -3dBW up link at L- and Ka-band.

NEVER LEO IN KA-BAND

In this study, use of a geostationary satellite (GEO) is assumed. Recently, many low earth orbit (LEO) systems are proposed and implementation is in progress. Before entering systems design, it should be emphasized that for Ka-band mobile or small aperture terminals systems, geostationary satellite is most appropriate. Figure 5 shows same beam coverage by LEO and GEO. Satellite communications are point (satellite) to area communications. Satellite antenna must cover required service area in order to support communications from any point in the service area. If GEO and LEO cover the same area as shown in Fig.5, transmitting power of the earth station or the satellite are same for GEO and LEO. It is easy to understand that point (satellite) to area communications, distance of the satellite from the earth do not affect required transmitting power. Small path loss of LEO is compensated by small satellite antenna gain of LEO. Therefore, small distance of LEO never assures small transmitting power for mobile terminal. Actually, small transmitting power is assured by small beam coverage(small cell). It is not dependent on satellite altitude. In principle, for the same signal, there exists required beam coverage, which is same for LEO and GEO. In Ka-band, required beam coverage diameter for voice communications using omnidirectional mobile terminal is extremely small. Therefore, in case of LEO, where satellite is orbiting, hand over between spot beams will become impossibly frequent. In Ka-band, use of omnidirectional terminal requires extremely small spot beam, which make LEO systems impossible.

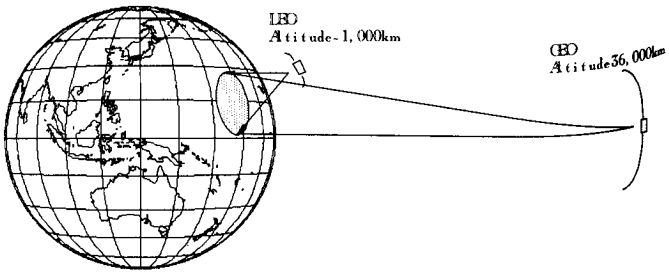


Fig.5 Same beam coverage means same transmitting power.

For multimedia communications at around 2Mbps, use of omnidirectional antenna becomes impossible from satellite link design requirement. In this case, use of the terminal with directional antenna become necessary. If directional antenna is used for earth station, in case of LEO, at least two tracking antennas are necessary to maintain continuous communications with orbiting satellites. This configuration is not realistic even in fixed satellite communications. Besides, tracking of the orbiting satellite is extremely difficult if the terminal is on a mobile platform. Therefore, it is apparent that it is impossible to envisage a LEO terminal concept in multimedia mobile satellite communications. In case of GEO, it is possible to envisage an inexpensive mobile terminal with directional antenna for multimedia mobile communications because satellite is stationary.
Also, using proposed power sharing multiple beams transmitting antenna concept, spot beams can be placed on the ocean or deserted area without loss of total power efficiency. In case of LEO, it is impossible to share beam power because one LEO satellite's whole coverage may be totally ocean or totally deserted area.

KA-BAND LARGE SATELLITE ANTENNA SYSTEMS

L-band GEO mobile satellite with 12m antenna will be launched in this or next year [6]. These systems support voice communications using omnidirectional mobile terminals. Omnidirectional mobile terminal has frequency independent antenna gain of around 0dB. Therefore, eirp of the mobile terminal is also frequency independent. From this frequency independent earth station eirp, required satellite antenna diameter to support omnidirectional terminal become also frequency independent. In principle, Ka-band 12m antenna will also support omnidirectional Ka-band mobile terminal. In the former section, author showed 10m antenna required for Ka-band omnidirectional voice terminal. Difference from above L-band 12m antenna is caused

by different up link margin and different up link bit rate [7].

Table 1 System parameters 3.5 and 10m antennas

Up Link (29.5GHz)						
Satellite Receive Antenna Diameter	3.5mϕ			10mϕ		
Spot Beam Coverage Diameter	144kmϕ			51kmϕ		
Transmission Bit Rate	4.8kbps	64kbps	1.544Mbps	4.8kbps	64kbps	1.544Mbps
Earth Station Antenna Diameter	2cmϕ	5cmϕ	15cmϕ	0.5cmϕ(0dB)	3cmϕ	10cmϕ
Earth Station Transmitting Power	0.5W	1W	3W	0.5W	0.5W	1W
Up Link Margin	8.9dB	8.6dB	9.1dB	5.2dB	10.3dB	9.9dB
Down Link (19.7GHz)						
Satellite Transmit Antenna Diameter	3.5mϕ			10mϕ		
Spot Beam Coverage Diameter	220kmϕ			76kmϕ		
Transmission Bit Rate	4.8kbps	64kbps	1.544Mbps	4.8kbps	64kbps	1.544Mbps
Earth Station Antenna Diameter	2cmϕ	5cmϕ	15cmϕ	0.7cmϕ(0dB)	3cmϕ	10cmϕ
Satellite Transmitting Power /ch	0.5W	1W	2W	0.5W	0.5W	1W
Down Link Margin	8.4dB	8.1dB	8.6dB	8.2dB	9.8dB	9.4dB

Assumed parameters; Satellite antenna aperture efficiency: 40%, Earth station antenna aperture efficiency: 50%,
Satellite system noise temperature: 600K, Earth station system noise temperature: 300K
Required Eb/No=5.7dB (at 10^{-4}), Link margin is defined at beam edge where -3dB from beam center.

Size of the Ka-band omnidirectional antenna will be smaller than 1cm. Such an extremely small antenna will enable implementation of extremely small size terminals. Table 1 shows 3.5m and 10m Ka-band large satellite antenna systems examples. Bit rate of the signal is selected as 4.8, 64 kbps and 1.544 Mbps. Receiving and transmitting antenna are assumed to have same diameter. Therefore, multiple beam patterns of up link and down link will be different. As stated in the later section, it is assumed that system design allows different up link and down link beam pattern. And it is also assumed that satellite antenna beam pointing variation can be absorbed in the system design. For beam to beam hand over and different transmit and receive cell pattern, existing cellular systems control technologies can be applied.

In this study, 3.5m antenna is considered as an example which is possible to implement by the present technologies. 10m antenna is considered as an example which can support omnidirectional voice terminal although its implementation is difficult by present technologies. 3.5m antenna has 144km spot beam on the earth. This size of spot beam allows use of 5cm 1W mobile terminal at 64kbps. For the same 3.5m antenna, down link spot beam diameter is 220km. The satellite transmitting power is 1W for the same terminal. As shown in the former section, if satellite antenna diameter is 10m, 4.8kbps /-3dBW omnidirectional mobile terminal can be used with 5dB up link margin. In this case, beam diameter on the earth is only 51km. It will be necessary to implement tremendous number of beams to cover intended service area.

Table 2 shows rough estimate of necessary number of beams to cover Japan and US by 3.5m and 10m satellite antenna. Rough estimate of number of beams is made dividing total area of each country by area of the spot beam. In these examples, use of omnidirectional mobile terminal will not allow 2 degrees satellite spacing as existing fixed satellite service. For this kind of mobile satellite service, use of a specific mobile satellite service band in Ka-band will be necessary.

Table 2 Rough estimate of necessary number of beams

Satellite antenna diameter	Beam coverage diameter	Number of beams	
		Japan	US
3.5 m	144 km	23	591
10 m	51 km	185	4,714

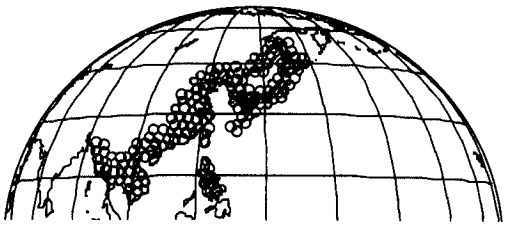


Fig.6 Example beam coverage by 3.5m antenna. The number of beams shown is 170.

POWER SHARING AMONG BEAMS

A large satellite antenna can support small mobile terminals with small transmitting power. However, Ka-band large antenna has extremely small spot beam coverage. Large number of beam is necessary to cover whole intended service area. In the satellite communication system with large number of beams, traffic in each spot is not uniform and also not static. Some beams will cover on the

ocean or on the mountain area where traffic will be very spontaneous.
Thus, it is not efficient to provide a fixed power transmitter for each spot beam. Transmitter for a large number multiple beam system must have capability to share transmitting power among beams. The author proposed "Multiport Amplifier" for this purpose in 1987(1). Multiport Amplifier consists of 2 set of multiport Hybrid and array of amplifiers between them. The concept was successfully applied in several L-band multiple beam mobile satellite systems. However, implementation of the Multiport Amplifier is difficult in Ka-band because Hybrids and cables loss become large. Another example of power sharing transmitter is using an active phased array multiple beam antennas. In case of multiple beam phased array, all beams are transmitted by common active array elements. Transmitting power of each beam is not fixed and can be changed according to the traffic without loss of efficiency. However, using a conventional multiple beam active phased array is difficult in Ka-band because beam forming network with phase shifter for large number of beams is extremely difficult.

KA-BAND MULTIPLE BEAM TRANSMITTER USING EQUAL PHASE SHIFT ACTIVE ARRAY

Figure 7 shows the proposed Ka-band offset reflector antenna with equal phase-shift active array in front of the primary feed horns. Use of a reflector is necessary to make a small spot beam on the earth. Use of direct radiating phased array is impossible because number of elements become too large to make a large aperture corresponding to a small spot beam. An ordinary direct radiating array radiates a plane wave. In the proposed equal phase-shift array, the array radiate a spherical wave similar to the radiation from the primary feed horn. The reflector and the primary feed horn are similar to the ordinary offset multiple beam antennas. The array element consists of a uniform electrical length Ka-band SSPA and rectangular horn at its input and output. There are no phase shifters in the array elements. These active array elements are space fed by primary feed horns. By assembling a large number of small power array elements, a very high power transmission becomes possible. Also, since all beams are transmitted by commonly used array elements, it is unnecessary to assume fixed power in each beam. This configuration support power sharing among beams.
Other features of this configuration are as follows.

By use of a small power light weight SSPA, weight of the whole transmitter will be decreased compared with utilizing Ka-band TWTAs.

It is not necessary to implement redundant SSPA units because failure of several elements only slightly affect total transmitting performances.

Low power simple equal phase-shift active array element is easy to implement in Ka-band.

By adopting power sharing capability among carriers, down link power control can be implemented as a powerful countermeasures to down link rain attenuation, [7].

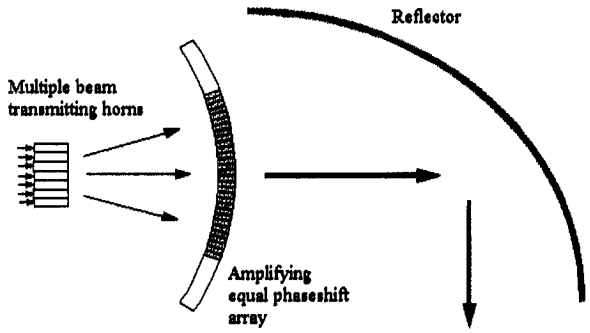


Fig.7 Ka-band multiple beam transmitter using equal phase shift amplifying array

KA-BAND AMPLIFYING ARRAY ELEMENT

In the configuration shown in Fig.7, the array elements are placed at equal distance from the focus of the reflector. The primary feed horns, which are placed on the focal plane of the reflector, point the center of the array cluster. On the other hand, the array element receiving horns point to the focus of the reflector. These arrangements are taken so as to obtain uniform array elements transmitting power distribution for all beams.
Figure 8 shows the concept of the amplifying array element. Received power is converted to a planer strip-line mode, amplified by Ka-band FET, and retransmitted by transmitting horn. Since planer strip-line has no cutoff frequency, its cross section can be reduced compared with the input and output horn. The space under the strip line will be used to dissipate heat and to supply power. The heat emanating from Ka-band FET will be diffused to perimeter and radiated (8).

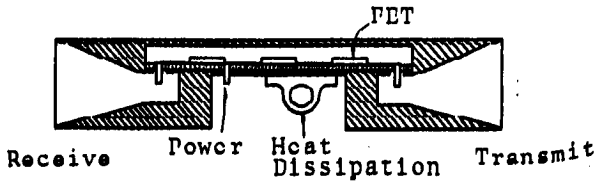


Fig.8 Ka-band amplifying array element concept.

ONBOARD PROCESSING AND ROUTING

If number of beams is hundreds to thousands, it is necessary to demodulate up link signals and to have function of routing signals onboard. If processing is not used, tremendous bandwidth is necessary in feeder link. Recently, large

capacity routing of packet signals become possible by applying ATM technologies. Up link signals are formatted as ATM packet, then sent at 4.8bps. Satellite receives the signal and demodulates using DSP technologies. Demodulated base band signals are routed by ATM technologies. Routing to feeder link or directly to down link spot beam is possible by the address in the ATM packet. Fig.9 shows configuration of onboard processing for hundreds to thousand beam satellite. Transmitting beams are amplified by common amplifying array.

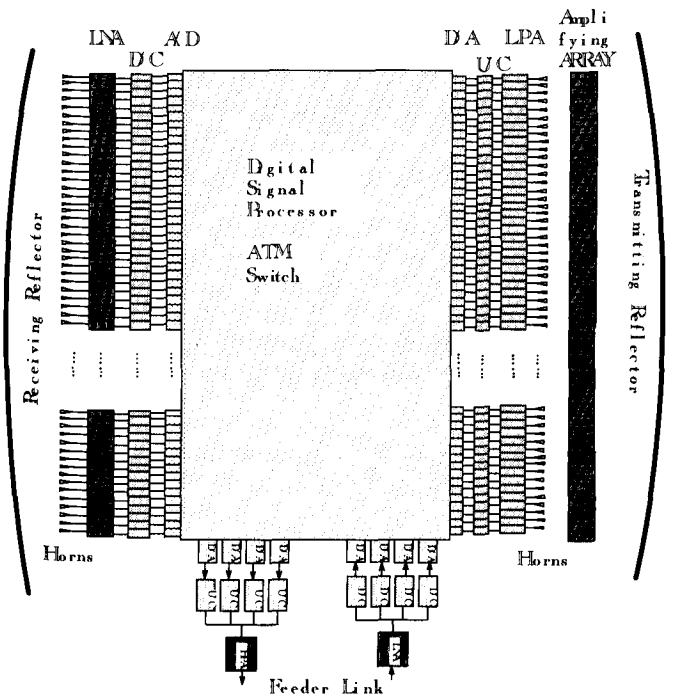


Fig.9 Ka-band hundreds to thousands beam mobile satellite concept.

APPLYING CELLULAR COTROL TECHNOLOGIES

In the proposed system, up link and down link beam pattern is different. Therefore, it is necessary to find down link beam where the mobile terminal exists. Finding down link beam is possible sending a control signal to all down link beams and receiving response from the mobile terminal. Also, beam to beam hand over is necessary by moving mobile terminal or pointing variation of the satellite. These locating and hand over technologies can be adapted from the existing cellular mobile communications technologies.

CONCLUSION

Ka-band multiple beam mobile satellite system with a large satellite antenna is studied. It was shown that, in principle, using 10m satellite antenna, omnidirectional mobile voice terminal could be implemented. It was also shown that 64kbps and 1.544Mbps multimedia mobile communications could be implemented by extremely small terminals. In these systems, beam coverage also became extremely small necessitating large number of beams. In order to implement large number multiple beam in Ka-band, offset reflector antenna fed through equal phase shift amplifying array is proposed. As an example, 3.5m and 10m antenna systems are studied. By uniform distribution of array element excitation, power sharing among beams will be attained.

REFERENCES

[1] S.Egami and M.Kawai, "An Adaptive Multiple Beam System Concept", IEEE Journal on Selected Areas in Communications, Vol.SAC-5, No.4, pp630-636, May 1987.

[2] D.J.Whalen, etal. "The American Mobile satellite Corporation Space Segment" AIAA 14th International Communications Satellite Systems Conference, pp.394-404, March 1992.

[3] S.Egami, "Application of Multi-port Amplifier to Personal Satellite Communications", International Mobile Satellite Conference, Ottawa, pp.67-72, July 1995.

[4] S.Egami and M.Kawai, "An Adaptive Multiple Beam Transmitter for Satellite Communications", IEEE Trans. on Aerospace and Electronic Systems, Vol.AES-23, No.1, pp11-16, Jan.1987.

[5] S.Egami "Satellite Link Requirement in Personal Satellite Communications", Space Communications, IOS Press, No.12, pp.105-114, Dec.1994.

[6] E. H. Kopp "Handheld Telephony from Geosynchronous Orbit", 17th International Communications Satellite Systems Conference, 98-1396, Feb. 1998.

[7] S.Egami "Closed Loop Satellite Access Power Control System for K-band Satellite Communications", IEEE Trans. on Aerospace and Electronic Systems, Vol.AES-19, No.4, pp.577-584, July 1982.

[8] G.Anzic et al., "Microwave Monolithic Integrated Circuit Development for Future Space Born Phased Array Antennas", AIAA International Communications Satellite Systems Conference, pp43-53, 1984.

Geo-Mobile Satellite System Air Interface Overview and Performance

Stephanie Demers, Chi-Jiun Su, Chandra Joshi, Xiaoping He, Anthony Noerpel, and Dave Roos
Hughes Network Systems
11717 Exploration Lane, Germantown, MD 20876
Email: sdemers@hns.com, csu@hns.com, cjoshi@hns.com, xhe@hns.com, anoerpel@hns.com, droos@hns.com

ABSTRACT

An air interface has been developed for a mobile satellite system planned for deployment in the year 2000. The interface definition draws its heritage from the internationally accepted and successful GSM terrestrial cellular specification, yet it incorporates many important features that optimize it to the peculiarities of the satellite channel. Some unique capabilities are provided which do not exist in the terrestrial system that are very important in the satellite system, such as integrated position-based services and single-hop terminal to terminal calling. The air interface is being standardized jointly by ETSI and TIA.

INTRODUCTION

Hughes Electronics has developed [1] a mobile satellite system that uses a geosynchronous satellite to provide continental coverage at low cost. The system consists of:

- one or more satellites in orbit nominally above the areas served by the satellites,
- one or more gateway stations providing interconnection with public land mobile networks and the public switched telephone network, and
- many user terminals which may be of many types, including handheld cellular-like phones, vehicular-mounted mobile phones, and fixed telephones.

The services provided by the system include the standard services offered by GSM as well as a few not provided by GSM such as the ability to make single-hop terminal to terminal calls. Many characteristics of the satellite channel have been taken into account in the design of the air interface.

One of the key features of the air interface includes its strong resemblance to terrestrial GSM at the upper protocol layers. This allows the integration of standard GSM services into the system by using as much as possible off-the-shelf components such as the mobile switching center (MSC).

The air interface is currently being developed into an international standard jointly by TIA and ETSI committees [2].

IDLE MODE BEHAVIOR

Introduction

A user terminal is in idle mode if it is not on a dedicated connection or in the process of establishing a dedicated connection. The GEM idle mode has four processes as shown in Figure 1: Spot Beam Selection and Re-selection, Public Land Mobile Network (PLMN) Selection, Position Determination and Location Updating (LU).

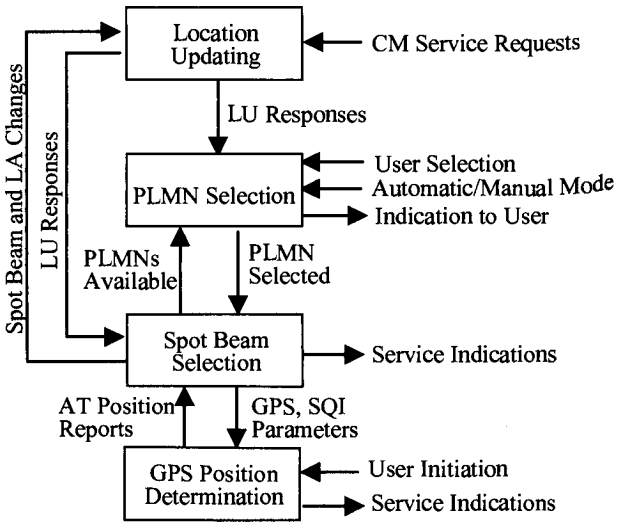


Figure 1. GEM Four Idle Mode Processes

GEM location updating and higher layer processes are similar to GSM. Nevertheless, GEM and GSM idle mode behaviors differ in several important respects.

A salient difference between a GSM cell and a GEM spot beam is its size. A GSM cell, being at most a few kilometers in radius, is wholly contained within a location area (LA), connected to a single MSC and associated with

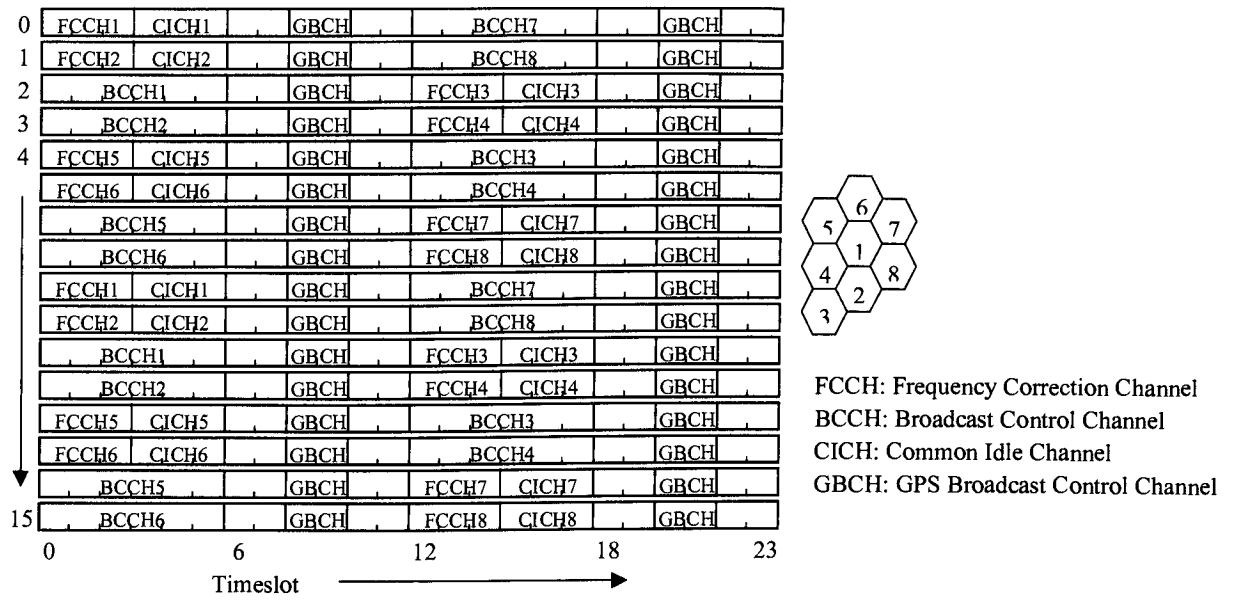


Figure 2. Time Orthogonality of Control Channels in a Multiframe

a single PLMN. A GEM spot beam, being approximately 100 km in radius, can subtend several countries, and therefore can be associated with several LAs and possibly PLMNs and be able to connect terminals to several national gateway stations, MSCs and visitor locations registors (VLRs).

Therefore position determination is an important feature of the GEM idle mode and is well integrated into the air interface protocol and system design. Thus the user terminal in a GEM system is required to perform Global Positioning System (GPS) position determination and report its position to the GEM network. In contrast, the user terminal position in a GSM system is only indicated by the accuracy of the cell selection.

Another difference is the need to conserve satellite power by ensuring that all terminals select the best spot beam from a signal strength perspective. Thus, in a GEM system, only the strongest or next to the strongest spot beam should be camped-on and the PLMN should be selected from the Broadcast Control Channels (BCCHs) of these spot beams. By contrast, in GSM, PLMN selection has priority over cell selection and any suitable cell of a preferred PLMN should be selected even if there are other stronger cells belonging to non-preferred PLMNs.

An advantage that a geo-mobile satellite system has over a terrestrial GSM system is that all spot beam broadcast control channels have the same source, the satellite, and therefore, at the terminal, all BCCHs traverse an identical propagation path with the same fading characteristics. Furthermore, time synchronicity is easily achieved.

The specific details of the GEM spot beam selection and re-selection procedure, position determination procedure, PLMN selection procedure and Location updating will be described in the following sections.

Spot Beam Selection and Re-Selection

The spot beam selection procedure ensures that a user terminal, when operating in a GEM system, nearly always selects the best spot beam. An accurate spot beam selection is crucial in a GEM system to minimize satellite power cost and timeslot assignment blockage, and to limit the number of spot beam re-selection procedures. Fast spot beam selection is desirable to provide a transparent procedure to the user.

To achieve a fast and accurate spot beam selection, several concepts were developed in the GEM system. For fast camp-on time, the network broadcasts system information in each beam that contains a list of neighbors BCCH frequency and timing as well as a complete list of all carrier frequencies used in the GEM system. The user terminal stores this information in non-volatile memory. When powering-on again, the terminal can first look at these stored lists of frequencies to speed up frequency and timing acquisition and spot beam selection. BCCHs from all seven spot beams are also time orthogonal and broadcast every half a multiframe so that the user terminal can measure them quickly and efficiently as shown in Figure 2.

Position Determination

Both standalone GPS and GEM assisted fast acquisition are supported in the GEM system. Fast GPS acquisition

considerably reduces the acquisition time by using a rough position estimate and a GPS Broadcast Control Channel (GBCH) providing satellite parameters for at least four GPS satellites in view to the terminal and GPS timing estimate. This rough position estimate is calculated by the user terminal based on relative power measurements from surrounding beams and satellite and beam center positions broadcast in the BCCH system information.

This position estimate combined with the GPS position, if available, is also used by the user terminal to validate the spot beam selection procedure once a spot beam has been selected and will trigger a new spot beam selection or its completion. The user terminal will also report its GPS position to the network in the Channel Request message if position information reporting is required. It is possible that the user terminal selects an incorrect beam due to measurement errors, for example. The network detects this error based on the reported GPS position, and redirects the user terminal to the correct beam. Therefore, the use of GPS by the user terminal and the network will insure near-perfect spot beam selection.

The position determination procedure will help determine the difference between the round-trip delay from the AT and the satellite, and the round-trip delay reported in the system information to the center of the selected spot beam. In the normal operating condition, a Random Access Channel (RACH) burst is assumed to arrive in a deterministic time window at the satellite. However, the user terminal is only aware of the round-trip delay from the center of the selected spot beam, broadcast by the network. As the user moves away from the spot beam center, its actual timing deviates from this known round-trip delay. This differential delay is consequently unknown to user terminal and can be as long as 6 milliseconds. As a result, the RACH burst could fall partially outside the RACH window for some beams. Measuring this differential delay accurately will allow the GEM system to operate with a small RACH window, conversely a large channel request message burst format.

PLMN Selection

Once spot beam selection is completed and suitable spot beams are identified, the user terminal determines all available PLMNs from the BCCHs of the suitable spot beams. The system information on a BCCH identifies all other BCCHs that are present in the same spot beam and all BCCHs of overlapping spot beams from other satellites within the same GEM system. The normal GSM PLMN selection procedure is then applied on those available PLMNs.

Location Updating

On selection of a PLMN, the mobility management (MM) layer of the AT initiates the Location Update procedure by issuing a radio resource (RR) Establishment Request to the RR layer. The RR sublayer of the AT then initiates an Immediate Assignment Procedure and sends a channel request message on a RACH containing the user terminal's GPS position. Based upon the reported GPS position and/or other information, the network may determine various position-based access errors or enforce position-based access restrictions by sending error causes in Immediate Assignment Rejection messages regarding spot beam and/or LAI selection. As shown in Figure 3, a spot beam may contain multiple LAIs and PLMNs. A user terminal in beam 2 in country 1, for example, may only be allowed to access the gateway station 2 (GS2) and not GS3. A rejection message will be sent by the network for a user terminal with a position not covered by the network, for a user terminal with an invalid position in the selected LAI, or for a user terminal selecting a spot beam with an invalid position. The network may also send a rejection message to direct the user terminal to register via a different GEM satellite.

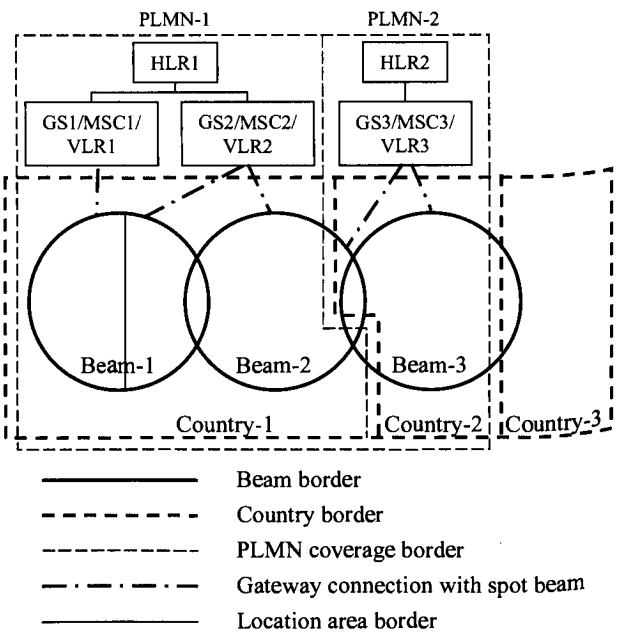


Figure 3. Various Reasons to Trigger Location Update

SINGLE HOP TERMINAL TO TERMINAL CALLING

If a call is established between two mobile terminals using two satellite hops, the delay from one user to the other will be approximately 540 milliseconds. This propagation delay may be unacceptable. To reduce this delay, a direct link can be provided between the two terminals through the satellite by establishing a connection with only a single

satellite hop. One of the innovations in the GEM system is the ability to establish single-hop Terminal-to-Terminal (TtT) calls.

The GEM TtT call procedure is based on GSM defined protocols with minimum changes required for the satellite environment. The scheme also supports GSM features such as authentication, encryption, user identity protection, roaming, call interception, and DTMF signaling.

In addition to the normal call setup procedures such as RR, MM and Call Management (CM) procedures, the crucial steps involved in the establishment of a TtT call are the connection of a TtT link at the satellite, the assignment of the link to the originating terminal and the assignment of the link to the terminating terminal. The GEM satellite is equipped with an on-board switch, which can cross-connect a direct link between two mobile terminals upon the request of a GS.

After the GS receives the call request from the originating mobile user terminal, it performs the call identification procedure and identifies the call as a TtT call. During the following MM and CM procedure, the MSC is informed of the arrival of a TtT call. Then, the GS initiates the establishment of the direct TtT link connection at the satellite. Upon the positive response from the satellite, the direct TtT link is first assigned to the originating terminal and then to the terminating terminal. After the verification of the direct TtT link between two terminals, a single hop voice connection is established between the two terminals. Our step-by-step establishment of the TtT link guarantees the reliable direct communication between two terminals with the minimal call setup delay and the optimal use of system resources. The configuration of radio resource during a TtT call is depicted in Figure 4.

During a single hop TtT call, since two terminals are communicating with each other through a cross-connected link at the satellite, a separate channel is required for sending signaling messages from the GS to each terminal. The channel is known as Terminal-to-Terminal Control Channel (TTCH). As the flow of signaling messages from the GS to a terminal is quite light during a TtT call, it is designed to be a point-to-multi-point control channel shared by several TtT calls for the efficient use of radio resource.

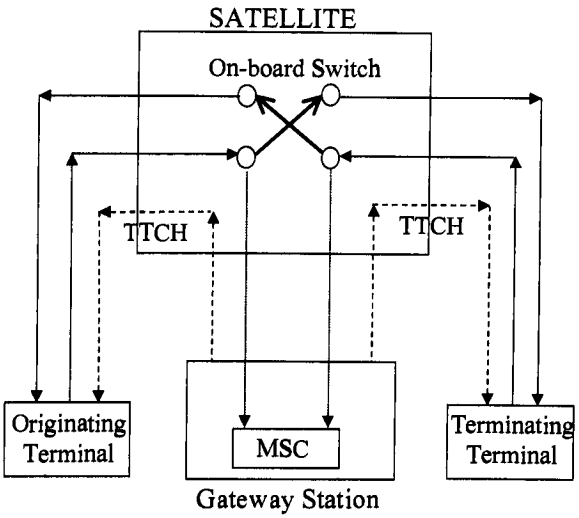


Figure 4. Radio Resource Configuration during a TtT Call

In a single TtT call, the ciphered communications begin first with both terminals talking to the GS in normal cipher mode, each with a separate cipher key. Then, after an encryption algorithm common to both terminals is selected, and a common cipher key is generated at the GS and transferred to the terminals, the mode is changed to a cipher mode with a common cipher algorithm and a common cipher key. In a single hop TtT call, the ciphered communication exists not only between the two terminals but also between each terminal and the GS for the purpose of interception. Therefore, in order to perform encryption and decryption properly with pseudo-random sequences at each terminal and the corresponding GS, terminals and the GS are instructed to act as a "mobile" or a "network" appropriately before the beginning of the cipher mode with a common cipher key.

One of the novelties in TtT signaling is the use of Service Access Point Identifier (SAPI) = 2. Since each terminal is communicating with the other terminal as well as the GS at the same time, SAPI=2 is used for signaling between the terminals so as to coexist with the signaling flow with SAPI=0 between a terminal and the GS. Link establishment with SAPI=2 between two terminals after the TtT link is assigned to both terminals, verify not only the physical connectivity through the satellite but also cipher synchronization between the two terminals. As a result, a reliable communication between two terminals is established.

Unlike GSM systems, symmetric power control is employed in a single hop TtT call with the regular exchange of power control message between two terminals. Two terminals act as peer entities in the control operations. Only for the synchronization of power control

message, one terminal is designated as the master and the other as the slave.

Timing and frequency synchronization during a TtT call involves a number of phases: synchronization at initial access, synchronization before a TtT link is assigned, synchronization during the transition to a TtT link and synchronization of the TtT link. Before a TtT link is assigned, in the forward link from a GS to a terminal, synchronization is performed at the terminal by tracking signals received from the GS. In the return link, any transmission timing drift is corrected by timing/frequency correction messages received from the GS, which keeps track of signal arrivals transmitted from the terminal. After the TtT link is assigned, in the forward link, the AT maintains an independent tracking loop for signals from the TtT link and TTCH channel. In the return link, TTCH-based observation is used as the reference for transmission. The network also monitors signal arrivals transmitted from each terminal and sends timing/frequency correction messages to the terminals on the TTCH to correct the timing and frequency drift.

PERFORMANCE

The GEM Common Air Interface (CAI) defines the interface between the access terminal and GEM satellite system. The GEM CAI reuses the GSM higher layer protocols to ease the integration with the terrestrial cellular system and reduce the development cost. The RR layer, the data link layer (DLL) and the physical layer protocols, however, are all optimized to the satellite link. A few important modifications including RACH design, Extended Channel Assignment signaling procedure, two-outstanding-L3-message approach, multiple set asynchronous balanced mode (SABM) and GEM DLL protocol are introduced in this chapter to exhibit the successful solution for mobile satellite communications.

GSM RACH only conveys 8 bits of information. Its purpose is to speed up the call setup. This design is appropriate in terrestrial cellular systems since a mobile user can only access a visible ground station. In the satellite environment, however, a mobile user can contact any GS that is in the satellite coverage. Due to the broad coverage of a GEO satellite, the GS selected by a mobile user may not provide the optimum routing. In order to reduce the call setup delay, an optimum GS should be selected at early stage of call setup. Therefore, GEM RACH is designed to carry all the necessary information to perform optimum routing. The RACH message includes random reference, MSC id, called number, establishment cause, number plan identification, retry counter, AT power class, HPLMN ID, etc, totaling 139 bits. This long RACH design is vulnerable to the noisy channel. Therefore, the success rate of access will be reduced. To solve this problem, RACH message is divided into two categories: class 1 and class2. Class 1 information (20 bits) is better protected than class 2. With a good visibility to the satellite, both class 1 and class 2 bits may be successfully received and an optimum routing decision can then be made. If a mobile user sends a call request in a location with bad visibility, only class 1 RACH is received. GS will immediately assign a channel to the mobile to get more setup information and stop the transmission on RACH. This signaling procedure is called extended channel assignment. The GEM RACH design and extended channel assignment together provide a solution to reduce call setup delay in the satellite environment.

GEM has a different physical layer design than GSM. In GSM, a signaling message has 456 coded bits (160 information bits) being interleaved into eight bursts. In GEM, a signaling message has 384 coded bits (56 information bits) being spread into four bursts. The coding scheme adopted for GEM FACCH enhances the channel robustness. The reason that we have the luxury to use the robust signaling channel is the GEM DLL design. GEM DLL has a window size of 16 in contrast to the window size of 1 in the GSM system. If the length of a L3 message is longer than 7 bytes, it will be segmented and a maximum of 16 segments (112 bytes) are allowed unacknowledged.

To fully utilize the big window size, an approach called two-outstanding-L3-message is proposed. In GSM, only one L3 message is allowed at a time to avoid the ambiguity in duplication detection when channel re-assignment happens. In most of the cases the L3 message may be shorter than the N201*k (N201 is the size of information in the frame and k is the window size). So the channel is under-utilized and the time needed for exchanging the call setup related message is high. Since the L3 header where duplication detection bit (N(SD)) locates is untouchable, this approach exploits the fact that only the MM and CM sub-layer messages have the N(SD) associated with them and RR sub-layer messages do not have any N(SD) associated with them. A SubLayer Flag is added in primitives from L3 to L2 to indicate to which sub-layer the message belongs. With this modification, a RR message can be transmitted with a MM or CM message without waiting for the acknowledgement. An example of this utilization is the mobile originated (MO) call setup where a setup message (CM layer) can be sent immediately after cipher mode complete message (RR layer).

Another modification in DLL protocol is called multiple SABM. With this approach, AT continuously sends multiple SABMs (up to eight) without waiting for a UA response or T200 time out. The T200 timer is set after the last SABM is sent. If there is no UA received until T200 time out, the link is released. This approach will reduce the

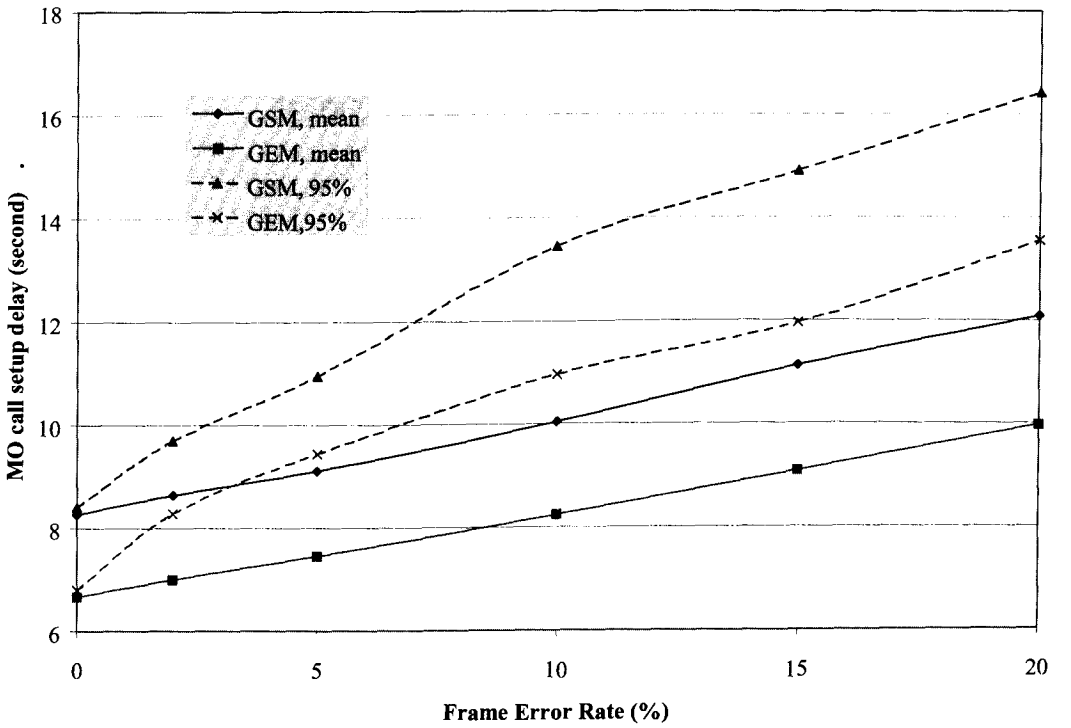


Figure 5. MO call setup delay vs. FER

call setup delay in error-prone channel and is also intended to speed up and maintain the synchronization.

There are additional modifications in GEM DLL, like group-reject (GREJ) in contrast to select-reject in the GSM system. The idea of GREJ is to acknowledge the received out-of-sequence frame and trigger the transmission of non-received frames. GREJ can detect a single error frame as well as several error frames in a row. Multiple GREJ S-frames are allowed to transmit. With GREJ, errors can be recovered quicker and redundant re-transmission is prevented.

The simulation is conducted using OPNET simulation tool. An OPNET model of the GEM CAI protocol is developed and extensive simulations are performed to investigate the performance of the proposed GEM CAI signaling protocols. The simulation validates the proposed protocols and also shows that the call setup signaling performance are satisfactory for the GEM satellite system. The MO call setup delay in comparing GEM DLL and GSM DLL protocols is shown as an example of performance investigation (Figure 5). The call setup time is the interval from transmission of the RACH message from the AT and receipt of call proceeding message from the GS by the AT. The performance shows approximately two seconds delay improvement using GEM DLL design. In typical channel condition (5% FER), 95% call setup delay is less than 9.5 second. The multiple SABM approach provides 0.5 second delay reduction for an error-prone channel.

CONCLUSION

A mobile satellite system has been designed with the benefits of the terrestrial GSM network, but optimized for the satellite channel. The air interface defines enhancements to GSM to allow effective use of the satellite. These enhancements include the ability to mitigate the long satellite delay inherent in terminal to terminal calls by providing single-hop terminal to terminal calling. Another enhancement is that position based services have been integrated into the air interface to support billing functions and emergency call routing. The air interface is being standardized jointly by ETSI and TIA as GMR-1 (Geo Mobile Radio-1).

REFERENCES

[1] Alexovich, J., et al, "The Hughes Geo-Mobile Satellite System", Proceedings of the International Mobile Satellite Conference, June 1997, pp. 159-165.

[2] ETSI Web Site, http://www.etsi.org/
ETSI GMR-1 Specifications: docbox.etsi.org/tech-org/ses/Document/ses/GEO_Mobile_Radio_interface_(GMR)/GMR_Specifications/GMR-1 Specs/

[3] TIA Web Site, http://www.tiaonline.org/
TIA GMR-1 Specifications: http://ftp.tiaonline.org/SAT-DIV/CAI_GMPCS

Timeslot Assignment Algorithm for Geo Mobile (GEM™) Satellite System

Wei Zhao, Steven P. Arnold
Hughes Network Systems
11717 Exploration Lane
Germantown, Maryland 20874, USA
wzhao@hns.com
sparnold@hns.com

ABSTRACT

As part of the traffic resource management technique in a time division multiple access (TDMA) based mobile satellite communication system, a mobile link carrier grouping algorithm provides an efficient way to reduce call blocking caused by mobile user delay variation within a spotbeam. The major idea is to divide the spotbeams of a satellite into a number of single offset coverage zones. User traffic generated from each zone is carried by a particular group of carriers. Since there is limited propagation delay variation within each offset coverage zone, a single transmit (Tx)/receive (Rx) burst offset on the satellite can be shared by all mobile users. Therefore, the call blocking probability can be greatly reduced.

INTRODUCTION

In a TDMA based mobile-satellite system, the timeslot assignment algorithm is part of the procedure to be performed at the Gateway Station (GS) during a call setup. This algorithm is fundamental in achieving high efficiency of frequency resource utilization and reducing the overall call blocking rate.

The main functionality of the Timeslot Assignment Algorithm in the GS is to assign a pair of channels on both forward and return links based on information received from the Access Terminal (AT) and information measured from the GS. These include required data rate, terminal type, call type (Terminal-to-Gateway or Terminal-to-Terminal), user spotbeam identity and user-satellite propagation delay, etc. A typical procedure in the call setup is shown in Figure 1.

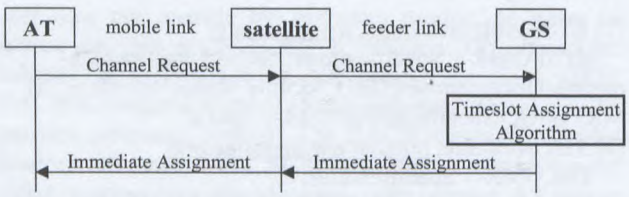


Figure 1 Procedure of call assignment

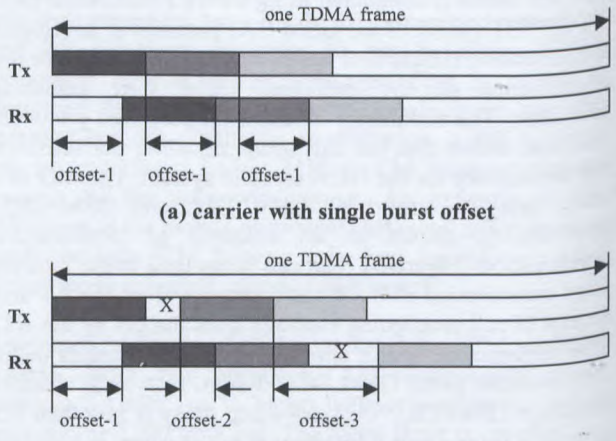
To reduce the overall call blocking rate, it is preferable that all calls assigned to the same carrier also use the same Tx/Rx bursts offset on the satellite. Thus calls can be packed close to each other in time, and high resource utilization efficiency can be achieved.

However, the particular features of a mobile satellite system make it difficult to do so: (1) For economic reasons, most of the ATs are equipped with a transmit/receive switch which impose constraints on the guard time to switch between transmission and receive frequencies; (2) Delay spread within a spotbeam can be very large such that a satellite burst offset which is appropriate to a user located on one edge of a spotbeam may not apply to another user located on the other edge of the spotbeam. As a result, the timeslot assignment algorithm has to apply multiple satellite offsets to spotbeams with large delay variation, and this is a major contribution to additional call blocking.

CARRIER GROUPING

Carrier Grouping Methodology

One of the keys to achieving high resource utilization efficiency is to use the same satellite burst offset among all calls using the same carrier. The burst offset is defined to be the timing difference between the beginning of the Tx burst and the beginning of the Rx burst. This is shown in the first graph of **Figure 2**.



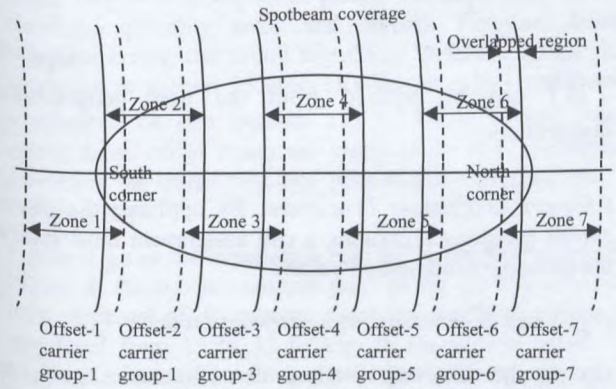
(a) carrier with single burst offset

(b) carrier with multiple burst offsets

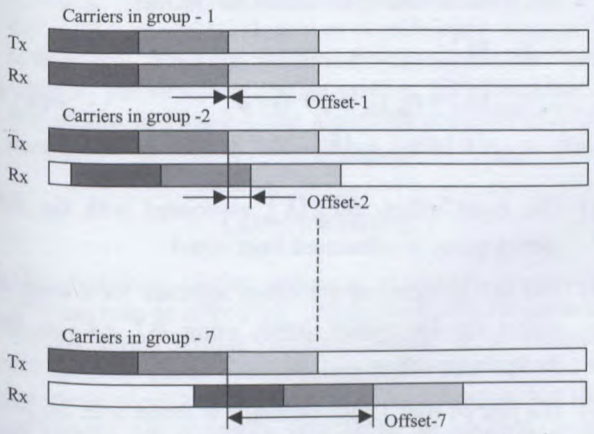
Figure 2 Burst timing on the satellite

Due to the large delay variation within the same spotbeam, users from different positions within the spotbeam must use different offsets to meet the non-overlapping requirement at the terminal. If these calls are packed into the same carrier, then the carrier cannot be fully utilized, resulting in increased call blocking probability. This is shown in the second graph of Figure 2, where 'X' stands for unused slots which cannot be assigned to any calls.

This paper addresses a carrier grouping algorithm which reduces the call blocking caused by the existence of multiple offsets in a spotbeam. To perform the carrier grouping, a spotbeam is divided into a number of single offset zones. Within each zone, a single offset is defined and this offset can be applied to all users within this zone. Meanwhile, the traffic resource pool is divided into the same number of groups. One group is associated with one single offset zone, and therefore one unique burst offset. This is shown in Figure 3. In the first graph of Figure 3, adjacent single offset zones are overlapped by a certain distance. This is to avoid an abrupt offset change due to AT-satellite relative motion during a call.



(a) offset zone arrangement within a spotbeam



(b) carriers in one group use only one burst offset

Figure 3 Carrier grouping methodology

The total number of single offset zones (for example, 7 zones in Figure 3) is determined by several factors: spotbeam size, frame structure, traffic burst length and the guard time required by the Access Terminal. This is also the number of carrier groups required for the spotbeam. Within each carrier group, the number of carriers can be derived from the mobile user geographical distribution.

To perform a call assignment, a user's propagation delay must be measured by the network at the time of call setup. Then a carrier group is selected by comparing this measured delay with the delay boundaries of each single offset zone. The call can then be assigned to a carrier in the selected carrier group and the offset associated with this carrier group must be applied to this call.

Carrier Grouping Algorithm

Operating Environment: Consider a TDMA based mobile satellite system with the following features:

(1) All timing in the system is referenced at the satellite, and both forward/return link frames are time aligned on the satellite reference point.

(2) Multiple spotbeam operation.

(3) Frame and traffic burst structures are defined as follows: the frame duration is T_f (ms), the timeslot duration is T_s (ms), and each frame has $N_s = T_f / T_s$ timeslots. Each traffic burst takes K timeslots, so that the burst length is $T_{ch} = K \cdot T_s$ (ms), where K is a constant number for all user traffic in the system.

(4) The timing constraints are outlined as follows: at the AT, the transmit/receive bursts can not overlap in time. In addition, a guard time of T_{gt} (ms) must be allowed for the AT to switch between transmit and receive frequencies. A traffic burst can wrap across the frame boundaries in both forward and return links.

(5) The total number of carriers dedicated to each spotbeam is N_c.

Delay Range of a Single Offset Zone: Based on Appendix-A, the maximum round trip delay variation R_z for each single offset zone can be given as

$$R_z = T_f - 2(T_{ch} + T_{gt}) \qquad \textbf{equ - 1}$$

where the unit of R_z is in ms.

Number of Offset Zones Required in a Given Spotbeam: Assume the offset zones are arranged such that one offset zone is centered over spotbeam's mid-delay point, and all other zones are symmetrically placed on both sides of the middle zone, as shown in Figure 3. If the range of propagation delay in the considered spotbeam is $t_d \in [t_{d\min}, t_{d\max}]$ (ms), and N_{os} is the total number of required offsets for that spotbeam, then

$$N_{os} = \begin{cases} 1 & if \quad 2(t_{d\max} - t_{d\min}) \le R_z \\ 1 + 2 \cdot ceil\left[\dfrac{(t_{d\max} - t_{d\min}) - R_z/2}{R_z - \Delta}\right] & if \quad 2(t_{d\max} - t_{d\min}) > R_z \end{cases}$$

equ - 2

where Δ (in ms) is the overlap between two adjacent single offset zones, measured in terms of round trip delay variation.

Offset Value Associated With Each Single Offset Zone: The value of the burst offset associated to a single offset zone is a function of the zone center propagation delay. Assuming the delay boundaries of the *kth* single offset zone are $[t_{dz\min}(k), t_{dz\max}(k)]$ and $t_{dz0}(k)$ is the delay at the zone center, then

$$t_{dz0}(k) = \frac{t_{dz\min}(k) + t_{dz\max}(k)}{2} \qquad \textbf{equ - 3}$$

From Appendix-B, the offset value $\Delta T_{os}(k)$ associated with the *kth* single offset zone for a geo-stationary mobile satellite system can be given as

$$\Delta T_{os}(k) = round\left[\frac{2t_{dz0}(k) - 6.5T_f}{T_s}\right] \qquad \textbf{equ - 4}$$
$$-\frac{N_{os}-1}{2} \le k \le +\frac{N_{os}-1}{2}$$

Single Offset Zone Delay Boundaries: Since the middle zone is centered over the spotbeam's mid-delay point, then the propagation delay at the center of the *kth* offset zone $t_{dz0}(k)$ can be calculated as

$$t_{dzo}(k) = t_{d0} + \frac{k(R_z - \Delta)}{2}$$
$$k = -\frac{N_{os}-1}{2}, \cdots, 0, \cdots, +\frac{N_{os}-1}{2} \qquad \textbf{equ - 5}$$

where t_{d0} is the propagation delay at the spotbeam mid-delay point.

Then delay boundaries of the *kth* single offset zone is $[t_{dz\min}(k), t_{dz\max}(k)]$ and

$$t_{dz\min}(k) = t_{dz0}(k) - R_z/4$$
$$t_{dz\max}(k) = t_{dz0}(k) + R_z/4 \qquad \textbf{equ - 6}$$

By comparing a user's propagation delay with the delay boundaries of each offset zone, an appropriate burst offset can be decided for this user.

Carrier Grouping: The carrier grouping algorithm takes two factors into consideration:

(1) Total number of carrier groups: when each carrier group is matched to one single offset zone, the number of carrier groups is equal to N_{os}, which is the total number of required offsets in the considered spotbeam.

(2) Total number of carriers in each carrier group: this number is proportional to the number of users located in the corresponding offset zone.

If $P(k)$ is the probability that a given call is generated from the *kth* offset zone, then the number of carriers within the *kth* carrier group is $n(k) = P(k) \cdot N_c$, where N_c is the total number of carriers in a spotbeam and is defined as

$$N_c = \sum_k n(k), \; k \in \left[-\frac{N_{os}-1}{2}, +\frac{N_{os}-1}{2}\right] \qquad \textbf{equ - 7}$$

$P(k)$ can be derived from the user geographical distribution.

Channel Assignment Procedure: By applying the above carrier grouping algorithm, a call assignment must follow the procedures addressed below:

(1) A call is generated with propagation delay t_d.

(2) At the Gateway Station, the *kth* carrier group is considered to be the preferred carrier group for this call if the following condition can be met

$$t_{\min}(k) \le t_d \le t_{\max}(k)$$
$$t_{\min}(k) = t_{dz\min}(k) + \Delta/2 \qquad \textbf{equ - 8}$$
$$t_{\max}(k) = t_{dz\max}(k) - \Delta/2$$

(3) The burst offset $\Delta T_{os}(k)$ associated with the *kth* carrier group is calculated from equ-4.

(4) The call assignment algorithm searches for a channel within the *kth* carrier group, using $\Delta T_{os}(k)$ as the Tx/Rx burst offset.

(5) If a pair of free Tx/Rx channels is found with the pre-defined burst offset, the channel is assigned to this call.

(6) If the search failed in the preferred carrier group, the call assignment algorithm searches for a channel within the entire resource pool available to the spotbeam. In that case, Tx/Rx burst timing relationship must be validated to avoid burst collisions at the Access Terminal.

RESULTS AND ANALYSIS

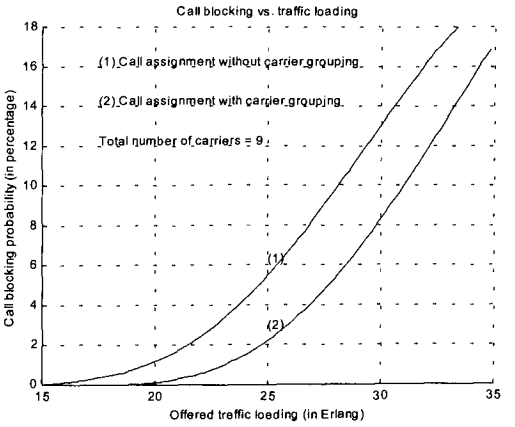
A simulation is performed for one spotbeam in a geo-stationary mobile satellite system. For a given traffic resource pool, call blocking probabilities are simulated with two different call assignment scenarios: with and without carrier grouping algorithm. Conditions applied to the simulation are given below:

The frame duration is $T_f = 40ms$, the timeslot duration is $T_s = 1.67ms$, and the number of timeslots in one frame is 24. Each traffic burst occupies 6 contiguous timeslots, so the maximum number of traffic channels contained in each carrier is 4. The Access Terminal is assumed to have a half-diplexer, where the guard time required between Tx and Rx bursts is $T_{gt} = 1.5T_s = 2.5ms$. The spotbeam has an angle of 0.7 degree as seen from the satellite. With a 15 degree spotbeam elevation angle, a 5.3 degree satellite inclination angle and 50% spotbeam coverage extension (due to the spotbeam pointing error and Access Terminal beam selection error), the round trip delay variation across the spotbeam is around 25.0ms. If adjacent offset zones are overlapped by one timeslot ($\Delta = 1.67ms$), then from equ-2, three offset zones are required for this spotbeam. Therefore the traffic resource pool shall be divided into 3 carrier groups.

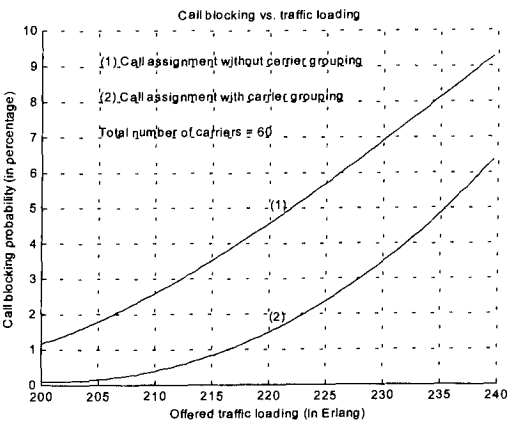
Figure 4 gives the simulation results. In the first graph of Figure 4, the traffic resource pool in the spotbeam has a total number of 9 carriers, and the traffic loading is simulated from 15 to 35 Erlang. In the second graph of Figure 4, the traffic resource pool in the spotbeam has a total number of 60 carriers, and the traffic loading is simulated from 200 to 240 Erlang. For each case, adopting the carrier grouping algorithm significantly reduces the overall call blocking probability throughout the entire range of the simulated traffic loading. By comparing the two graphs in Figure 4, the algorithm provides additional benefits for spotbeams with a larger traffic resource pool.

CONCLUSIONS

The described carrier grouping algorithm provides an efficient way to effectively utilize the traffic resource in a TDMA based mobile satellite system. It is especially effective for spotbeams with large spotbeam size and low spotbeam elevation angle, since delay spread within those spotbeams are generally very large. In addition, a geo-stationary satellite constellation is considered to be the best candidate, since in such a system, the burst offset is not subject to significant change during a call. Some of the concepts presented here are subjects of patent applications.



(a) with 9 carriers



(b) with 60 carriers

Figure 4 Simulation results of call blocking probability

Appendix-A: Delay range of a single offset zone

This analysis takes Figure 5 as reference. In Figure 5, the first two rows are downlink and uplink burst timing on the satellite. The last two rows are downlink and uplink burst timing at the Access Terminal. k_d and k_u are mobile downlink and uplink burst positions. They all range from 0 to $N_s - 1$ and t_d is the user propagation delay.

Assume the starting frame number is *n*. For a given Access Terminal position, the propagation delay is t_d. If the Rx/Tx burst offset at the satellite is $\Delta t_{os} = \Delta T_{os} \cdot T_s$, where ΔT_{os} is the offset measured in timeslots, and T_s is the timeslot duration measured in ms, then the start time t_{tm1} of the transmission burst at the Access Terminal can be given by

$$t_{tm1} = (m-n)T_f + k_d T_s - t_d + \Delta t_{os} \qquad \textbf{equ - 9}$$

Therefore the propagation delay t_d is a function of the transmission time t_{tm1}:

$$t_d = (m-n)T_f + k_d T_s + \Delta t_{os} - t_{tm1} \qquad \textbf{equ - 10}$$

In order to meet the Access Terminal guard time requirement, the Tx burst at the terminal must leave enough guard time T_{gt} from the Rx bursts on both sides of the Tx burst. Assuming a single offset on the satellite can support propagation delay in the range of $[t_{dz\min}, t_{dz\max}]$, then the range of t_{tm1} can be given by

$$\begin{aligned} t_{tm1\min} &\le t_{tm1} \le t_{tm1\max} \\ t_{tm1\min} &= t_{dz\max} + k_d T_s + T_{ch} + T_{gt} \\ t_{tm1\max} &= t_{dz\min} + T_f + k_d T_s - T_{ch} - T_{gt} \end{aligned} \qquad \textbf{equ - 11}$$

where T_{ch} and T_{gt} are the traffic burst duration and guard time duration, both measured in ms.

Since

$$\begin{aligned} t_{dz\max} &= (m-n)T_f + k_d T_s + \Delta t_{os} - t_{tm1\min} \\ t_{dz\min} &= (m-n)T_f + k_d T_s + \Delta t_{os} - t_{tm1\max} \end{aligned} \qquad \textbf{equ - 12}$$

and by substituting equ-10 into equ-11, we have

$$\begin{aligned} 2t_{dz\max} &= (m-n)T_f + \Delta T_{os} - T_{ch} - T_{gt} \\ 2t_{dz\min} &= (m-n)T_f + \Delta T_{os} + T_{ch} + T_{gt} - T_f \end{aligned} \qquad \textbf{equ - 13}$$

Let R_z be the range of the equal offset zone measured in terms of the round trip delay variation, then from equ-12, we have

$$R_z = 2(t_{dz\max} - t_{dz\min}) = T_f - 2(T_{ch} + T_{gt}) \qquad \textbf{equ - 14}$$

Appendix-B: Offset value calculation

The offset calculation can be performed according to two steps: first, determine the position of each offset zone, and then calculate the offset based on the propagation delay in the offset zone center.

Assuming that $t_{dz0}(k)$ is the propagation delay from the satellite to the center of the *kth* single offset zone, and $[t_{dz\min}(k), t_{dz\max}(k)]$ is the offset zone delay boundary, then

$$t_{dz0}(k) = [t_{dz\min}(k) + t_{dz\max}(k)]/2 \qquad \textbf{equ - 15}$$

If the satellite burst offset corresponding to the *kth* offset zone is $\Delta t_{os}(k)$ in ms, then by substituting equ-12 into equ-14, we have

$$\Delta t_{os}(k) = 2t_{dz0}(k) - \frac{2(m-n)-1}{2} T_f \qquad \textbf{equ - 16}$$

To achieve frame/timeslot synchronization on the satellite, the offset $\Delta t_{os}(k)$ must be an integer of a timeslot duration. Then $\Delta T_{os}(k)$, the burst offset in number of timeslots can be given as

$$\Delta T_{os}(k) = round\left[\left(2t_{dz0}(k) - \frac{2(m-n)-1}{2} T_f\right)/T_s\right] \qquad \textbf{equ - 17}$$

In equ-16, the operation *round(x)* takes the nearest integer around x. For a geo-stationary satellite constellation, we have $m - n = 7$, and therefore

$$\Delta T_{os}(k) = round[(2t_{dz0}(k) - 6.5T_f)/T_s] \qquad \textbf{equ - 18}$$

Acknowledgment

The authors gratefully acknowledge Dave Roos of the Hughes Network Systems and Jason Lee of the Hughes Space Craft, for providing them with constructive suggestions and advice during their design and simulation work.

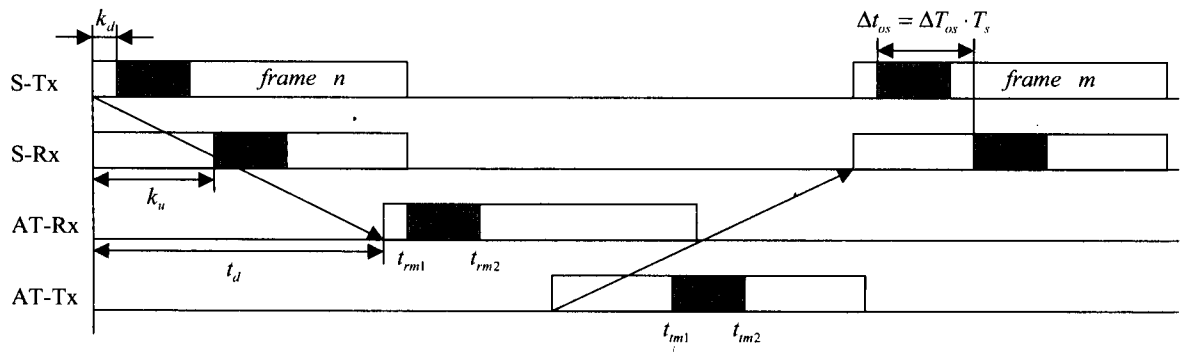


Figure 5 Burst timing on satellite and Access Terminal

Enhanced Throughput for Satellite Multicasting

Daniel Friedman and Anthony Ephremides*
Center for Satellite and Hybrid Communication Networks
Institute for Systems Research
Univerisity of Maryland
College Park, MD 20910
{danielf,tony}@isr.umd.edu

ABSTRACT

Faithful information delivery in satellite multicasting requires appropriate error control. If multicast automatic-repeat-request (ARQ) is employed, a retransmission does not benefit receivers which do not require it, and consequently the throughput suffers greatly as the number of receivers increases. This performance degradation might be alleviated substantially by conducting retransmissions through terrestrial paths from the transmitter to each receiver instead of through the multicast satellite link. By sending a retransmission directly to the receiver(s) which requires it, higher throughput can be provided in such a *hybrid* network than in a pure-satellite network. In this work, we examine the throughput improvement provided by the hybrid network.

INTRODUCTION

Satellites are excellently suited for distributing information simultaneously to multiple locations. As in nearly all communication systems, some sort of error control scheme is required in satellite multicasting to assure satisfactory fidelity of the information provided to each destination. Error control schemes may be broadly classified as forward error correction (FEC) or automatic-repeat-request (ARQ), and both can be applied for satellite communication. FEC has been used in satellite/space communication for decades, having grown from successful application by NASA for communication with interplanetary probes [1, 2]. However, satellite channel characteristics vary with time, and at any given time multiple receivers may perceive different channel qualities. Applying FEC for satellite multicast communication accordingly requires using an error-correcting code strong enough to protect data against the worst-case channel impairments. Unfortunately the error correction capability provided by powerful FEC code comes at the cost of sending many check symbols which constitute overhead in the communication. Further, this overhead penalty is exacted even at times of good channel quality, since FEC is not an adaptive error control technique. This is particularly troubling since good channel conditions will be experienced a majority of the time when using a well-designed satellite link [3].

ARQ protocols adapt to different channel qualities by retransmitting data only as needed. Also, an error-detecting code capable of detecting t or fewer errors in k information symbols requires fewer overhead symbols than would an FEC code designed to correct t errors in the same k symbols [1, 2]. Hence ARQ can provide high fidelity with less overhead than FEC during times of good channel quality, which tend to prevail as mentioned above. A drawback of ARQ not suffered by FEC is the need for a feedback channel, but this requirement is often an acceptable concession for achieving information transfer with excellent fidelity.

A difficulty arises in applying ARQ in multicast settings. The typical problem in a multicast ARQ system is that since retransmissions are sent over the multicast channel, those retransmissions required by only a few receivers do not benefit the other receivers. The other receivers wait unproductively during such retransmissions for a new information frame. Accordingly the throughput for the system falls drastically as the number of receivers increases. Furthermore, if one receiving station is a "poorer listener" than other stations, i.e. it suffers a relatively high frame error rate, then the throughput to all stations is essentially limited to the throughput achievable to that poorer listener [4].

The throughput might be improved considerably if the retransmissions could somehow be sent only to the receivers

*This work was supported in part by the Center for Satellite and Hybrid Communication Networks under NASA Grant NAGW-2777 and by the Institute for Systems Research under National Science Foundation Grant EEC 940234.

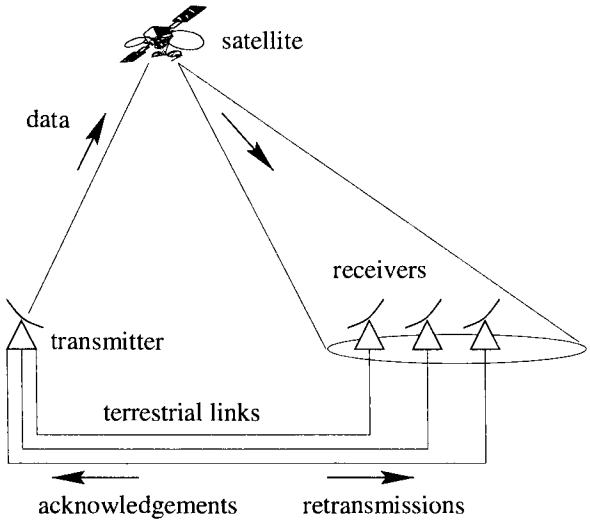


Figure 1: Multicasting in a hybrid network.

which need them. It is natural, then, to suggest supplementing a satellite multicast system with a set of point-to-point terrestrial links between the transmitter and each receiver, as depicted in Figure 1. In such a system, retransmissions may be sent terrestrially instead of via the multicast satellite link, and an improvement in throughput may be so achieved. Furthermore, if the ARQ acknowledgements are set terrestrially as well, then the receiving stations do not require satellite transmission capability and the cost of such stations may be correspondingly reduced.

In this work, we examine the throughput offered by such a *hybrid* (satellite and terrestrial) network configuration for unicast and multicast selective-repeat ARQ operation. In the next section we calculate approximately the throughput for unicasting and multicasting in pure-satellite and hybrid networks. Numerical examples are presented in the following section. We immediately extend these examples to discuss the efficiency of the protocol in the two network architectures. We conclude with a discussion of additional considerations to be regarded in applying the hybrid network for ARQ multicasting.

ANALYSIS

Point-to-Point Case

We first examine unicasting in a pure-satellite network, and make the following assumptions and notational definitions:

1. Infinite supply of information frames available for transmission, so the transmitter is never idled for want of a fresh frame to send.
2. Unlimited buffer size, unlimited window size; ideal selective-repeat ARQ protocol.
3. All acknowledgements are delivered without errors.
4. The probability a frame sent via the satellite link arrives in error at the receiver is p_s, while the terrestrial link frame error rate is p_t.
5. An ARQ information frame sent either via satellite or terrestrially comprises h header (overhead) bits and ℓ information bits.
6. Acknowledgements are sent only for frames received without errors.
7. The satellite channel bit transmission rate r_s exceeds r_t, the terrestrial channel bit transmission rate.
8. In the hybrid network, all retransmissions are sent terrestrially.

We remark that the first of these assumptions is implicit in most selective-repeat ARQ throughput analyses although it is rarely mentioned.

The ultimate purpose of the ARQ system is to deliver error-free information frames in proper order to a consuming process at the receiver. Hence we define as our performance measure the throughput, ν, calculated as the expected number of information bits released per second to the consuming process.

Pure-Satellite Network: To calculate the throughput in unicast and multicast pure-satellite networks, define β as the expected number of frames sent by the transmitter per frame delivered to all receivers. With this definition, β is a measure of *inefficiency* (while its reciprocal is a measure of efficiency).

This inefficiency measure for a pure-satellite architecture with one receiver we shall denote as $\beta_{1,satellite}$, and is given by [1, 2]:

$$\beta_{1,satellite} = \sum_{i=1}^{\infty} i\,(1-p_s)\,p_s{}^{i-1} = \frac{1}{1-p_s}. \qquad (1)$$

The throughput in this setting, $\nu_{1,satellite}$, is then

$$\begin{aligned} \nu_{1,satellite} &= \left(\frac{\ell}{\ell+h}\right)\frac{1}{\beta_{1,satellite}} \\ &= \left(\frac{\ell}{\ell+h}\right)(1-p_s)\ . \end{aligned}$$

Hybrid Network: We may model the protocol operation in the hybrid network as the queueing system shown in Figure 2. We do not consider propagation delays in this model

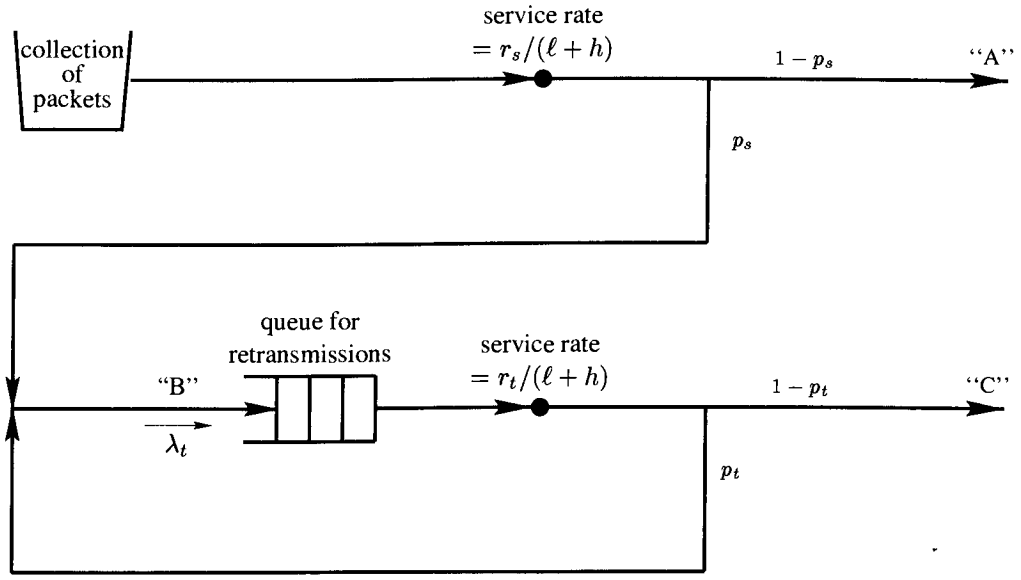


Figure 2: Queueing model for hybrid network.

since we have assumed unlimited window size and unlimited buffering. In this model, information frames are sent continuously via satellite to the receiver. With probability $1 - p_s$ a frame sent via the satellite link is successfully received. Hence the average rate at which frames are delivered successfully via the satellite link is the average frame flow rate at point "A" in the figure, $r_s\,(1 - p_s)\,/(\ell + h)$ frames per second.

A frame which is which is corrupted in satellite transmission is queued at the transmitter for retransmission. A retransmitted frame may be successfully received with probability $1 - p_t$, or the retransmission may be unsuccessful, in which case the frame is queued for another retransmission. Hence a frame may be retransmitted multiple times before it is successfully delivered.

Denote as λ_t the average frame flow rate at the input to the retransmissions queue, point "B." We will say the system is *stable* if the average input rate to the retransmissions queue is less than the rate at which retransmissions can be sent, i.e.

$$\lambda_t < r_t/(\ell + h), \tag{2}$$

and *unstable* otherwise.

Now if the system is stable, conservation of flow indicates $r_s p_s/(\ell + h) + \lambda_t p_t = \lambda_t$ or

$$\lambda_t = \frac{r_s}{\ell + h}\left(\frac{p_s}{1 - p_t}\right),$$

which in combination with (2) implies another expression of the stability condition,

$$\frac{r_s}{\ell + h}\left(\frac{p_s}{1 - p_t}\right) < \frac{r_t}{\ell + h}. \tag{3}$$

The corresponding average rate at which frames arrive successfully at the receiver via the terrestrial link is the average frame flow rate at "C," $\lambda_t\,(1 - p_t)$. If the system is unstable, then this same flow rate is limited to $r_t\,(1 - p_t)\,/(\ell + h)$.

Hence we have $\nu_{1,hybrid}$, the throughput in a single-receiver hybrid architecture:

$$\nu_{1,hybrid} = \begin{cases} \frac{\ell r_s}{\ell + h} & \text{, if stable} \\ \frac{\ell r_s}{\ell + h}(1 - p_s) + \frac{\ell r_t}{\ell + h}(1 - p_t) & \text{, if unstable} \end{cases}$$

These results may be combined with (3) to yield the more compact expression

$$\nu_{1,hybrid} = \frac{\ell}{\ell + h}\min\{r_s, r_s\,(1 - p_s) + r_t\,(1 - p_t)\},$$

in which the first term in the minimization corresponds to stable operation and the second corresponds to unstable operation.

It should be noted that Figure 2 does not necessarily represent an *implementation* of a hybrid network. In particular, one might expect prolonged unstable operation would lead to overflow of a finite retransmissions buffer. However, if the system is implemented with a common window for frames sent on the satellite and terrestrial links, then it is

possible to assure no overflow of frames from the retransmissions buffer during prolonged unstable operation. Even so, the window size cannot be unlimited in an implementation, and so the flow of fresh information frames on the satellite link may have to be cyclically suspended and resumed to allow for many terrestrial retransmissions if the system is unstable in the sense described above.

Point-to-Multipoint Communication

For analyzing multicast networks, we preserve the assumptions of the point-to-point analysis and add the following:

1. There are $M > 1$ receivers. (The results presented here also apply for $M = 1$.)
2. The noise processes experienced by all receivers are independent and identical.
3. There is no competition among receivers for access to the acknowledgment channel.
4. The ARQ protocol operation is according to the Dynamic Retransmissions Group Reduction (DRGR) technique described in [5]. The essential feature of this multicast selective-repeat ARQ protocol is that the transmitter maintains a history of which stations have acknowledged which frames. Accordingly, if receiver $m \in \{1, 2, \ldots, M\}$ has positively acknowledged receipt of frame $\mathcal{F}$, an acknowledgement is not required from m for any retransmissions of $\mathcal{F}$ which may be required for other receivers in the network.

We define the throughput of the multicast system as the average of the unicast throughputs to all receivers.

Pure-satellite Network: In the multicast pure-satellite network, the transmitter continuously sends frames via the satellite multicast channel to the M receivers, which generate respective acknowledgments to send to the transmitter. Upon receiving acknowledgments from the receivers, the transmitter retransmits the frame if one or more receivers so request through their acknowledgements. Otherwise a new frame is sent.

Let m_j denote the number of receivers which successfully receive a frame $\mathcal{F}$ after exactly j multicast transmission attempts to deliver $\mathcal{F}$. Also let $\gamma(j)$ denote the probability with which the frame $\mathcal{F}$ is successfully delivered to all M receivers with j or fewer transmissions. This probability may be found by counting all possible combinations of the number of transmissions required to deliver frame $\mathcal{F}$ to each of the M receivers, given $\mathcal{F}$ was transmitted j times, yielding

$$\gamma(j) = \sum_{m_1=0}^{M} \cdots \sum_{\substack{m_j=0 \\ \sum_{h=1}^{j} m_h = M}}^{M} \left[\binom{M}{m_1, m_2, \cdots, m_j} \times \prod_{k=1}^{j} \left[{p_s}^{k-1} (1 - p_s) \right]^{m_k} \right]$$

where the multinomial coefficient is given by

$$\binom{M}{m_1, m_2, \cdots, m_j} = \frac{M!}{m_1! m_2! \cdots m_j!}.$$

An simpler way to calculate $\gamma(j)$ is to consider the complement of the events of the destinations not receiving $\mathcal{F}$ after j transmissions, yielding

$$\gamma(j) = \left(1 - {p_s}^j\right)^M.$$

By its definition, $\gamma(j)$ is the cumulative distribution function for the number of transmissions required to successfully deliver a frame to all receivers. Then we may calculate $\beta_{M,satellite}$, the expected number of frames sent per frame delivered to all M receivers in the pure-satellite network, as:

$$\beta_{M,satellite} = \sum_{j=1}^{\infty} j[\gamma(j) - \gamma(j-1)] \qquad (4)$$

Hence the throughput for multicasting in a pure-satellite network, $\nu_{M,satellite}$ is

$$\nu_{M,hybrid} = \left(\frac{\ell}{\ell + h}\right) \frac{1}{\beta_{M,satellite}}$$

with $\beta_{M,satellite}$ calculated as above. We remark that evaluating (4) with $M = 1$ yields (1), and so the expressions for β and ν in the single– and multiple–receiver cases of the pure-satellite network are consistent.

Hybrid Network: Although a multicast transmitter in a hybrid network must keep track of more information and service more receivers with retransmissions than its unicast counterpart, the receivers in the multicast setting are the same as the unicast receiver. Hence $\nu_{M,satellite}$, the throughput for multicasting in a hybrid network, is the same as for unicasting in the hybrid network:

$$\begin{aligned} \nu_{M,hybrid} &= \nu_{1,hybrid} \\ &= \tfrac{\ell}{\ell+h} \min\{r_s, r_s(1-p_s) + r_t(1-p_t)\} \end{aligned}$$

In particular, the throughput in the hybrid network is independent of the number of receivers in the network.

NUMERICAL EXAMPLES

We now turn to some numerical examples to better understand the throughput expressions derived above. For these examples, we will make the following further assumptions:

1. Binary symmetric channel (BSC) models characterize the terrestrial channels and the logical satellite channels between the transmitter and each receiver. The crossover probabilities (bit-error rates, BERs) are q_s for all logical satellite channels and q_t for all terrestrial channels.

2. The channel bit transmission rates are $r_s = 1536000$ and $r_t = 33600$ bits per second in the satellite and terrestrial channels, respectively.

3. There are $\ell = 1776$ information bits and $h = 32$ overhead bits in all ARQ information frames, whether sent via satellite or via a terrestrial link. (The value of h was chosen supposing the ARQ frame has a 16-bit sequence number and a 16-bit CRC for error detection. The value of ℓ was chosen to maximize the throughput in a point-to-point satellite network, which is the reference network for comparison purposes. This maximization is calculated by a straightforward differentiation method presented in [6]. For this maximization, q_s was taken to be 10^{-5}, the median value of the satellite link BERs examined below.)

4. In calculating $\beta_{M,satellite}$, we approximated the infinite summation of (4) by truncating the summation at the minimum j such that $\gamma(j) > 1 - 10^{-10}$. (We justify this truncation not only as a fair approximation, but also because, in an actual network, a station which requests retransmissions too frequently would likely be recognized by the transmitter as suffering from excessive noise, and would accordingly be disconnected from the communication.)

Calculated throughput values for point-to-point communication are presented in Figure 3. The reduction of throughput with increase in number of receivers in the pure-satellite network is clearly indicated in the figure. The figure also shows the throughput in the hybrid network meets or exceeds that in the pure-satellite network and is independent of the number of receivers.

Of course, achieving the higher throughput of a hybrid network requires the terrestrial links from the transmitter to each receiver. This need for additional bandwidth naturally prompts asking about the efficiency of the hybrid network. Several efficiency measures can be defined, each with its own merits and pitfalls. We choose to define the throughput efficiency, η, as the throughput divided by the sum of the network link bandwidths. That is, the efficiency in the pure-satellite network of M receivers is $\eta_{M,satellite} = \nu_{M,satellite}/r_s$, while the efficiency in a corresponding hybrid network is $\eta_{M,satellite} = \nu_{M,satellite}/(r_s + Mr_t)$. Applying these definitions to the setting and results of the foregoing examples yields the results shown in Figure 4. For the hybrid network, $\eta_{M,hybrid}$ decreases as M increases, as is evidenced in the figure. The figure also indicates the pure-satellite network is more efficient than the hybrid network unless the satellite link BER is fairly high (say, 10^{-4}), and only up to some number of receivers (say, 100)—at which point the efficiency of the pure satellite network again exceeds that of the hybrid network. So, while the hybrid network provides better throughput than the pure-satellite network, it does so at greater cost, according to at least the efficiency measure defined above. As remarked earlier, our efficiency measure is but one of several which can be defined.

ADDITIONAL CONSIDERATIONS

The inherent problem in ARQ multicasting, as stated in the Introduction, is that retransmissions sent over the multicast channel do not benefit stations which do not require them. Consequently the throughput suffers as the number of receivers increases. In this work we have suggested a solution to this problem, namely retransmissions should be sent over a system of point-to-point terrestrial links between the transmitter and each receiver. However, many considerations remain to be studied.

Perhaps foremost among these concerns is how to increase the efficiency of the protocol's operation in the hybrid network. At times when the satellite link experienced by one receiver is good that receiver's terrestrial channel may be underutilized. This suggests that if the transmitter can deduce the approximate quality of the satellite link to each receiver, perhaps initial transmissions of some frames can be conducted on the terrestrial links to increase the throughput and efficiency.

We must also examine the effect of packet lengths on throughput. While the frame length which maximizes throughput in a point-to-point satellite network is easily calculated ([6]), the optimal frame length for unicasting in a hybrid network, and for multicasting in satellite and hybrid networks, remains to be found. Adaptively changing the frame length may offer a throughput advantage, particularly at high bit error rates in the satellite channel.

Important implementation issues arise in multicasting, and these must be examined. In addition to well-known multi-

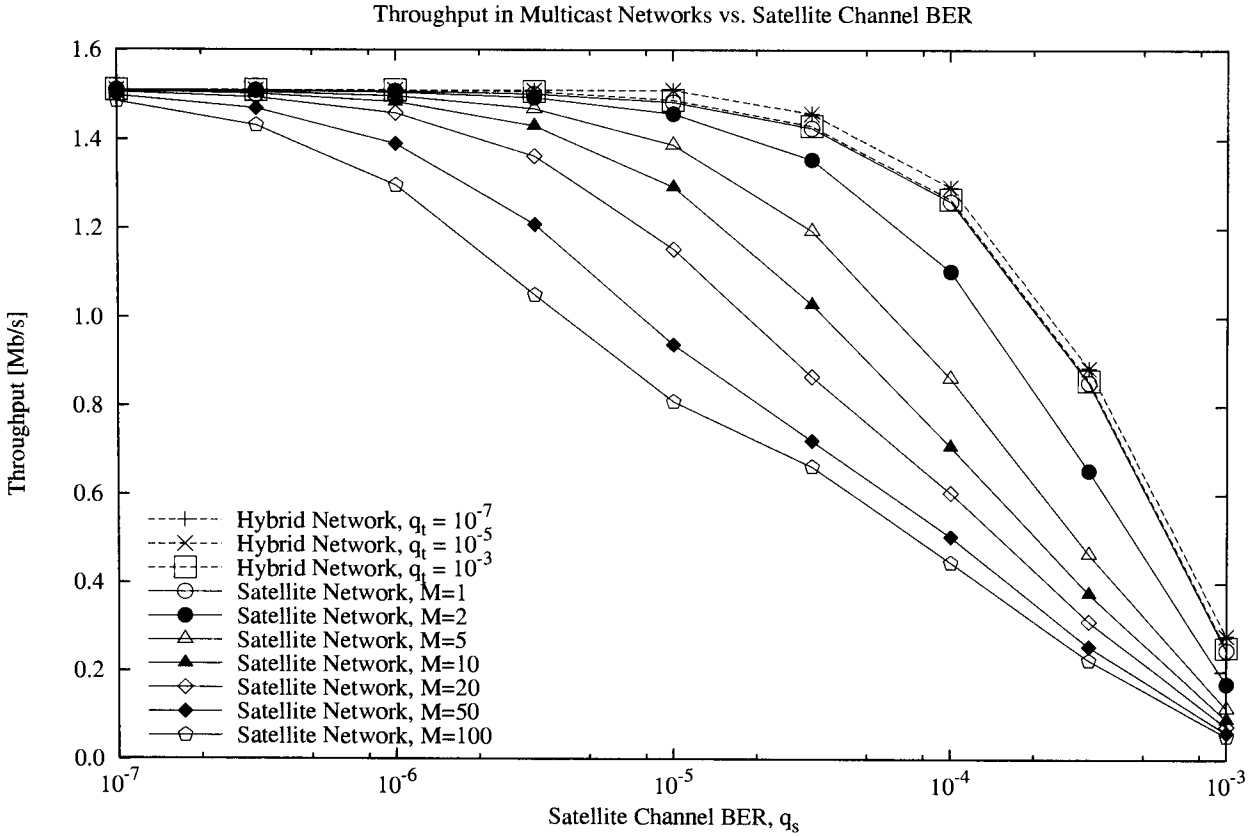


Figure 3: Throughput in point-to-multipoint networks.

cast network concerns such as overwhelming the transmitter with acknowledgements, some are unique to the hybrid network. In particular, it is not clear if the transmitting station should have a single protocol governing operation on all links, or if the operation should be split into a protocol for the satellite link and separate protocols for the terrestrial links. How this split should be made and how the pieces of the transmitter's operation should be connected should similarly be explored.

We have also not yet studied terrestrial network topologies other than a star topology. Our proposed solution does not necessarily preclude other configurations. On the contrary, other topologies are not only acceptable, but perhaps even desirable. In particular, suppose the terrestrial network is a tree of terrestrial links, with the transmitter at the root node and a receiver at each non-root node. Such a tree could not only support multicasting in a hybrid network as we have described above, but would also allow a retransmission request sent by one receiver node to be serviced by the nearest ancestor node having the requested frame. The transmitter's load in servicing retransmission requests would then be reduced. Such operation in a tree-shaped terrestrial is similar to the operation of the Reliable Multicast Transport Protocol [7, 8].

Similar possibilities arise if the terrestrial network is a wireless network, as in, for example, the case of mobile receiving nodes. For example, mobile receivers, with omnidirectional antennas, can broadcast retransmission requests to other receivers possibly nearby and receive frames over the terrestrial wireless channel. A terrestrial tree for retransmissions, albeit a continuously changing tree, is perhaps applicable for mobile receivers as well.

Hybrid ARQ schemes for multicasting, which employ FEC techniques for improving throughput have appeared in the literature recently, and these suggest possibilities in the context of hybrid networks [9, 10, 11]. (The reader is cautioned that the term "hybrid ARQ," which is the standard term in the literature for ARQ schemes incorporating FEC, is not related to our term of "hybrid network" for a parallel arrangement of satellite and terrestrial networks.) In [9], for

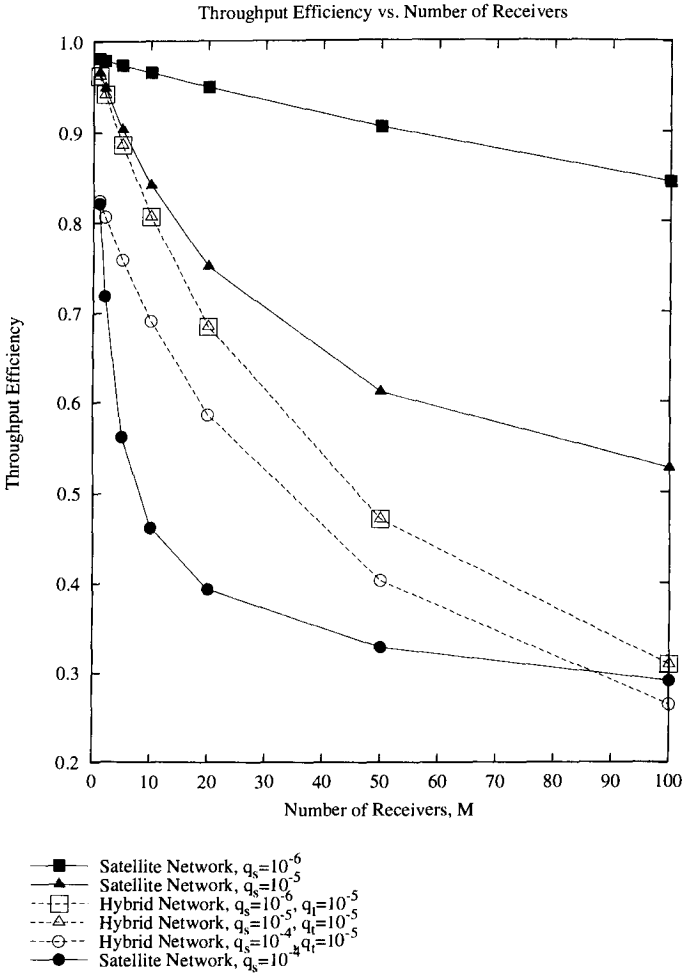


Figure 4: Throughput efficiency in point-to-multipoint networks.

example, an adaptive type-II multicast hybrid ARQ scheme is proposed. Rate-compatible BCH codes are used for error correction in this scheme. Each time another retransmission is requested for a particular frame, the transmitter sends an increasing number of parity digits, which, when combined with the original data frame, from a series of BCH codewords of decreasing rate. Employing such an FEC technique would not only improve throughput in a hybrid network, it would reduce terrestrial retransmissions and so increase the terrestrial bandwidth possibly available for original transmissions, as described at the beginning of this section.

There are clearly many aspects of multicast ARQ to explore. In addition to exploring such aspects, we intend to also consider how to tolerate and/or recover from errors in systems where the multicasted information has delay constraints, such as voice and video multicast systems. Because of the delay constraints, ARQ is not suited well for error control in such settings, and other schemes for mitigating error effects must be devised.

*

[1] S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*. Prentice-Hall, 1983.

[2] S. B. Wicker, *Error Control Systems for Digital Communication and Storage*. Prentice-Hall, 1995.

[3] T. Pratt and C. W. Bostian, *Satellite Communications*. Wiley, 1986.

[4] Y. Yamauchi, "On the packet radio multicast scheme for the personal communications era," in *International Conference on Communication Systems (ICCS '94), Singapore*, pp. 576–580, IEEE, 1994.

[5] K. Sabnani and M. Schwartz, "Multidestination protocols for satellite broadcast channels," *IEEE Transactions on Communications*, vol. 33, pp. 232–240, Mar. 1985.

[6] M. Schwartz, *Telecommunication Networks: Protocols, Modeling, and Analysis*. Addison-Wesley, 1987.

[7] S. Paul, K. Sabnani, J. Lin, and S. Bhattacharyya, "Reliable multicast transport protocol (RMTP)," *IEEE Journal on Selected Areas in Communications*. To appear in special issue on Network Support for Multipoint Communications. (Also available at http://bell-labs.com/users/sanjoy/rmtp2.ps.)

[8] S. Paul, K. Sabnani, and D. Kristol, "Multicast transport protocols for high-speed networks," in *Proceedings of the International Conference on Network Protocols*, pp. 4–14, 1994.

[9] A. Shiozaki, "Adaptive type-II hybrid broadcast ARQ system," *IEEE Transactions on Communications*, vol. 44, pp. 420–422, April 1996.

[10] H. Liu, Q. Zhang, M. E. Zarki, and S. Kassam, "Wireless video transmission with adaptive error control," in *1996 International Symposium on Information Theory and its Applications (ISITA '96), Victoria, British Columbia*, pp. 371–374, 1996.

[11] H. Zhao, T. Sato, and I. Kimura, "A hybrid-ARQ protocol with optimal adaptive error control for multidestination satellite communications," in *International Conference on Communication Systems (ICCS '94), Singapore*, pp. 420–424, 1994.

GeoMobile (GEM™) Satellite System Physical Layer Overview.

Yezdi Antia
Hughes Network Systems,
11717 Exploration Lane, Germantown MD 20876.
E-mail: antia@hns.com

ABSTRACT

This paper deals with an overview of the physical layer of the GEM™ system. The paper describes the various logical channels that are part of the system and gives the modulation and coding which is used to carry them. It explains the multiple access scheme. It describes the RF power control mechanism and the timing and frequency synchronization used by the system. The terminal characteristics are outlined and various radio link measurements enumerated.

INTRODUCTION

The Hughes GeoMobile (GEM™) system consists of one or two geo-synchronous satellites with an associated ground-based advanced operations center (AOC) and a spacecraft operations center (SOC), one or more ground-based gateway stations (GSs), and a number of user terminals (UTs) or access terminals (ATs). User terminals communicate over the satellite with the terrestrial public switched telephone network (PSTN) via the GSs and with each other under control of the GSs. The AOC provides system wide resource management and control functions and each SOC provides on-orbit satellite operations for one satellite.

The GEM™ system delivers mobile telephone, data, fax, and messaging services to users in the designated coverage region using digital communication techniques. Digital communications between subscribers or between subscribers and the public telephone users is provided via a high-powered geo-synchronous satellite and one or more gateway ground stations. The system supports subscriber roaming among gateways, between the GEM™ systems and Global System for Mobile Communications (GSM) terrestrial cellular systems, using GSM roaming procedures.

This paper concentrates on describing the physical layer of the GEM™ system.

LOGICAL CHANNELS[1].

The logical channels associated with the GEM™ system may either be a traffic channel (TCH) or a control channel. The control channel is further divided into common control channels (CCCH) and dedicated control channels (DCCH).

1 Traffic Channel: These are intended to carry encoded speech or user data which could be fax or data. These are all bi-directional channels.
   a) TCH3: This channel carries normal speech and has a gross information rate of 5.2 kbps
   b) TCH6: This channel carries 4.8 kbps user data and has a gross information rate of 10.75 kbps.
   c) TCH9: This channel carries 9.6 kbps user data and has a gross information rate of 16.45 kbps.

2 Control Channels: These are intended to carry signaling or synchronization data.
   a) Broadcast Channel: This is a downlink only channel and consist of the following
      i) Frequency Correction Channel (FCCH): It is used by the UT for system information synchronization and carries information for frequency correction of the UT.
      ii) GPS Broadcast Control Channel (GBCH): It carries global positioning system (GPS) time information and GPS satellite ephemeris information.
      iii) Broadcast Control Channel: (BCCH): It is used to broadcast system information to the UTs.
      iv) Cell Broadcast Channel (CBCH): It is used to broadcast short message service cell broadcast information to the UTs on a spot beam basis.
   b) Common Control Channel (CCCH): This consists of the following
      i) Paging Channel (PCH): This is a downlink only channel and used to page UTs.
      ii) Random Access Channel (RACH): This is a uplink only channel and is used to request a channel (SDCCH or TCH) allocation.
      iii) Access Grant Channel (AGCH): This is a downlink only channel used to allocate a standalone SDCCH or TCH.
      iv) Basic Alerting Channel (BACH): This is a downlink only channel and used to alert UTs.
   c) Dedicated Control Channels (DCCH): These are channel resources that are dedicated for that UT. The are all bi-directional except the TACCH which is downlink only.

i) Slow TCH6-associated control channel (SACCH6).
ii) Slow TCH9-associated control channel (SACCH9).
iii) Fast TCH3-associated control channel (FACCH3).
iv) Fast TCH6-associated control channel (FACCH6).
v) Fast TCH9 associated control channel (FACCH9).
vi) Standalone dedicated control channel (SDCCH).
vii) Terminal-to-terminal associated control channel (TACCH). This channel can be shared among a subset of terminal-to-terminal calls and is not necessarily dedicated to a single terminal-to-terminal call.

RADIO FREQUENCY BANDS[4]

Operational frequencies in the mobile band may be anywhere within the 34 MHz L-band 1.525 GHz-1.559 GHz (downlink) and 1.6265 GHz-1.6605 GHz (uplink); each carrier will be centered on an integer multiple of 31.25 kHz. To minimize the time spent by ATs during spot beam synchronization, identification, and selection, a subset of RF carriers called broadcast control channel (BCCH) carriers is used by the network to broadcast BCCHs. L-band RF carriers are configured for each spot beam, depending on traffic demand, frequency reuse considerations, and available spectrum as a result of coordination with other systems using the same spectrum. Any RF channel can be used in any spot beam.

Feederlink frequencies are in the C-band range of 6.475 to 6.725 GHz (uplink) and 3.400 to 3.625 GHz (downlink).

MULTIPLE ACCESS[2]

The access scheme used is Time Division Multiple Access (TDMA) and Frequency Division Multiplex (FDM). The multiple access unit is a timeslot which has a duration of 5/3 msec. Twenty-four timeslots form a TDMA frame, 40 msec in duration. At the satellite, the TDMA frames on all of the radio frequencies in the downlink of each spot beam will be aligned. The same also applies to the uplink. The timeslots within a TDMA frame will be numbered from 0 to 23, and a particular timeslot will be referred to by its timeslot number (TN).

TDMA frames will be numbered by a frame number (FN). The frame number will be cyclic and have a range of 0 to 313,343. The frame number will be incremented at the end of each TDMA frame. The complete cycle of TDMA frame numbers is defined as a hyperframe. The need for a hyperframe arises from the requirements of the encryption process, which uses FN as an input parameter.

Other combinations of frames include see Figure 1:

- **Multiframes.** A multiframe consists of 16 TDMA frames. Multiframes are aligned so that the FN of the first frame in a multiframe, modulo 16, is always 0.
- **Superframes.** A superframe consists of four multiframes. Superframes are aligned so that the FN of the first frame in a superframe, modulo 64, is always 0.

BURST MODULATION[3]

The modulating symbol rate is 23.4 ksps for all burst types, except for the Broadcasting Alert Channel (BACH). The symbol period T is defined as 1/23.4 msec. For BACH, the symbol rate is 300 sps (1200 bps).

The modulation scheme used for all traffic and control channels except for DKAB, BACH and FCCH whether they are uplink or downlink is one of the following schemes.

1. $\pi/4$-CQPSK (coherent quadrature phase shift keying) which is pulse shaped using a root raised cosine filter with a rolloff factor of 0.35. This is used by the TCH3, TCH6, TCH9, common control and broadcast control channels.
2. $\pi/4$-CBPSK (coherent binary phase shift keying) which is pulse shaped using a root raised cosine filter with a rolloff factor of 0.35. This is used by the FACCH3 and SDCCH channels. This modulation scheme along with some additional FEC allows these channels to operate in a disadvantaged channel condition compared to the normal traffic channels. All the call setup signaling can take place while the UT may not be optimally deployed for traffic operation.

A novel Dual Keep Alive Burst (DKAB) is transmitted during periods of speech inactivity and is proposed by the GEM™ System to save the battery life, satellite power, reduce cochannel interference, add comfort noise and maintain the power control and timing/frequency synchronization (more information maybe found in a related paper in this conference proceedings). The modulation scheme used by DKABs is $\pi/4$-DBPSK (differential binary phase shift keying) which is pulse shaped using a root raised cosine filter with a rolloff factor of 0.35.

The BACH is modulated using a 6-PSK which is pulse shaped using a root raised cosine filter with a rolloff factor of 0.35. The FCCH burst is a real chirp signal spanning three slots.

CHANNEL CODING[2]

The channel coding consists of the following basic units. The exact configuration of these varies from channel to channel, depending on the size of the information bits, the

amount of coding gain that the channel needs to achieve required performance, etc.

1 Outer Code - Cyclic Redundancy Check which is 8, 12 or 16 bits of parity depending on the channel.
2 Inner Code -
   a) Convolutional Coding. Most channels use a convolutional coder with constraint length of 5 and rates of 1/2, 1/3, 1/4 and 1/5 as required by the particular channel. The TCH3 channel which caries speech data is coded using a constraint length 7 convolutional encoder and utilizes tail biting (circular encoding) to avoid overhead of the tail bits.
   b) Golay Coding: The power control messages that are embedded into various burst use the (24,12) systematic golay encoder. The golay decoder used is a soft decision golay decoder which gives additional gain over the traditional hard decision decoding.
   c) Reed-Solomon Coding: The BACH uses a systematic (15,9) Reed-Solomon code generated over the Galois Field GF(2^4).
3 Puncturing. Various puncture masks are used to fit the coded bits into the physical channel bit carrying capacity. The masks are designed to minimize the performance degradation from the unpunctured case.
4 Interleaving: It can be intra-burst or inter-burst interleaving which is based on block interleaving methods with pseudo random permutations and is channel dependent.
5 Scrambling: The scrambler adds a binary pseudo-noise sequence (masking sequence) to the input bit stream to randomize the number of 0s and 1s in the output bit stream. The masking sequence is generated by a linear feedback shift register.
6 Encryption: Certain channels have data encryption to prevent eavesdropping.

TERMINAL TRANSMIT AND RECEIVE CHARACTERISTICS[4]

The AT transmitter and receiver characteristics are as follows.

1. The transmitter output power is measured in terms of EIRP and ranges from 5 to 9 dBW, depending on the terminal type which may be a handheld, vehicular with or without an adjustable antenna or a fixed terminal.
2. The transmitter has been characterized in detail [4] specifying the antenna radiation pattern, its polarization, carrier off EIRP, power control range and accuracy, adjacent channel interference and spurious emissions.
3. The receiver has also been characterized in detail [4] specifying the antenna radiation pattern, its polarization, figure of merit, sensitivity, selectivity, intermodulation and blocking characteristics.
4. The receiver sensitivity is defined under various channel conditions. The conditions are Static (AWGN) Channel which represents strong direct line-of-sight to the satellite, Rician fading channel with K factor (direct to multipath ratio) of 9 dB and fade bandwidth of 10 Hz which represents a typical handheld with the user walking and Rician fading channel with K factor of 12 dB and fade bandwidth of 200 Hz which represents a typical vehicular terminal moving at 60 mph.

RF POWER CONTROL[5]

The power control in the GEM™ system is performed for all the active traffic channels in the GS-to-AT direction, the AT-to-GS direction and in the AT-to-AT configuration. There is no power control for the common control channels. The AT transmits the RACH and SDCCH at full power.

RF power control is employed to minimize the transmit power required at the AT or the GS while maintaining the quality of the radio links. By minimizing the transmit power levels, the unnecessary power source drain at the satellite and the AT is prevented, and the cochannel interference due to the signals received from different ATs is reduced.

The power control aims at a fast transient response to mitigate sudden shadowing events, a steady-state condition that accurately achieves the designated received signal quality, and a robust operation with respect to the error conditions. Power control is governed by many different parameters, e.g., target signal quality, power control loop gain, etc. These parameters provide the capability to adjust the power control response for different channel conditions, terminal types, operation policies, etc.

The power control mechanism has two ends. One of the two power control ends is an AT. The other end is either a GS or an AT. In the open-loop power control mechanism, each power control end increases the transmit power if the quality of the received signal suddenly deteriorates by a designated amount. The use of open loop power control assumes a useful degree of statistical correlation between the receive and the transmit links. In closed loop power control, the receiver end estimates the quality of the signal received from the transmitter end and, based on the estimated signal quality, conveys to the transmit end a request for attenuation relative to the maximum transmit power level. Closed loop power control is performed in both the downlink (the control of the GS transmit power based on the received signal quality measurement at the AT) and in the uplink (the control of the AT transmit power based on the signal quality measurement at the GS).

IDLE MODE TASKS[5]

While in idle mode, an AT shall implement the spot beam selection and reselection procedures. These procedures ensure that the AT is camped on a spot beam from which it

can reliably decode downlink data and with which it has a high probability of communications on the return link. An accurate spot beam selection is crucial to minimize satellite power cost and timeslot assignment blockage, and to limit the number of spot beam re-selection procedures.

For the purposes of spot beam selection and reselection, the AT shall be capable of detecting and synchronizing to a BCCH carrier and reading the BCCH data at reference sensitivity and reference interference levels. The AT is also required to maintain an average of the received signal strength for all the monitored frequencies. (More information maybe found in a related paper in these conference proceedings)

RADIO LINK MEASUREMENTS[5]

The AT will make radio link measurements for the RF power control, radio link failure criterion, idle mode beam selection/reselection procedures, idle mode selection criteria, and user strength indication. The GS will make radio link measurements for RF power control, radio link failure criterion, and time and frequency synchronization processes. These measurements are as follows.

1. Received Signal Strength Indicator (RSSI): The RSSI is the rms value of the signal received at the antenna.
2. Signal Quality Indicator (SQI): The SQI is defined as an estimate of the ratio of the desired signal power to the noise and interference power in the received burst.
3. Link Quality Indicator (LQI): The LQI is the amount of reserve link margin, with respect to the target signal quality. A negative value of LQI indicates that the target signal quality is not being met by the indicated value.
4. Receive Symbol Time and Carrier Frequency Offsets.
5. Interference Plus Noise to Noise Ratio (INR): It is the ratio of the interference and the noise power to the noise power. It indicates the increase in the additive noise floor due to the presence of the interference.

SYSTEM SYNCHRONIZATION[6]

GEM™ is a multi-spot beam, multicarrier, synchronous system where the timing and frequency on the satellite serve as the reference to synchronize the TDMA transmissions for the ATs, the network GSs and other network elements. The satellite includes a switch designed to provide single-hop, terminal-to-terminal (TtT) connectivity at L-band.

Timing Synchronization

The general requirement for AT timing synchronization is that the AT will transmit signals that are time aligned and frame number aligned with the system timing on the satellite reference point.

The whole system is synchronized on the satellite. The network adjusts FCCH and BCCH transmission so that each of these channels leaves from the satellite antenna at the predefined system timing. An AT derives its local timing reference from the signals received from the satellite. By listening to the FCCH, both timing and frequency synchronization can be achieved for CCCH channels.

From a cold start, ATs initially search for and acquire the FCCH sent in each spot beam. The AT's frame timing is then synchronized to system timing.

In idle mode, after initial timing acquisition, the AT needs to track system timing continuously to compensate the timing drift caused by its local oscillator frequency uncertainty and the relative motion between the satellite and the user.

At initial access, an AT accesses the network using a RACH offset pre-calculated for the spot beam center. This RACH offset is distributed by the BCCH in each spot beam. The round trip delay variation caused by the difference of AT position relative to the beam center will be detected from the network, and this value will be passed to the AT as a timing correction via the AGCH. After receiving the AGCH, the AT will be able to transmit such that timing of burst arrival on the satellite is nominal.

At the beginning of a call, to achieve frame/timeslot synchronization on the satellite, a transmission frame offset relative to the start of downlink reference frame is provided from the network. During a call, both AT transmitter and receiver adjust their burst timing to maintain the frame/timeslot synchronization. The AT receiver timing is maintained by using its internal timebase. For the AT transmitter, a closed loop synchronization scheme is adopted. Any transmission timing drift at the AT will be detected from the network by comparing the actual burst arrival with the expected arrival, and a timing correction is passed to the AT if the difference exceeds a threshold defined by the network.

Frequency Synchronization

Both forward and return link signals are required to align their nominal frequencies on the satellite. The task of frequency synchronization is to precompensate the transmission signal to align the nominal frequency on the satellite and to track the received signal in frequency to achieve effective demodulation.

The AT frequency alignment is achieved by correcting transmission frequency with messages provided by a network. RACH frequency is set up by messages provided over the BCCH. SDCCH/TCH frequency is corrected with corrective factors given over the AGCH. During a call, frequency correction is provided through FACCH (TCH3) or SACCH (TCH6/TCH9).

ACKNOWLEDGEMENTS.

The author would like to acknowledge the contributions of all the people who were involved with the design and definition of the physical layer specification of the GEM™ system.

REFERENCES:

[1] GEM 05.02 (HNS): "GEM™ Satellite Telecommunications System (Phase 2); Multiplexing and Multiple Access; Stage 2 Service Description."

[2] GEM 05.03 (HNS): "GEM™ Satellite Telecommunications System (Phase 2); Channel Coding.

[3] GEM 05.04 (HNS): "GEM™ Satellite Telecommunications System (Phase 2); Modulation."

[4] GEM 05.05 (HNS): "GEM™ Satellite Telecommunications System (Phase 2); Radio Transmission and Reception."

[5] GEM 05.08 (HNS): "GEM™ Satellite Telecommunications System (Phase 2); Radio Subsystem Link Control."

[6] GEM 05.10 (HNS): "GEM™ Satellite Telecommunications System (Phase 2); Radio Subsystem Synchronization."

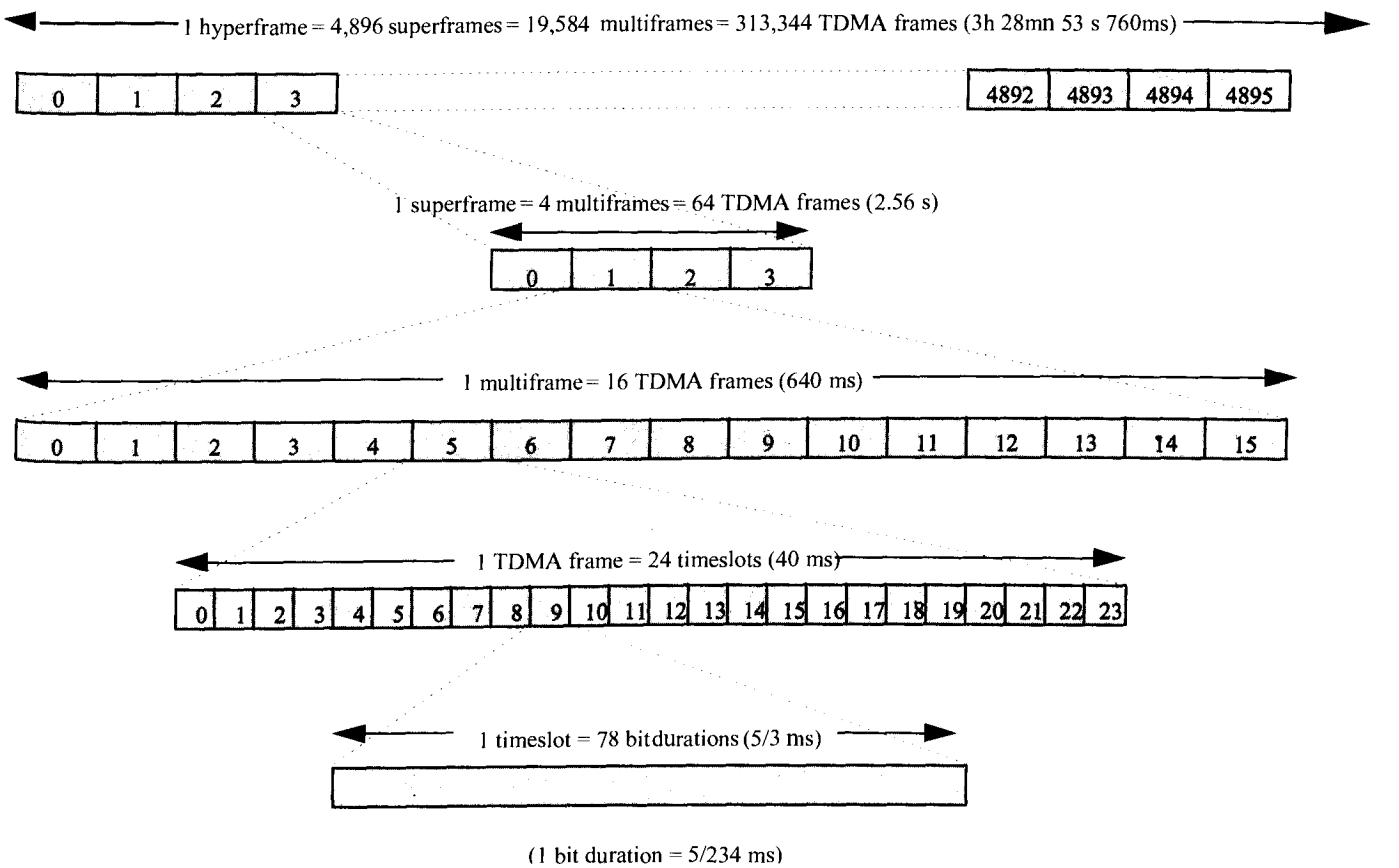


Figure 1 . Frame and Timeslot Structures.

Synchronization for Geo-Mobile (GEM™) Satellite TDMA Transmission System

Ludong Wang
Hughes Network Systems
11717 Exploration Lane, Germantown, MD 20876
Tel: (240) 453-2148
Email: lwang@hns.com

Abstract

Presented in this paper is a robust synchronization scheme for Geo-Mobile (GEM™) Satellite up-link transmission system. Composed of estimators and link outage detector, the proposed scheme is capable of providing reliable timing and frequency synchronization through serious channel impairment. Characteristic of robustness and reliability, it can be extended, partially or as a whole, to other related applications.

1. Introduction

In the up-link of Geo-Mobile (GEM™) Satellite TDMA transmission system, traffic channel bursts are transmitted from user terminals (UT) to satellite gateway station subsystem (GSS). Timing and frequency may drift randomly due to the mobile UT. In the GSS receiver, overall transmission performance and quality largely depend on the synchronization of these two temporal parameters. To achieve reliable synchronization, estimators of both symbol timing and frequency offset must be robust for fading channel conditions and other unpredictable channel impairment. In addition, complexity of these estimators is also constrained for practical implementation.

In GEM™ system, there are very few known symbols available during traffic transmission. Non-data-aided approaches are reasonable options for better performance. On the other hand, direct estimation from each individual burst exhibits random jitter, which actually does not reflect the true temporal variation but noise. When mixed types of bursts are received and estimation from unfavorable bursts is relatively poor, as in GEM™ system, smooth tracking procedure is indispensable.

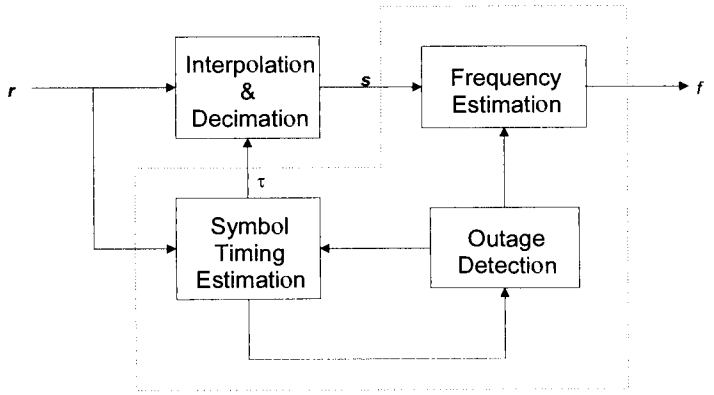


Figure 1. Synchronization Architecture of GEM™ Traffic Transmission

On the other hand, occasional link outage is inevitable in GEM™ applications. Temporary loss of transmitted signal may incur loss of synchronization. Reliable detection of such occurrence is crucial for both maintenance of synchronization and resumption of transmission.

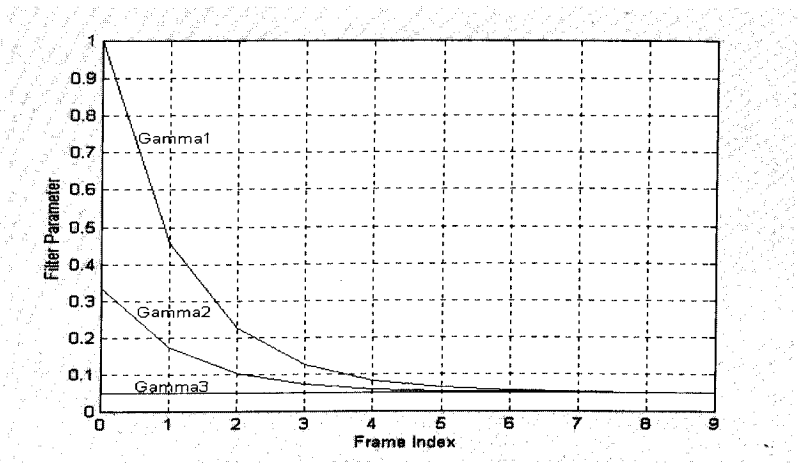


Figure 2. Dynamic Profile of Filter Parameter γ

Based on these system constraints and practical concerns, a coordinated synchronization architecture is proposed for GEM™ [1]. As illustrate in **Figure 1**, it performs symbol timing and frequency estimation, and link outage detection. Simulation demonstrates that the proposed architecture is functionally comprehensive, and satisfies the design requirement.

In the following sections, the proposed architecture is described with simulation examples. The emphasis is focused on the tracking procedures and the adjustment instructed by outage detection.

2. Symbol Timing Estimation

The square timing estimator [2] is simple for implementation. However the application is usually hindered by the problem of "symbol cycle ambiguity", which occurs when the unknown timing offset is close to half symbol cycle. This problem becomes serious when the signal-to-noise (SNR) is low. One way to alleviate this disturbing phenomenon is using Kalman filter prior to the phasor inversion [2]. On the other hand, direct estimation of timing from each individual burst is actually a noisy observation of a deterministic process. To achieve optimal estimation and eliminate the occurrence of "symbol cycle ambiguity", the square timing estimator is modified and

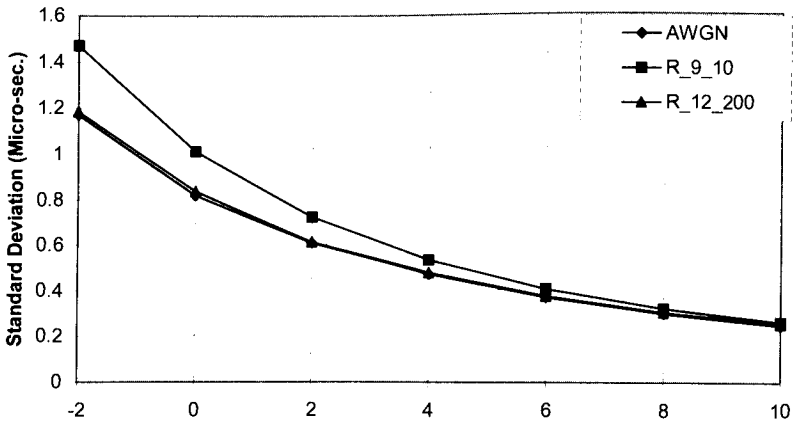


Figure 3. STD of Timing Estimation

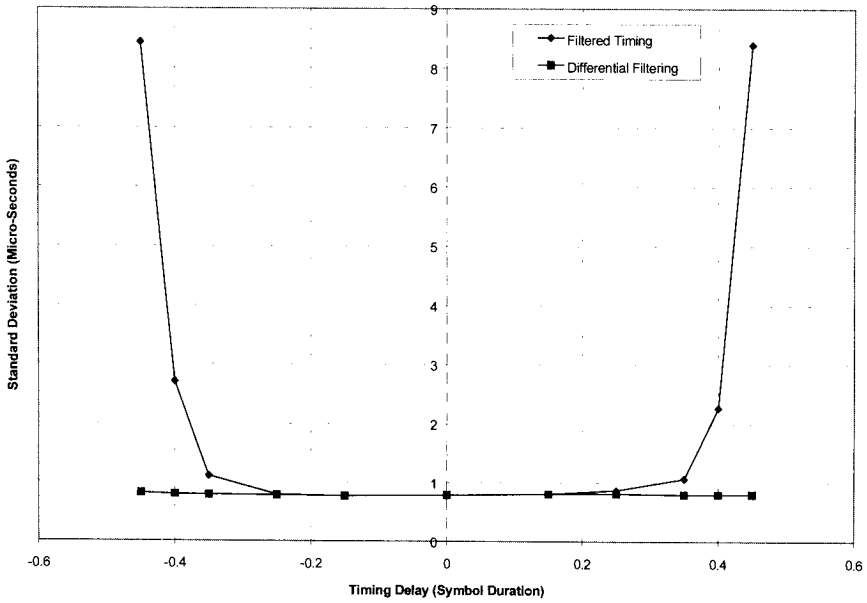


Figure 4. STD of Estimation vs. Timing Offset

Kalman filtering is applied [3][4] in a different manner from that proposed in [2].

Instead of symbol timing τ, differential timing $\hat{\delta}_k$ is estimated from each individual burst. Based on this differential quantity, the unknown timing is obtained by Kalman filtering,

$$\hat{\tau}_k = f\left(\hat{\delta}_k, \hat{\tau}_{k-1}\right) \tag{1}$$

The transfer function of the above procedure is,

$$H_t(z) = \frac{\gamma z}{z-1} \tag{2}$$

where γ is the parameter which determines the filter both bandwidth and gain.

As given in (2), h_t is actually an IIR low-pass filter. Different bandwidths are needed for acquisition and tracking respectively. A smooth transition from acquisition to tracking is controlled by a variable γ

$$\gamma(i) = \rho\left(1-\gamma_{steady}\right)\cdot \rho_0^{\xi \cdot i} + \gamma_{steady}, \quad i = 0, \cdots\cdots \tag{3}$$

where i is the burst index. In **Figure 2**, three trajectories are shown in shown. γ_I is used for filter initialization. It starts with 1 such that the filter is initialized with direct estimation. With γ_I exponentially decreased, the filter output quickly settles down to the optimal estimation. γ_{II} is identical to γ_I except its smaller initial value. This is used to gradually settle a transition discrepancy from relatively poorer estimation to that of more confidence. γ_{III} is a constant for static tracking. Another optional value of γ is 0, which is used during outage as described in section 4.2. In general, various filtering can be easily implemented by selecting different profiles of γ.

Standard deviation (STD) of estimation under AWGN and two Rician fading channel conditions are as shown in **Figure 3**. The Rician channels are designated by the notation R_k_b, where k is the Rice factor and b is the fading bandwidth. The advantage of using differential filter (2) is illustrated by STD of estimation vs. timing offset in **Figure 4**. It can be seen that the STD with no differential filtering increases when the timing offset is bigger than 0.25T. In contrast, the STD with differential estimator is independent of actual timing offset. As long as the burst-to-burst timing variation is less than half symbol cycle, which is always true in practical application, there is no occurrence of "symbol cycle ambiguity".

3. Frequency Estimation

For frequency estimation, the method based on the discrete Fourier transform (DFT) is well known of its superior overall performance. In many applications, however, the use of DFT is prohibited by its lengthy computation. This is especially true when the unknown frequency has large uncertainty range as in GEM™

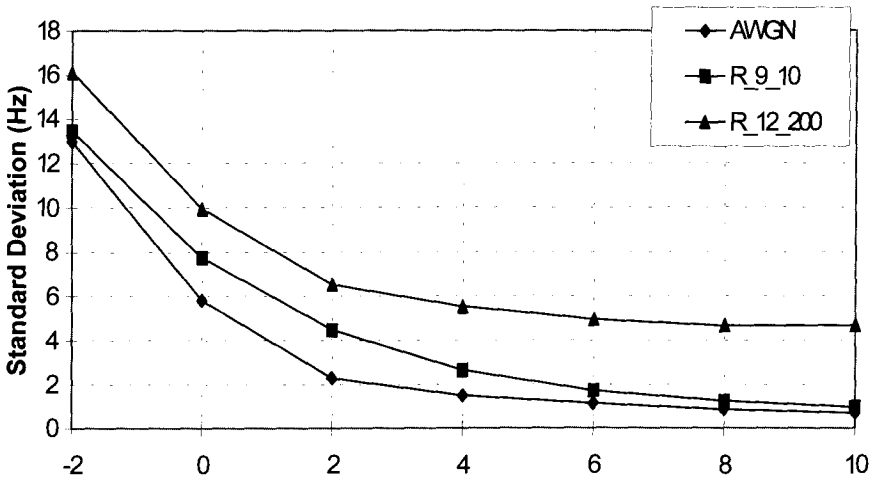


Figure 5. STD of Frequency Estimation

application. If DFT is not performed in a sufficiently wide bandwidth, frequency synchronization is lost.

To achieve the superb DFT estimation performance with reduced computation, the conventional DFT procedure is concatenated with a tracking filter, and frequency search is performed adaptively [6][7].

Let $\Omega(f)$ denote a bandwidth centered at f. The DFT-based frequency search is

$$\tilde{f}_i = \max_{f \in \Omega(\hat{f}_{i-1})} \left| F(s, f) \right| \tag{4}$$

where s is the modulation-stripped signal. $\tilde{f}_i$ is then low-pass filtered, and the final estimation is

$$\hat{f}_i = g\left(\tilde{f}_i, \hat{f}_{i-1}\right) \tag{5}$$

The transfer function of the above filtering procedure is,

$$H_f(z) = \frac{\gamma z}{z - (1-\gamma)} \tag{6}$$

Similar to (2), parameter γ determines both bandwidth and gain.

With a confined bandwidth $\Omega(f)$, computation of conventional DFT is effectively reduced. On the other hand, sufficient $\Omega(f)$ must be maintained to cover any possible burst-to-burst frequency variation. Depending on the burst duration, the tracking procedure can have a pull-in range larger than $\Omega(f)$ [8][9]. Therefore $\Omega(f)$ is not necessarily wider than the maximum burst-to-burst frequency variation.

The filtered estimation process of (4) and (5), called closed-loop estimation, is actually an auto-regressive tracking process. The initial state f_0 is determined by majority-vote among the estimates from the first N<20 bursts.

In addition to the closed-loop estimation, (4) and (5) is implemented in an open-loop manner. By selecting γ=0, this option is only used during outage, as described in section 4.2.

The STD curves of frequency estimation under AWGN and two Rician fading channel conditions are shown in **Figure 5**.

4. Link Outage Detection and Synchronization Maintenance

This procedure performs two functions: outage detection and coordinating synchronization.

4.1 Outage Detection

During normal transmission, the estimated timing is characteristic of a deterministic process. In contrast, when an outage occurs, the estimation process exhibits a pattern of "random walk".

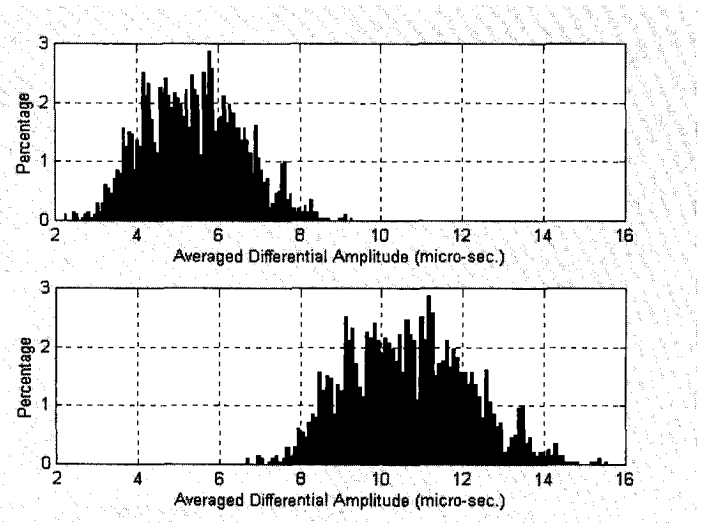


Figure 6. Histograms of Amplitude of Differential Timing

Quantitatively a "random walk" has the feature of large variance. The piece-wise variance can actually be approximated by a detection metric

$$\hat{\Phi} = E\left[\left|\hat{\delta}_i\right|\right] \tag{7}$$

where δ_k is the differential timing which is readily available in the differential timing estimation as described in section 1. $E[]$ is a multiple observation procedure. Two typical histograms of the detection metric Φ are as shown in **Figure 6**, where $E[]$ is performed over 20 bursts. The upper portion is of normal transmission at the lower end of operation range while the lower portion is of outage.

It is noted that the two histograms in **Figure 6** are not completely separated in distribution. A status-dependent dual-threshold detection procedure is thus applied. Let χ_1 and χ_2 denote two pre-determined thresholds, and $\chi_1 > \chi_2$. A binary state variable $D(i)$ is defined as 0 for outage and 1 for normal transmission. During normal transmission, χ_1 is used to detect outage,

$$\text{When } D(i-1)=1, \quad D(i)=\begin{cases} 0, & if \left|\hat{\Phi}(i)\right| > \chi_1 \\ 1, & if \left|\hat{\Phi}(i)\right| \le \chi_1 \end{cases} \tag{8}$$

and the χ_1 is used during outage,

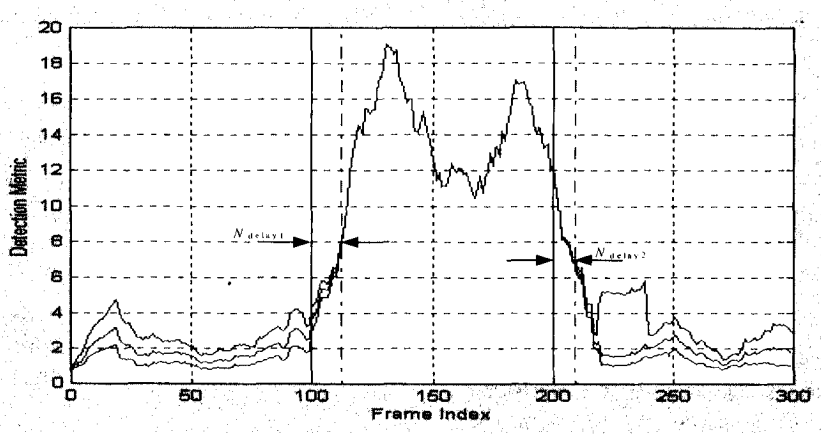


Figure 7. Profile of Detection Metric

$$\text{When } D(i-1)=0, \quad D(i)=\begin{cases}0, & if\ |\hat{\Phi}(i)|>\chi_2\\ 1, & if\ |\hat{\Phi}(i)|\le\chi_2\end{cases} \quad (9)$$

Three detection profiles of metric $\hat{\Phi}(i)$ over one outage are shown in **Figure 7**. They are obtained under Rician fading channel with Eb/N0=-0.5 dB, 2 dB, and 5 dB respectively. The true occurrence of outage is indicated by a pair of vertical solid lines. With χ_1=8 and χ_2=7, the actual detection of the start and end of the outage is as shown by the alternating-dot-dash vertical lines.

In **Figure 7**, N_{delay1} and N_{delay2} are the detection delays due to statistical averaging. Simulation shows the average delays are $\overline{N}_{delay1}=13$ and $\overline{N}_{delay2}=10$ bursts.

4.2 Maintenance during Outage

Based on the detection status $D(i)$, both timing and frequency estimators perform either dynamic tracking or static maintenance, as illustrated in the diagram **Figure 8**. Dynamic tracking is the estimation process during normal transmission as described in section 2 and 3. Static maintenance is a special procedure to prevent no-target tracking during outage.

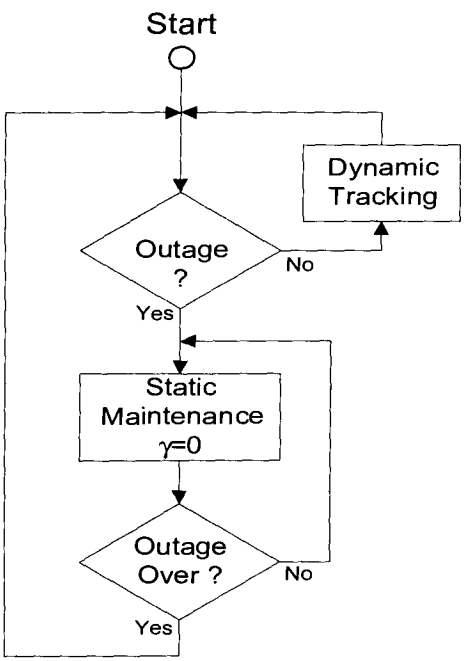


Figure 8. Flow Chart of Synchronization Control

When outage is detected by (8), dynamic tracking procedures are halted, and static maintenance is performed by setting γ=0 in (1) and (5). That is, the estimates prior to outage is taken as the temporary tracking target. This process continues until the end of outage. With the detection switched by (9), regular tracking process resumes for synchronization of normal transmission.

5. Conclusion

The presented synchronization architecture with its tracking and detection procedures is designed for practical implementation and channel conditions in Geo-Mobile applications. Its simplicity, flexibility, and robustness have been demonstrated by simulation. It has potential to be extended, partially or as a whole, to other related applications.

Acknowledgment

The author is grateful to Y. Antia, L. Shi and J. Dai for discussions on the related topics. Some of the materials presented in this paper are the subjects of pending patent applications.

References

[1] Wang, L., "Synchronization Maintenance and Control of GEM Traffic Channel," Hughes Network Systems Internal Memo, Sept. 29, 1998.

[2] Oerder, M., and H. Meyr, "Digital Filter and Square Timing Recovery," *IEEE* Trans. on Commmunications, Vol. 36, No. 5, May 1988.

[3] Wang, L., "A New Symbol Timing Estimation Scheme," Hughes Network Systems Internal Memo, April 29, 1998.

[4] Wang, L., "Performance Analysis of a differential recursive loop for symbol timing estimation," Hughes Network Systems Internal Memo, May 4, 1998.

[5] Haykin, S., Adaptive Filter Theory, Prentice-Hall, 1996.

[6] Wang, L., "An Extended DFT Frequency Estimation Scheme," Hughes Network Systems Internal Memo, April 17, 1998.

[7] Wang, L., "The Optimum Tracking Capability of the Recursive Loop in Frequency Estimation," Hughes Network Systems Internal Memo, May 18, 1998.

[8] Oppenheim, A.V., and R.W. Schafer, Discrete-Time Signal Processing, Prentice-Hall, 1989.

[9] Wang, L., "Burst-to-burst Frequency Error," Hughes Network Systems Internal Memo, Jan. 6, 1999.

Novel Dual Keep Alive Burst in the GEM™ System

Jerry Qingyuan Dai

Hughes Network Systems
11717 Exploration Lane
Germantown, MD 20876
Tel: (301) 601-2639
Email: jdai@hns.com

ABSTRACT

A great variety of information must be transmitted between the Gateway (GW) and the User terminal (UT) in the GEM™ System, specifically, user data and control signaling. This information is mapped onto the physical TDMA bursts. A novel Dual Keep Alive Burst (DKABs) is transmitted during periods of speech inactivity and is proposed by the GEM™ System to save the battery life, satellite power, reduce cochannel interference, add comfort noise, estimate the interference level, maintain the power control and timing/frequency synchronization.

This paper will discuss the DKABS' estimation and demodulation. The Non-data aided Tone Estimation algorithm is used for the timing synchronization. Because of the nature of the differential modulation, the demodulated data in DKABS does not depend on the frequency estimation performance, and the data-aided frequency estimation is chosen for DKABS frequency estimation. Timing and frequency are synchronized using extremely short DKABS in the large timing/frequency offset and large Doppler frequency channels.

1. DKABS BURST STRUCTURE AND SIGNAL MODEL

The proposed DKAB is a 117 symbol long $\pi/4$ differential binary phase-shift keying (DBPSK) modulation burst. Two KABs are separated as shown in Figure 1.

Each KAB has 4 information symbols (m0-m3) and one reference symbol which is required by the differential modulation. There are guard symbols surrounding the KABS, each guard time contains 2.5 symbols. The DKABs burst structure can be configured by choosing the parameters p1, p2 and M. It can been seen that only 10 DKABs symbols are used for a 117 symbols burst, therefore it saves both satellite power and user terminal's power.

Differential binary phase-shift keying modulation is

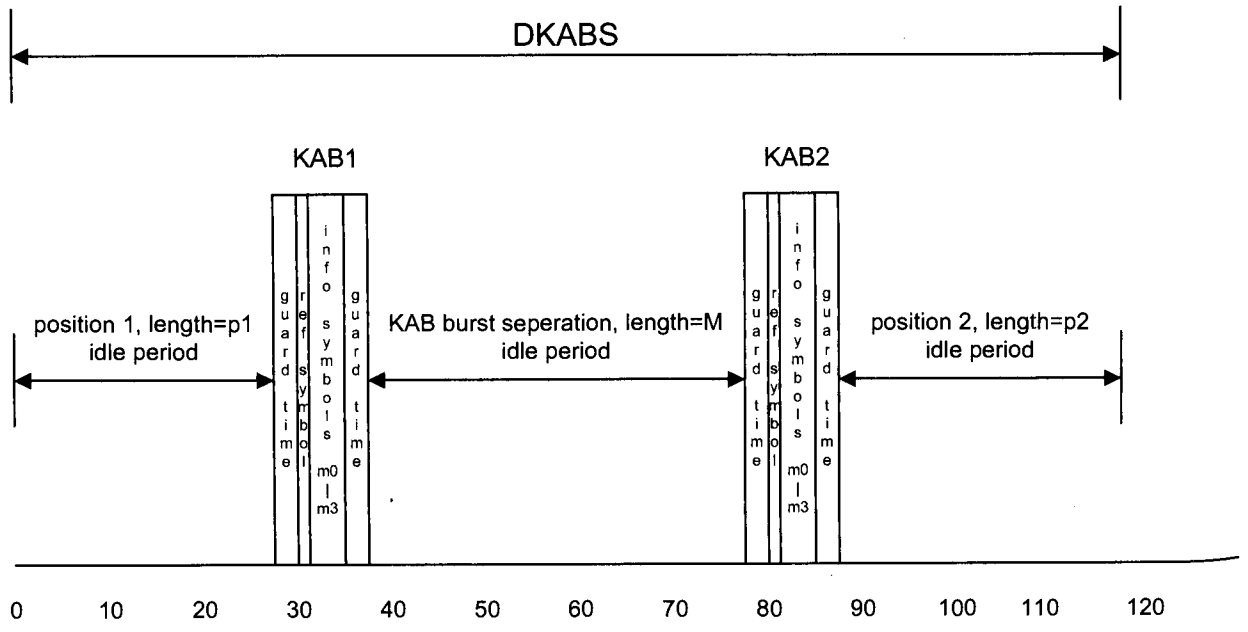


Figure 1 DKABS Burst Structure

selected to combat the frequency uncertainty. The complex envelope of the transmitted signal is defined as follows:

$$x(t) = \sum_k \alpha_k \cdot h(t - kT)$$

where $x(t)$ is the shaping filter α_k is the $\pi/4$-DBPSK coded information symbol defined according to the following table:

Data symbols d_k	Modulation symbols α_k
0	$\alpha_{k-1} e^{j\pi/4}$
1	$\alpha_{k-1} e^{-j3\pi/4}$

The $\pi/4$-DBPSK transmitted symbol can also be represented as

$$\alpha_k = \alpha_{k-1} \cdot e^{j(\pi \cdot d_k + \pi/4)}$$

A digital receiver for DKAB is shown in Figure 2. Since the frequency offset is small ($fT<<1$), timing can be recovered prior to frequency estimation without performance degradation.

As shown in Figure 2, the sampled signal $\{r(kT_s)\}$ is filtered by a matched filter $\{h(kT_s)\}$ as

$$z(kT_s) = \sum_i r(kT_s - iT_s) \cdot h(iT_s)$$

where T_s is sampling period.

The timing estimator estimates the symbol timing error (dt), and the optimal symbol timing is achieved by using the timing interpolation. The Demodulation and Frequency Estimation is implemented separately since the nature of the differential demodulation which does not require the phase information.

2. DKABS DIGITAL PROCESSING

The DKAB digital receiving consists of 3 issues:

- Timing Estimation
- Frequency Estimation
- DKAB Differential Demodulation

The associated digital processing algorithms are discussed in this section.

A. DKAB Timing Estimation

It can be shown [1] that if the matched filter output with timing error is

$$z(lT + \varepsilon T)$$

then the likelihood function for unknown timing error ε is

$$L(\varepsilon) = \sum_l \left| z(lT - \varepsilon T) \right|^2$$

The Maximum likelihood estimate $\hat{\varepsilon}$ is

$$\hat{\tau}_j = -\frac{T}{2\pi} \arg(c_1^j)$$

where

$$c_1^j = \sum_l \frac{1}{M_s} \sum_{k=0}^{M_s - 1} \left| z(lM_s + k)T_s \right|^2 e^{-j(2\pi/M_s)k},$$

$$M_s = T / T_s$$

and j is the burst index.

B. DKABs Frequency Estimation

The frequency offset estimation is required for the system frequency control loop. With the frequency offset, the matched filter output is

$$z_n = \alpha_n \cdot g_0 \cdot e^{j\theta_n} + v_n$$

$$\theta_n = 2\pi \cdot f \cdot n \cdot T + \theta_0$$

where f is the frequency offset and θ_0 is the initial unknown phase.

The DKABs use differential modulation, therefore its BER

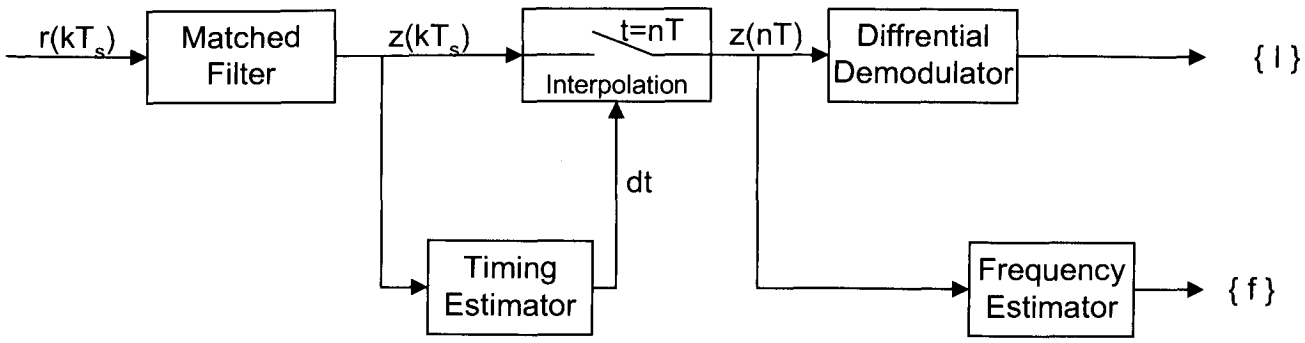


Figure 2 DKABS Digital Receiver Signal Flow

does not depend on the frequency offset. With less than 2% BER at operating SNR (5.5dB), the data aided frequency estimate method is proven to be more attractive.

After demodulation, information symbols are demodulated as $\hat{\alpha}_n$, the modulation data can be removed by

$$y_n = \hat{\alpha}_n^* \cdot z_n$$

$$= \hat{\alpha}^*{}_n \cdot \alpha_n \cdot g_0 \cdot e^{j\theta_n} + \eta_n$$

when $\hat{\alpha}^*{}_n \cdot \alpha_n = 1$, we have

$$y_n = g_0 \cdot e^{j\theta_n} + \eta_n$$

Two phases are estimated independently for two DKABS as

$$\phi_1 = a\tan\left[\frac{\sum \mathrm{Im}(y_n)}{\sum \mathrm{Re}(y_n)}\right], \text{ for KAB1}$$

$$\phi_2 = a\tan\left[\frac{\sum \mathrm{Im}(y_n)}{\sum \mathrm{Re}(y_n)}\right], \text{ for KAB2}$$

The frequency offset can be estimated by measuring the slope of those phases as

$$\hat{f}_j = \frac{\phi_2 - \phi_1}{2\pi \cdot MT}$$

where M is two DKABs distance in symbols from their middle points and j is the jth DKABS.

C. π/4-DBPSK Demodulation

After timing adjusting, the matched filter output is sampled at the symbol rate, *1/T*, as

$$z_n = \alpha_n \cdot g_0 + v_n$$

where

$$\alpha_n = \alpha_{n-1} \cdot e^{j(\pi \cdot d_n + \pi/4)}$$

The decision variable is the phase of

$$U_n = z_n \cdot z_{n-1}^*$$

$$= \alpha_n \cdot \alpha_{n-1}^* \cdot g_0^2 + v_n'$$

$$= e^{j(\pi^* d_n + \pi/4)} \cdot g_0^2 + v_n'$$

Now rotating phase by -π/4, the decision variable becomes

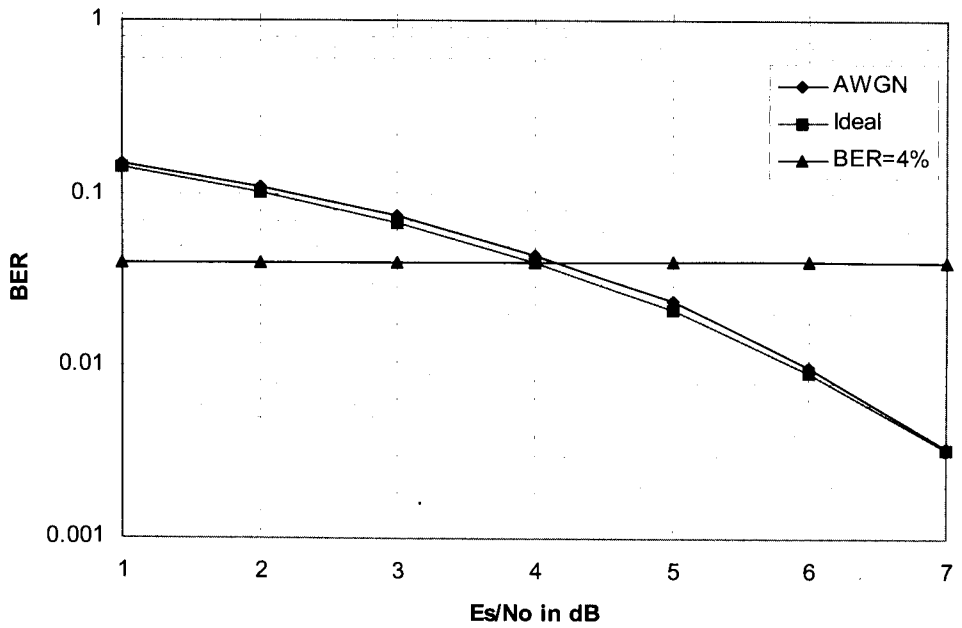
$$V_n = U_n \cdot e^{-j\pi/4}$$

$$= e^{j\pi^* d_n} \cdot g_0^2 + u_n$$

$$= (1 - 2d_n) \cdot g_0^2 + u_n, \; d_n \in (0,1)$$

The decision rule is:

$$\hat{d}_n = 1, \text{ if } V_n < 0$$



$$\hat{d}_n = 0, \text{if } V_n > 0$$

Simulations were done for the DKABS demodulation over the AWGN channel. Figure 3 shows that the DKABS demodulation performance is about same as ideal receiving.

3. CONCLUSIONS

A novel DKABS burst structure and its digital receiving is discussed in this paper. The DKABS can save both the GW and UT power, maintain the frequency and timing synchronization, and carry the power and comfort noise. The simulation results show the DKABS demodulation achieves about the same performance as the ideal demodulation.

REFERENCE

[1] H. Meyr, M. Moeneclay, S. Fechel, "Digital Communication Receivers", John Wiley & Sons, INC, 1998.

Market Trends in Global Satellite Communications - Implications for Canada

Stéphane Lessard
Canadian Space Agency
6767 route de l'Aéroport
Saint-Hubert, Québec J3Y 8Y9
stephane.lessard@space.gc.ca

ABSTRACT

Space-based tele-communications systems will continue to change our lives as we enter the 21st century, increasingly providing access to all types of information anywhere on the planet. Virtual reality entertainment, video on demand, expanded tele-health and tele-education, global corporate networks and mobile consumer devices able to carry converged voice, data and video information are just some of the developments we can expect as the Global Information Infrastructure takes shape.

Space will continue to play a key role in the implementation of this information services revolution. As a result, the global satellite communications market will continue to grow, and change, rapidly. The 1997 WTO agreement on Basic Telecommunications is just one of the factors pushing the fast expansion of this market. Other factors include the growing role played by the private sector in the financing of projects and the provision of services, as well as the on-going industrial restructuring. Satellite constellation projects have fuelled these trends.

Canadian companies can, and are, capturing a significant share of this global market, despite some challenges. The Canadian Space Agency (CSA) and other government agencies are committed to working with their industrial partners to sustain Canadian industrial global competitiveness.

A. Global SatCom market size and growth profile

In 1997, the SatCom market segments were as follows:

- Space segment: US$ 11B (this includes launch and insurance services)
- Ground segment: US$ 18.1B
- Services: US$ 28B

As a result of a variety of factors, future growth is expected to be significant. The market outlook for the period (1996-2006) is as follows:

- Space segment: US$ 60-80B
- Ground segment: US$ 120-150B
- Services: US$ 400B

According to Merrill Lynch, mobile, multimedia and DTH communications will account for 62% of total sector growth by 2007. Between 270 and 350 satellites will be launched between 1998 and 2007, worth US$ 30-39B.

B. Key market characteristics

The SatCom market has the following characteristics:

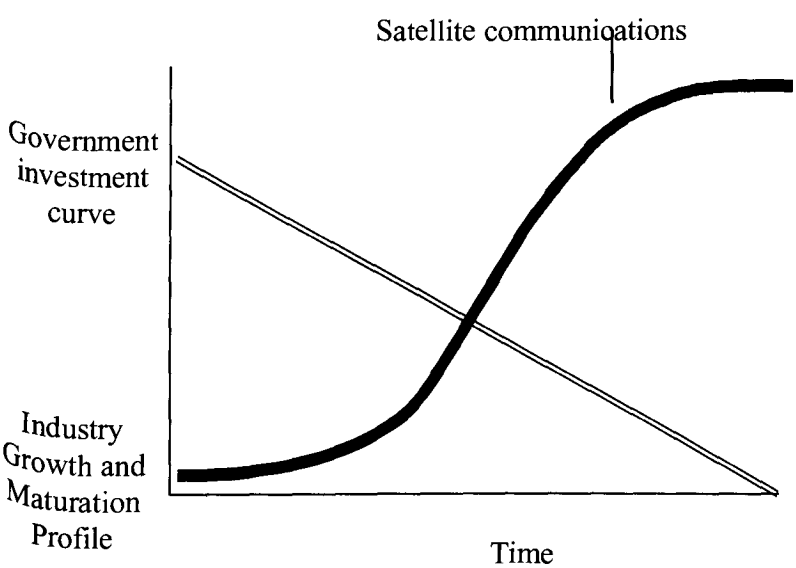
- SatCom constitutes a key part of the very

large total global communications market [1].

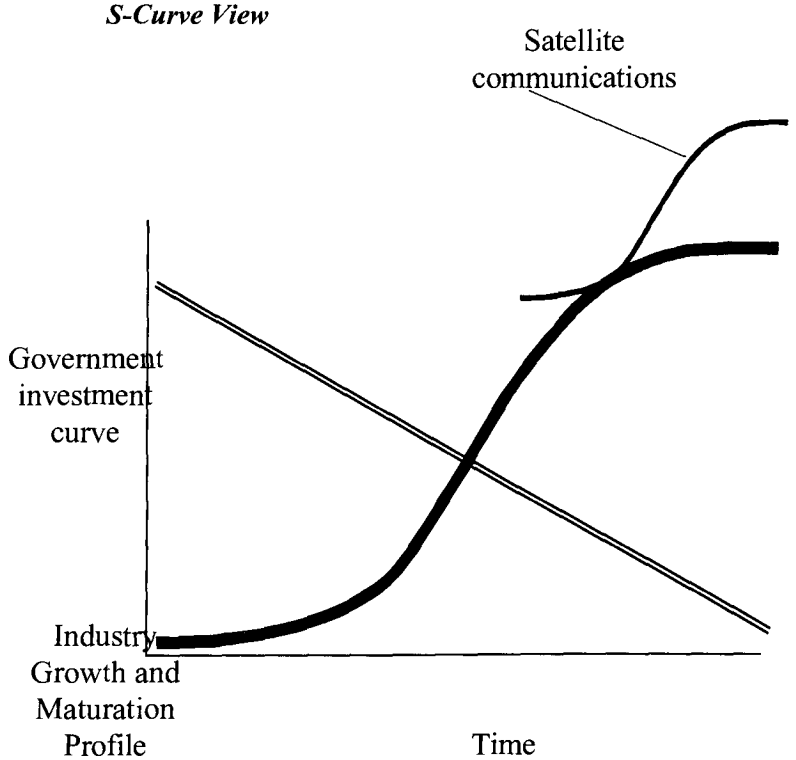
- The satellite market is dominated by a few large US (Boeing, Motorola, Hughes, Loral, Lockheed-Martin) and European (Alcatel-Aérospatiale-Thomson, as well as Matra Marconi-Daimler-Chrysler Aerospace-Alenia) prime contractors.
- Services constitute the largest share of the infrastructure and satellites.
- Canadian firms must compete in a truly global and very fast changing market place.

The SatCom sector is the most commercial and private sector oriented of all space sectors. This could lead to the conclusion that the SatCom sector has reached a high level of maturity (top end of the S-curve shown in the first graph below), and that government investments are less critical. We argue that, on the contrary, the sector is currently at the beginning of a new S-curve (second graph below), in light of the heavy technological, product and applications developments required for multiple new projects aiming at providing new and converged services to consumers. This argues for continued government investment in this sector, in particular for leading edge R&D.

Satellite Communications Industry Life Cycle (Kondriatieff S-Curve): ***Traditional View***



Satellite Communications Industry Life Cycle: ***New S-Curve View***



C. Key change drivers

Major factors driving the fast pace of change in the global SatCom arena include:

- **Users:** new applications are emerging constantly, and interactivity, affordability, globality and mobility are key requirements;
- **Technological convergence:** digitalisation pushes the integration of communications, content (including entertainment) and informatics, creating new, integrated communications platforms and consumer products and services;
- **Globalization:** new global and regional systems are being introduced rapidly;
- **Deregulation:** markets for telecommunications services are being deregulated, globally through the efforts of the World Trade Organization (WTO) and nationally.

D. General market trends

Some of the key general trends which can be observed on the global SatCom scene are as follows:

- **The markets are growing fast (with regional variations)**: Revenue growth to 2005 is expected to be driven by (in descending order): broad band multimedia systems, DBS, GMPCS, with fixed satellite services increasing more steadily. Ground terminals market will expand dramatically, as a result of strong demand for terminals in the DBS, mobile telephony, and later on the multimedia markets. Service markets are expanding everywhere, with many new regional and global systems. Strongest growth in the recent past was achieved in the Asian region. Long term forecasts are positive, but over capacity [2] and the regional economic crisis have taken their toll [3].

- **The markets are being deregulated**: the Feb. 15, 1997 WTO Agreement on Basic Telecommunications has promoted global deregulation in the SatCom sector (it has fuelled privatization, as well as reduced monopolies and foreign ownership and service restrictions). But the Agreement does not cover broadcasting (including direct broadcasting [4]). In the US, the markets are very deregulated, but restrictions remain [5]), while Europe is moving towards a fully liberalised, single telecommunications market. In Canada, monopolies are being phased out, and investment rules are relaxed [6].

- **Global communications are being increasingly privatized**: one can see a shift from government-supported consortia (Intelsat-Inmarsat) to private companies like PanAmSat, Orion and DirecTv (Hughes), and Luxembourg's SES (Astra). Increasingly, governments use commercial services, even for the military.

- **The satellite industry is restructuring:** we are witnessing a complete industrial shake up among large prime contractors. This industrial restructuring is well advanced in the US, and two main industrial poles are emerging in Europe, one led by Matra Marconi (with Daimler-Chrysler Aerospace, Alenia and the missile/rocket operations of Aérospatiale) and Alcatel (with the satellite operations of Aerospatiale and Thomson). Mergers and acquisitions may eventually also extend to Asia. All primes are streamlining manufacturing facilities, to maximize economies of scale, and several are diversifying into carriage and direct service provision [7].

- **Constellation projects are changing the face of the SatCom sector:** services (data, messaging, voice, multimedia) are being offered to the global community, in a flurry of low, medium and geostationary orbit projects (so called Little LEOs, Big LEOs and broadband - "Internet-in-the-sky"), which together represent the single largest market opportunity for Canadian industry. These projects call for established prime contractors (Hughes, Lockheed-Martin, Loral and Alcatel e.g.) to become service providers. Constellations projects fuel the change to the mass production of components by manufacturers, and lead to a loss of projects' clear national identity, as investors, suppliers, service providers and financiers come from around the world.

- **The capital markets have been favourable to commercial projects, but with caution recently:** Since 1993, more than US$ 19B has been invested by the money markets into satellite ventures (much of this is attributable to constellations ventures). However, commercial and technical uncertainties over Iridium and other Big LEOs have made it more difficult for companies to raise capital in the last half of 1998 and in early 1999[8].

- **Converging towards the Network Society**: Cost decreases in communications and computing, computing power increases (Moore's Law) and the convergence of computers, telecommunications and content (incl. entertainment) via digitalization lead to a new information era.

- **The role of satellites in the networks of the future:** Satellites are not the only game in town to provide advanced communications to

customers, with the relentless expansion and upgrades of ground wireless networks, the fast expansion of fibre optic and cable systems, and the introduction by telephone companies of Digital Subscriber Line (DSL) technology. Satellites may have a natural competitive edge over ground networks, for direct digital broadcasting and interactive multimedia service, and as extensions of ground cellular/PCS networks for mobile communications. Key features of satellites are mobility, ubiquity, direct link to users, rapid deployment and reconfiguration of networks, and insensitivity to distance (particularly important in Canada).

E. Overview of sector trends

- **Fixed Satellite Services (FSS) will remain the backbone of regional and global communications**: This is the industry's most mature sector. The next generation of systems will see larger and more powerful satellites. Revenues for FSS services should grow from US$ 9B in 1996 to US$ 25B in 2006.

- **Digital Audio Broadcasting will emerge as a major new market**: Revenues for these services should grow from nothing in 1996 to US$ 48B in 2006. Key players include Worldspace, CD Radio Inc., WCS Radio and American Mobile Radio Corp.

- **Direct Broadcasting Systems (DBS) are growing fast into a large and mature market**: There has been an explosion of services worldwide, with revenues expected to increase from US$ 9B in 1996 to US$ 37.5B by 2007. US firms and European firms lead the way, with Hughes' DirecTv in the US and SES Astra in Europe being the most successful. There is an on-going battle with cable for markets in developed nations (especially urban centres), whereas in developing nations, satellites often provide the most cost effective means of delivering video signals.

  The US is pushing for market entry agreements (has signed with Mexico and Argentina, the latter in June 1998), but the exclusion of broadcasting from major trade agreements (WTO, NAFTA) has allowed Canada to ensure that Canadians receive DBS services from Canadian providers.

- **Global Mobile Personal Communications: unprecedented investments, but uncertain market prospects**: This is the largest single market for satellite, ground and service suppliers [9]. OrbComm and Iridium are already operational, with Globalstar (1999) and ICO (2000) next in line. The US Federal Aviation Administration's conservative scenario expects three Little LEOs and four Big LEOs to be deployed between 1998-2010; the optimistic scenario says four Little LEOs and five Big LEOs. The difficulties being experienced by Iridium in early 1999, and the relentless advance of ground wireless systems, cause some observers to question whether there is a real commercial case to be made for Big LEO projects.

- **Interactive Broadband Satellite Services: more talk than action so far**: Gradually, communications networks in North America will carry more data than voice signals [10]. The current circuit-switched infrastructure (designed to carry voice) is inadequate to accommodate this growth, and incumbent (and new) carriers are examining their options for new infrastructure (cable, DAL, broadband satellites, fibre).

  It is expected that satellites will play an important role in delivering interactive broadband services, in conjunction with ground systems. Pioneer Consulting estimates that more than US$ 76 B will be invested in broadband satellite systems between 1998 and 2010. By 2007, these systems should serve more than 40 M subscribers, or 29% of the total market, for total revenues (residential and business) of about US$ 75B[11].

  This explosive growth fuelled by demand for high speed interactive multi-channel TV, Internet browsing, movie and software down loading, personal telecommerce, and telepresence (teleworking, teleconferencing, tele-education, tele-medicine, virtual library).

  The US FAA's conservative scenario expects

two global LEO broadband systems to be deployed between 1998-2010; optimistic scenario says three can survive on the market.

Because of market pressures, broadband projects are susceptible to mergers (e.g. the unofficial merge of Teledesic and Celestri) or abandonment - we can expect Skybridge and Cyberstar to merge (common parentage of Loral and Alcatel).

F. Assessment of Canada's position on the world's SatCom stage

- **Canada's position in the global market for satellite sub-systems**: Canadian companies have recognized competitive capabilities as sub-system suppliers (EMS Technologies Canada [12], COM DEV, e.g.). Some Canadian companies have built good stable relationships with key US, European primes. Canadian companies must pursue their efforts, with a focus on long term supply agreements, clear product strategies and "commodity" production.

- **Canada's position in the global market for ground terminals and systems**: Canadian companies are well positioned in this market (Nortel, Spacebridge, Telesat, Nanowave, Norsat [13], e.g.). The next big market could be Internet access.

- **Canada's position in the global market for applications and services**: Telesat is Canada's domestic satellite carrier, and the world's 10th operator on revenue. It owns and operates the Anik satellites. It will soon launch the Nimiq DBS satellite. Through subsidiary TMI, it owns and operates the MSAT satellite. Telesat provides consulting services to clients worldwide. Teleglobe is Canada's primary international communications provider, using satellites and cable to provide intercontinental connections. Telesat and Teleglobe face challenges resulting from the loss of their domestic monopolies. ExpressVu and Star Choice are making good progress enlisting subscribers to DTH service in Canada [14], despite strong competition from cable and a continuing grey market for the reception of US programming. There is good, and increasing, participation by Canadian companies in constellation projects [15].

- **Industry-Government partnerships**: The Canadian Space Agency is committed to working with its industrial partners to sustain their global competitiveness, through advice, strategic R&D and information, and market positioning support. Other government players can also make a difference: IC/CRC, EDC, DFAIT, DND. Close coordination among government, and between government and industry, is key to success on world markets.

REFERENCES

[1] The US FCC estimates the worldwide telecommunications market will reach US$ 1 trillion by 2001, up from US$ 550B in 1996. It also predicts that the space telecommunications market will reach US$ 41B by 2001, or just 4% of the overall market. Growth is fuelled by the development of the Global Information Infrastructure (GII).

[2] Since 1993, the number of satellite transponders has nearly tripled to more than 1,100, and in 1997, a record 14 satellites carrying 250 new transponders were launched into orbit above Asia. Transponder leases, which typically cost $1 M-$4 M/ year, have dropped in price by 20%-30% in Q1 and 2, 1998.

[3] Bangkok-based Asia Broadcasting and Communications Network Public Co. cancelled the purchase of 2 L-Star DTH satellites from Loral, while Pasifik Nusantara of Jakarta cancelled the purchase of the Multi Media Asia (M2A) satellite. Loral cut 300 jobs as a result. Asia Satellite telecommunications Holdings reported in August 1998 a 15.2% fall in profit.

[4] However, the Agreement is important to broadcasters because new emerging technologies from telephone companies or other suppliers can potentially change the face of broadcasting. Also, there are indications that the new WTO Round (to be launched by January 1, 2000) could include Audio-Visual Services, which would cover broadcasting.

[5] In November '97, the FCC did away with ECO and DISCO 2 tests, opening further access for WTO countries. But the "public interest" test remains (i.e. systems licensed by WTO member states will be subject to a review that includes analysis of national security, law enforcement and trade policy implications).

[6] Teleglobe lost its monopoly in October 1998, and Telesat will lose its monopoly in March 2000. Foreign ownership limits for Teleglobe have been lowered to 47.6%.

[7] For example, Matra Marconi with EAST or WEST, Motorola with Iridium, Loral with Globalstar, Loral Skynet and Loral Orion, Hughes with DirecTv and Panamsat.

[8] In July 1998, OrbComm postponed its IPO because of low demand; also, ICO global's August 1998 IPO only raised less than one third of the expected sum (US$ 120M rather than the expected US$ 380M).

[9] 800 of 1,100 satellites expected to be launched in next 10 years will be for mobile communications; by 2003, GMPCS terminals could account for 43% of total ground segment sales, or US$ 12.1B; services revenues expected to grow from zero in 1998 to US$ 31.6B by 2007.

[10] Expectations are that, by 2001, data will represent over 75% of all telecommunications traffic in the US.

[11] By 2007, broadband satellites are expected to have a share of multimedia traffic about equal to cable and DSL, according to Pioneer Consulting.

[12] Grouping the Ste-Anne-de-Bellevue facilities formerly belonging to Spar Aerospace, and the company known until recently as CAL Corp.

[13] Nortel and Spar won a major contract in December 1998 from SES to provide the Ka/Ku band ground segment for the Astra system.

[14] ExpressVu had about 125,000 customers in October 1998, while Star Choice had about 110,000 at the end of August 1998. Wood Gundy estimates that the two satellite operations will attract 1.2 million subscribers by 2003.

[15] BCE, Telesat, COM DEV and Stratos Mobile in Iridium, CanCom in Globalstar, and Spar and COM DEV in Skybridge, for example.

S-UMTS in the Wireless Information Society: The Challenges Ahead.

(The views expressed herein are those of the authors, and do not necessarily reflect the views of the European Commission)

B. Barani, J. Schwarz da Silva, J. Pereira, B. Arroyo-Fernández, D. Ikonomou.
European Commission , DG XIII-F.4
Avenue de Beaulieu, 9 B-1160, Brussels, Belgium.
Bernard.Barani@bxl.dg13.cec.be

ABSTRACT

Telecommunication is emerging as a global business and as an important factor in the globalisation and networking of economic activities. The rapid development and increased use of communication networks and technologies are underpinning globalisation of the economy and reinforcing these trends. Bandwidth, network connectivity and global access are considered to be essential ingredients of the new industrialised world, with telecommunications networks accelerating the scope of an increasingly border-less world economy in key sectors, by shifting an increasing proportion of economic activity on-line and fostering the emergence of the Information Society.

From the technological point of view, there is no doubt that wireless technologies will play a major role in the successful implementation of the Information Society. At the end of 1997, at world-wide level the number of subscribers to mobile communications systems was estimated to be 200 Million with projections pointing to a net addition of about 700 new Million subscribers in the period of the next 5 years.

In Europe, digital mobile communications, in particular through GSM and DCS-1800, are providing pan-European mobility in personal communications at unprecedented levels. On a world scale there were at the end of 1997 some 70 Million subscribers growing at a rate of close to 3 Million every month. Today, more than 200 GSM networks are in live commercial operation in over 100 countries world-wide.

Building up on this success, the next window of opportunity for growth of the mobile and wireless industry resides in the development of true broadband mobile systems, known as UMTS in Europe or IMT.2000 in its ITU version. The basic premise upon which work is being carried out, is that by the turn of the century, the requirements of the mobile users will have evolved: they will wish to avail themselves of the full range of broadband multimedia services provided by the global information infrastructure, whether wired or wireless connected.

In Europe, the ground work for UMTS started in 1990. R&D programmes sponsored by the European Union such as RACE (1990-1994) and ACTS (1994-1998) played a major role in developing and experimenting the technology, including satellite UMTS access. ACTS work culminated at the January 98 meeting of ETSI, that endorsed UMTS terrestrial standards developed under the ACTS framework. This European proposal has been submitted to ITU in the context of the IMT-2000 family of standards. In addition, the European commission was instrumental in the creation of the UMTS Forum, the now world recognised forum for next generation mobile communication developments. Finally, the European Commission has put in place the regulatory framework that will allow the first UMTS licenses to be awarded in 2002, with the corresponding UMTS Decision approved by the European Council and Parliament early 99.

How do mobile satellite systems fit in that picture?
Satellite communication systems have an inherent capability to fulfil the requirements of the Information Society deployment, in particular through their global access capability. In spite of these promises, it can be noted that the satellite situation regarding UMTS/IMT-2000 is quite different from that of the terrestrial systems. If Europe has been able to generate a consensus on a terrestrial UMTS standard which is widely supported by key US and Japanese players, the same can not be said for satellite UMTS.

Before such a satellite based system can be successfully deployed, there are still a number of issues to further investigate, such as:

- What market for S-UMTS, and what positioning strategy;
- What are the spectrum requirements;
- Is the regulatory framework adapted;
- What are the missing technological bits;
- How can R&D help.

Against this background, the paper reviews the situation, with a particular focus on Europe, in relations to the above open questions. It also provides an outline of the S-UMTS supporting actions expected to be implemented in Europe in the context of the 5th Framework R&D Programme of the European Union (1999-2002).

MARKET PERSPECTIVES

Market perspectives for Multimedia Mobile Services have been extensively carried out in Europe through the UMTS Forum. Studies of the detailed market potential have concentrated on two aspects:

- The volume of revenues that may justify operator's investment in third generation systems;
- The volume and characteristics of data streams with a view to defining the medium to longer term spectrum requirements.

Markets for Terrestrial Systems

Before reviewing the potential markets for satellite based mobile multimedia, it is useful to consider the figures outlined for the terrestrial markets. In Table 1, the world-wide market forecast for the physical users of terrestrial mobile services including multimedia is outlined.

Mobile users in millions at year end	2000	2005	2010
Europe, EU15	113	200	260
North America	127	190	220
Asia Pacific	149	400	850
Rest of the world	37	150	400
Total	426	940	1730

Table 1: World-wide Mobile Market Forecast [1]

The figures outlined in above tables show that significant growth of the mobile markets are expected over the coming decade. Europe, Japan and North America are regions that will most likely face market saturation in terms of physical users by the year 2010. On the other hand, many countries in Asia Pacific/Africa and South America are expected still to be far from reaching saturation in terms of mobile users in the year 2010.

The UMTS Forum has paid particular attention to the market situation of Europe, with the objective of using the results as a basis of the traffic model for the spectrum calculations. Table 2 shows the market forecasts for the number of physical users of mobile services and, out of them, the number of physical users of mobile multimedia (MM) services and their assumed penetration rates.

Year	Population in millions	Physical users in millions	Penetration	Thereof users of MM service (millions)	Penetration
2005	385	200	0.52	32 (from which 20 use High MM)	0.08 (0.05 for High MM)
2010	387	260	0.67	90	0.23

Table 2: EU 15, Penetration of Future Mobile Services

For the multimedia type users, three types of traffic have been distinguished, as outlined in table 3. Building on these figures and assumptions, the following results were achieved for the European prediction in Y2005:

Total mobile market:

Users: *200 million*
Service revenues: *104 billion ECU per year*
Traffic: *6320 million Mbytes/month -- 32 Mbytes/user/month*

Mobile multimedia segment:

Users: *32 million*
Service revenues: *24 billion ECU + 10 billion for terminal revenues*
Traffic: *3800 million Mbytes/month -- 119 Mbytes/user/month*

The UMTS Forum therefore conclude that the annual market revenues in Europe for mobile multimedia will be at least 34 billion Euro (services and terminals) by the year 2005 with at least 32 million physical users using mobile multimedia services. In these figures the uses of enhanced 2nd generation mobile services are included.

From that perspective, it appears that mobile multimedia represent 16% of the users and 23% of the revenues in the year 2005. Traffic requirements of that market segment will represent 60% of the total. Every multimedia user will generate significantly more traffic than today's mobile user, but they cannot be expected to pay an equivalent multiple of current tariffs, which implies that tariffs will not be proportional to the traffic volume or to the used spectrum. This is one of the major challenge facing the communications and computing sectors for the successful roll out of terrestrial broadband mobile systems.

Market for Satellite Systems

The UMTS Forum has also worked towards the definition of potential markets for MSS systems offering broadband multimedia services. The underlying assumption is that less than 20% of the world's land area will be covered by terrestrial cellular networks within the envisaged time scale of UMTS/IMT-2000. Satellite systems, which provide world wide or large regional coverage may thus have the potential to complement terrestrial UMTS/IMT-2000 to provide complete coverage. In turn, this may also spur further demand for terrestrially based UMTS/IMT-2000 services. In the specific case of the EU15, more than 80% of the population can be expected to be covered by terrestrial UMTS/IMT-2000 in 2010. Price differences between terrestrial and satellite services will play an important role in

Services	User net bit rate [kbps]	Coding factor	Asymmetry factors	Call duration (effective) [s]	Service bandwidth[1] [kbps]	UMTS Switch Mode[2]
High interactive MM	128	2	1/1	180 (144)	256/256	CS
High Multimedia (MM)	2000	2	0.005/1	53	20/4000	PS
Medium Multimedia	384	2	0.026/1	14	20/768	PS
Switched data	14,4	3	1/1	156	43,2/43,2	CS
Simple messaging	14,4	2	1/1	30	28,8/28,8	PS
Speech	16	1.75	1/1	120 (60)	28,8/28,8	CS

[1] The service bandwidth is the product of columns 2, 3 and 4
[2] CS is circuit switched and PS is packet switched

Table 3: Services classes and attributes

the usage of these services, but the difference in price is not likely to significantly change in the future.

Market analysis for MSS has been supported by MSS operators. The classes of services that have been considered are slightly different from those considered for the analysis of the terrestrial market. MSS services have been categorised as follows:

1. Speech - quality basic speech at 8/16 kbit/s
2. Low-speed data - predominantly messaging and e-mail (without attachments) type services at 9.6/16 kbit/s
3. Asymmetric services - this includes the predominantly one way services including file transfer, database/LAN access, Intranet/Internet WWW, E-mail (with attachments), image transfer etc. Rates of transmission will be up to around 144 kbit/s. This corresponds approximately to the Medium (and High) Multimedia services defined for terrestrial UMTS/IMT-2000
4. Interactive Multimedia - predominantly relating to videoconferencing and videotelephony at data speeds of around 144 kbit/s. This corresponds approximately to the High Multimedia services as defined for terrestrial UMTS/IMT-2000.

Classes of services 1&2 will not be offered by first generation of S-PCS systems but have been classified as 'Non Multimedia'. Classes of services 3&4 have the potential to offer broadband services and have been classified as 'Multimedia'. Compared to the terrestrial scenarios, it has been considered in the analysis that the maximum offered bit rate at S band would be of 144 kbit/s, whilst the terrestrial systems are conceived with a maximum bit rate of 2Mbit/s. This is also a difference with some other analysis outlined in ITU TG 8/1 documents, considering that satellite services could be as high as 384 kbit/s.

The forecasts further distinguished between 12 traffic environments, which are hereafter listed:

1. Rural pedestrian (handheld, portable, transportable)
2. Rural vehicular (car, truck, train, bus)
3. Rural fixed
4. Remote pedestrian
5. Remote vehicular
6. Remote fixed.
7. Off-shore maritime (yacht, tug boat)
8. Deep sea maritime (freighter, tanker, ocean liners)
9. Personal/Corporate aeronautical
10. Passenger aeronautical
11. Localised base station (a cell covering an indoor area not covered by terrestrial UMTS/IMT-2000, such as a train, bus or building, with an outside repeater to the satellite)
12. Terrestrial fill-in (traffic resulting from coverage of areas which eventually will be covered by terrestrial UMTS/IMT-2000)

It should also be noted that Fixed Satellite Service (FSS) systems (those systems providing services >~ 1Mbit/s to fixed installations) are specifically excluded, as these systems are assumed to be outside the bounds of UMTS/IMT-2000.

Table 4 outlines the results obtained in EU 15 and world-wide, both in terms of potential users and in terms of generated traffic [2]. Users have been segmented into non-multimedia users (those requiring non-multimedia services only, as defined above) and multimedia users. These forecasts have been derived from forecasting the total MSS demand with UMTS/IMT-2000 forming a part of this. It is expected that all the forecast MSS multimedia users will be UMTS/IMT-2000 compliant while only a portion of the MSS non-multimedia users will be UMTS/IMT-2000 compliant.

The demand for UMTS/IMT-2000 satellite services differs from region to region and depends, *inter alia*, on population density and developed terrestrial infrastructure. In the EU15, the demand may decrease in the long term, as countries complete their terrestrial UMTS/IMT-2000 coverage and combine with 2nd generation cellular systems. This effect is not taken into account in traffic figures for 2010 for EU15, where it can be expected that UMTS coverage will be high. Therefore, the demand for 2010 in EU15 may be considered as overestimated.

Year	Worldwide 2005	Worldwide 2010	EU 2005	EU 2010
MSS Subscribers (000s)				
Non-Multimedia	4,875	7,500	609	938
Multimedia	6,585	10,975	395	659
	11,460	18,475	1,004	1,596
Average Usage per subscriber (kBs per month)				
Non-Multimedia				
Voice	8,709	8,491	8,709	8,491
Low Speed Data	6,208	5,587	6,208	5,587
Multimedia				
Voice	1,194	1,561	1,194	1,561
Low Speed Data	2,584	3,380	2,584	3,380
Asymmetric	26,154	34,247	26,154	34,247
Interactive	1,781	2,334	1,781	2,334
Total Annual Traffic (Million MB's)				
Non-Multimedia				
Voice	509	764	64	96
Low Speed Data	491	736	45	63
Multimedia				
Voice	94	206	6	12
Low Speed Data	204	445	12	27
Asymmetric	2,067	4,510	124	271
Interactive	141	307	8	18
Total	3,506	6,968	259	486
Annual Traffic (Mill. MB's) - excluding non UMTS/IMT-2000 compliant traffic				
Non-Multimedia				
Voice	34	123	4	15
Low Speed Data	33	119	3	10
Multimedia				
Voice	94	206	6	12
Low Speed Data	204	445	12	27
Asymmetric	2,067	4,510	124	271
Interactive	141	307	8	18
Total	2,573	5,710	158	354

Table 4: World-wide satellite market and traffic volumes

It can thus be concluded that:

- *The world market for users MSS (including multimedia MSS) will be 11.5 million users by the year 2005, rising to 18.5 million by 2010.*
- *There will be 1 million MSS users in Europe by 2005, rising to 1.6 million by 2010. 0.4 million of these will be multimedia users in 2005, rising to 0.7 million multimedia users in Europe by 2010.*
- *Total traffic levels (multimedia + non-multimedia) for the European MSS market will reach 22 million Mbytes/month in 2005, rising to 40 million Mbytes/month in 2010.*

It can also be noted that an independent study performed by ICO and sponsored by the European Commission [3] concludes with similar results, as exemplified in Figure 1.

Although the market has been segmented slightly differently in the case of the ICO study, the study concludes that the mobile multimedia market over the 43 CEPT countries would be in the order of 400 000 users in 2005, with the assumption that ICO would capture 40% of the total market, i.e. 1 million of users. This figure fits remarkably well with the findings of the UMTS Forum for Europe.

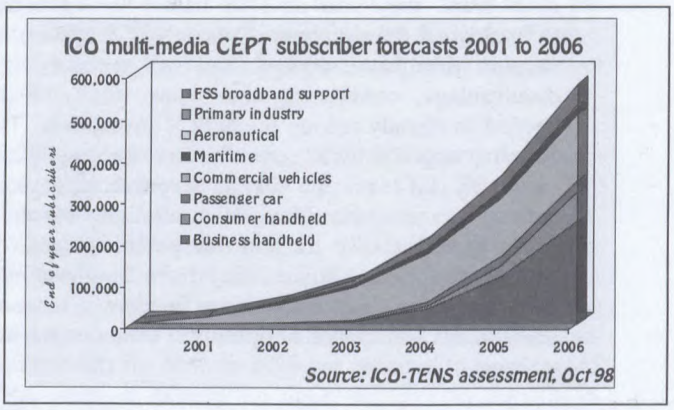


Figure 1: ICO SAT market assessment for mobile multimedia by satellite

Preliminary Conclusions on Markets

From the above figures, the following conclusions regarding S-UMTS can be outlined:

- As in the traditional 2nd generation mobile case, satellite communications are expected to represent a small fraction of the total business. European figures for Y2005 show that the satellite penetration is not expected to represent more than 1/30th of the terrestrial user penetration and not more than 1% in terms of traffic. On this very competitive markets, satellite communication will probably move from a niche market situation, as is the case for second generation system, to a 'niche of a niche' market.

- The small size of the targeted market makes it in turn more difficult and more risky to be positioned on the S-UMTS market. ICO study has shown that almost all the targeted market segment are 'elastic', meaning that sensitivity to price is high. Competition with the terrestrial sector will thus not be easy, should the business models retained for first generation S-PCS systems be applied to S-UMTS.

- The satellite sector is facing a couple more difficulties: first, it has to be outlined that market prediction for the terrestrial mobile sector have so far been systematically underestimated. The reality has shown that the terrestrial sector grows faster than anticipated. The same can not be said for MSS, where the original predictions have still to be validated. Recent examples show that it will not be easy for the satellite industry to live up to their expectations.

- Two types of entrants can be considered for the S-UMTS markets: those already established on the S-PCS market, and new entrants. The former will have the advantage, provided they develop successfully on the S-PCS market, of having already an established customer base. They will thus have the possibility to 'educate' the market and to develop optimised strategies based on real customer experience. On the

other hand, they will have to further invest in the technological development of their S-PCS system, to support wider band services. This will probably be a disadvantage, considering that some time will be needed to already recoup the S-PCS investment. The cost of upgrade will certainly not be negligible, especially if it requires a change in spacecraft/payload technology over the whole constellation, which is likely to be the case for non transparent systems. In that context, first entrants may have less customer experience, but should have more freedom in terms of investments and choice of optimised technologies and systems.

- In that context, it seems that the most viable route for the satellite industry would be to closely associate itself with the terrestrial industry. The 'chance' here is that it will be fairly costly to develop terrestrial UMTS systems, even if the technology is compatible with an evolution of second generation systems such as GSM. First UMTS movers may seriously consider the possibility to deploy UMTS, in a first phase, only in very densely populated areas, such as city centres or business islands. The satellite complement could here play a role to its fullest extent, by providing the coverage complement, not only in developing countries, as has often been the approach for S-PCS systems, but also in developed regions of the world. This however requires from the satellite sector a new approach to the terrestrial sector, and a real willingness to create partnerships with the terrestrial sector, with a 'mutual benefit' and risk sharing approach. This would correspond to a real evolution from the S-PCS situation, where satellite is not considered 'strategic' for the realisation of the terrestrial business cases.

SPECTRUM PERSPECTIVES

There is no doubt that spectrum, particularly at S-band, in the 2GHz region is a scarce resource and that competition from the various players is expected to intensify, especially in the context of the upcoming WRC 2000 Conference. Also, regional approaches are different: Europe has expressed its preference for a priority to terrestrial systems, whilst the USA are traditionally supportive to spectrum allocation to satellite services. A simplified Spectrum situation, as inherited from WRC 92 and further complemented with regional allocations is depicted on figure 2.

In Europe, the UMTS Forum and the CEPT ERC Task Group 1 are currently working on the identification of additional spectrum requirements and on the candidate bands [4], [5]. Based on above market and traffic analysis, the following results have been obtained:

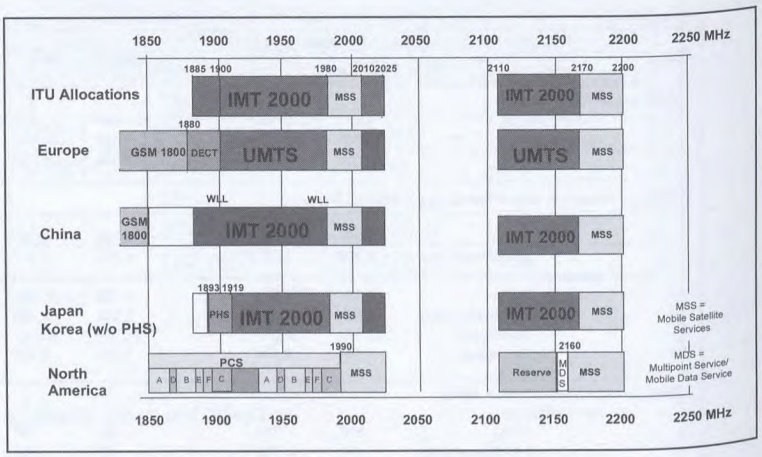


Figure 2: Current 2GHz Spectrum Plans

Terrestrial situation

The spectrum requirements in Europe over the coming ten years period are depicted on figure 3.

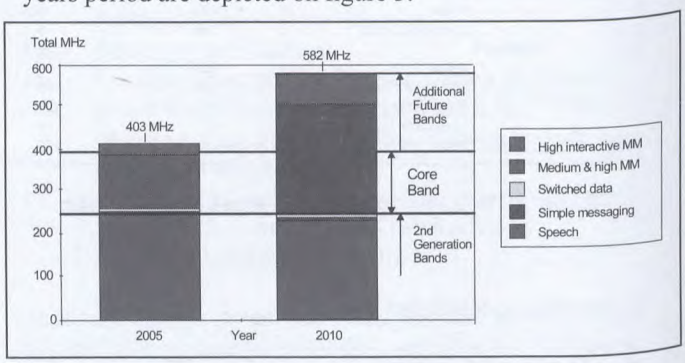


Figure 3: Terrestrial Spectrum Requirements in EU15

The results show that current allocation is sufficient to cover spectrum needs until Y2005. This also assumes that bands already assigned to 2nd generation systems will be used by third generation systems. In the longer term, expected growth of the multimedia traffic clearly shows that additional allocation is required. The results obtained in Europe suggest that an additional spectrum allocation of about 180 MHz is needed.

Satellite Situation

A similar exercise has been carried out for the MSS bands. Results are reported on figures 4 and 5, respectively for the EU15 and world-wide scenarios.

The results have been obtained using the methodology depicted in ITU Recommendation R.M.1391. In this process, it can be noted that the maximum considered carrier rate was of 144 kb/s, with extra coding requirements in the order of 30 kb/s. Each carrier is thus assumed to require around 200 kHz of bandwidth. Number of beam cluster, Fc, is assumed to be 1 for existing systems whilst it has been considered equal to 2 for future systems capable of more spectrum efficient usage. Delay

factor, Fd, is equal to one for non delay tolerant traffic whilst it has been assumed comprised between 2 and 5 for asymmetric data services being unaffected by delays over the transmission network.

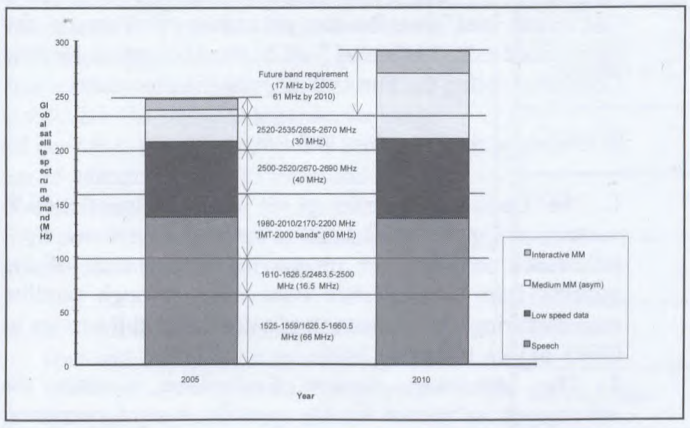


Figure 4: MSS Spectrum Needs, EU 15

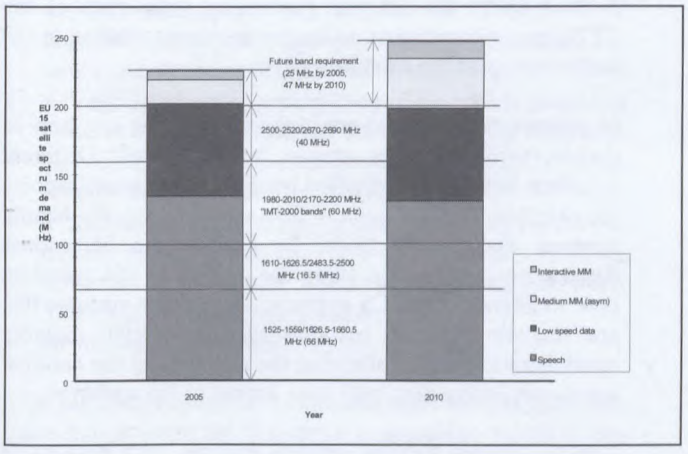


Figure 5: MSS Spectrum Needs, Hot Spot

The above two figures are clearly outlining the additional MSS spectrum requirements in the medium to longer term. These are summarised in table 5.

	EU 15		*Global Hot Spot*	
	Year 2005	**Year 2010**	**Year 2005**	**Year 2010**
Non IMT 2000 MSS	2x93 MHz	2x79 MHz	2x93 MHz	2x79 MHz
IMT 2000 MSS	2x19 MHz	2x44 MHz	2x30 MHz	2x66 MHz
Total	**2x112 MHz**	**2x123 MHz**	**2x123 MHz**	**2x145 MHz**
Additional Spectrum Needs	***25 MHz***	***47 MHz***	***17 MHz***	***61 MHz***

Table 5: Additional MSS Spectrum Needs

The above results obviously differ according to the considered region, as exemplified by the differences of the EU scenario and the Global Hot Spot scenario, based on US type of environment. Compared to other regions, it also has to be noted that the band 2520-2535 MHz and 2655-2670 MHz are not available for MSS in Europe, which explains the larger European additional requirement in 2005.

The requirements outlined by the UMTS Forum have been to a large extent endorsed by the ITU TG 8/1. The proposed text for the CPM report [6] in relation to WRC 2000 agenda item 1.6.1 (IMT 2000) outlines an extra spectrum need of 160 MHz in the three ITU regions for terrestrial systems in 2010 (to be compared with the 180MHz derived by the UMTS Forum) whilst the total requirements for MSS in 2010 are reported to be of 2x145 MHz, which is exactly the figure reported by the UMTS Forum. It is however clear that it will be difficult to accommodate such a request, corresponding in total to an extra 220 MHz across the 400MHz to 3Ghz region.

Candidate bands

The bands currently considered for MSS 'IMT 2000 ' extension include traditional MSS frequency bands such as the 1525-1559/1626.5-1660.5 MHz or 1610-1626.5/2483.5-2500 MHz frequency bands. These bands are already used either by traditional systems or by new global S-PCS systems, and it will thus take some time before they can become available to S-UMTS type systems. Other bands are also contemplated, such as the 2520-2535 MHz or 2655-2670 MHz. These bands present however two difficulties: a) some countries are using these bands for Multipoint Distribution Systems; b) the terrestrial community has also identified these bands as a possible extension for terrestrial systems.

In Europe, the CEPT position favours the usage of these bands for the terrestrial component of IMT 2000, whilst recognising that other regions of the world may want to reserve it for the satellite component. In any case, it is quite unlikely that the extra requirements of the MSS community may be satisfied at the WRC 2000. With the current approach, the total amount dedicated to MSS will remain roughly the same, whilst some allocation would be specifically earmarked for the satellite component of IMT 2000. This situation is further complicated by the large number of satellite systems that have already been advanced published by the ITU for the MSS frequency bands between 1 and 3 GHz.

Currently, there are filings for more than 150 systems in the 1.5/1.6 GHz bands, more than 50 systems in the 1.6/2.4 GHz bands, almost 100 systems in the 2 GHz bands and about 75 systems in the 2.5/2.6 GHz bands. Some of these systems have been filed in more than one of the bands. It is expected that some systems will not be implemented due to financial or other reasons. Nevertheless, the number of

filings demonstrate the very large interest in providing MSS in the 1 – 3 GHz range. In particular, the filings in the 2 GHz region correspond mainly to traditional S-PCS systems, which reduces S-UMTS spectrum accordingly and renders the prospects of deploying true S-UMTS systems even more remote.

ACCESS TO CRITICAL RESOURCES

Access to Spectrum and licenses

Access to critical resources such as spectrum, licenses and markets are key success factors for any wireless telecommunication networks, and in particular for satellite based systems. The specific nature of satellite based communication infrastructures, with coverage naturally extending far beyond the geographical boundaries of a single country, is raising in turn regulatory challenges different from those of terrestrial networks.

The European Commission has for a long time recognised this specificity of satellite based systems, especially for those based on global constellations such as LEO or MEO systems. At the initiative of the European Commission, the EU Council and Parliament adopted in early 97 the Decision 710/97/EC, the 'S-PCS Decision', calling for the co-ordinated introduction of S-PCS in the European Union. The Decision puts in place the necessary instruments to obtain national licenses within a commonly agreed framework. According to its provisions, the Commission mandates CEPT to develop harmonized use of frequencies, harmonized conditions, free movement of terminals, harmonized authorization procedures

The mandate issued to the CEPT covers MSS frequency bands, i.e. the 1.6/2.4 GHz and 1.9/2.1 GHz bands, which thus includes the S-UMTS or S-IMT2000 bands. The Decision also includes the possibility for the European Commission to enforce binding measures, if required. The overall licensing/spectrum access process is depicted on figure 6 .

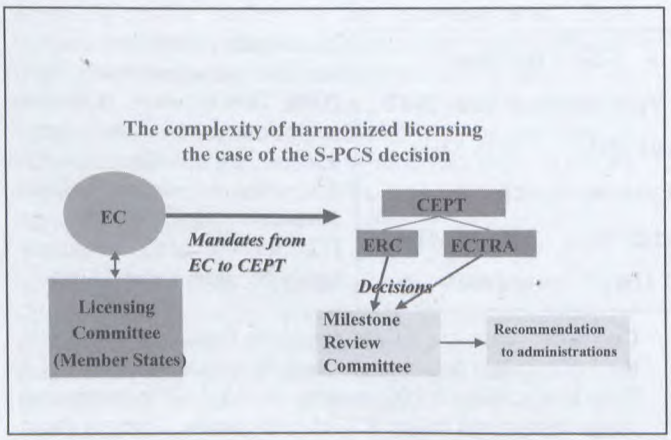


Figure 6: The S-PCS Licensing Process in Europe

In response to the mandate, the CEPT/ERC and CEPT/ECTRA have produced a set of Decisions addressing the authorisation, frequency availability and free circulation of GMPCS systems and. The Decisions are the basis of the ERC/ECTRA Milestones Review Committee and describe the procedure for running the harmonised introduction of S-PCS, the Milestones Review Committee being the body which runs the procedure.

In summary, the procedure is following:

1. The Decisions contain a set of 'milestones' which measure the maturity of satellite systems. There are eight milestones covering the phases of development of the systems from the Advance Publication through satellite manufacturing and the launch of satellites to full service in CEPT before 1 January 2001.
2. The Milestones Review Committee assesses the information submitted by the satellite system operators, i.e. whether the milestones criteria is met, and gives recommendations to the member states of the CEPT regarding the authorisation of that satellite system.
3. It is up to the national regulatory authorities of the CEPT member states to make any decisions relating to the authorisation of the satellite system.

In general, the principal questions of licensing are: how to decide, which systems are to be authorised. Different solutions have been identified by different administrations, for example the first come - first served basis, the beauty contests and the auctions. In Europe the Milestones Review process is seen to be the answer to the question, how to licence GMPCS systems. Only those systems that are realistic will be recommended to CEPT member administrations and, following the decision of the national regulatory authorities, will have access to the spectrum.

The Milestones Review process ensures with reasonable certainty, that the same satellite systems obtain access to the same piece of spectrum in CEPT member countries. It also makes it easy to avoid wasting spectrum to paper satellites. The deficiency is that the process does not include means of responding to a situation, where spectrum scarcity occurs.

As far as we know, Europe is the only multi-country area of the world, where such a harmonised introduction scheme has been implemented. In other regions of the world the decisions will be made on a country-by-country basis. Some global guidelines do, however, exist. In accordance with Opinion 5 of the first World Telecommunications Policy Forum, a group was established to facilitate the implementation of GMPCS in developing countries.

Looking more closely at the situation of the 2GHz band, CEPT has been requested to implement co-ordinated frequency harmonisation and asked by the ERC to identify candidate systems for the band. In Europe, there are currently only TDMA

candidates for that band. Consequently, a portion of this band has been reserved for TDMA systems, while leaving the other band of the band 'to be decided', in case other candidate systems would appear at a later stage. It has to be clearly understood that, according to that process, the band has not been assigned to any system in particular, but rather to a particular usage of the band. In contrast, the US FCC released on 25 March a Notice of Proposed Rulemaking, [7] outlining the following four possible approaches to regulate the usage of the 2GHz MSS spectrum:

- a "flexible band" arrangement of 3 core and 2 expansion spectrum bands, which groups applicants in segments according to technology used (TDMA, CDMA, combined TDMA/CDMA), allocates 2.5 MHz uplink/2.5 MHz downlink for TDMA and 12.5 MHz uplink/12.5 MHz downlink for CDMA and provides for expansion for new system requirements between segments;
- a "negotiated entry arrangement," by which proposed systems are licensed across the 2 GHz MSS band and co-ordination and dispute resolution are to be dealt with by them or with FCC assistance;
- a traditional band arrangement, allocating spectrum equally among applicants (3.75 MHz uplink/3.75 MHz downlink with 0.625 MHz guard bands between TDMA and CDMA operations);
- an auction of spectrum.

This NPRM is open for comments until the 26 June and takes into account the 9 systems proposed to the FCC for operation in this band, following the 2 GHz MSS processing round that the FCC initiated in 1997. These systems are listed in table 6.

Applicant	Service Link Spectrum Request	System	Technology
Boeing	uplink: 8.25 MHz at 1990-1998.25 MHz downlink: 8.85 MHz in 2170-2185 MHz	16 NGSO	CDMA
Celsat	uplink: 25 MHz in 1990-2025 MHz downlink: 25 MHz in 2165-2200 MHz	1 GSO	TDMA/ CDMA
Constellation II	uplink: 45 MHz at 1980-2025 MHz downlink: 35 MHz at 2165-2200 MHz	46 NGSO	CDMA
Globalstar	uplink: 35 MHz at 1990-2025 MHz downlink: 35 MHz at 2165-2200 MHz	4 GSO; 64 NGSO	TDMA/ CDMA
ICO	uplink: 30 MHz at 1985-2015 MHz downlink: 30 MHz at 2170-2200 MHz	10-12 NGSO	TDMA
Inmarsat Horizons	uplink: 45 MHz at 1980-2025 MHz downlink: 40 MHz at 2160-2200 MHz	4 GSO	TDMA
Iridium Macrocell	uplink: 35 MHz at 1990-2025 MHz downlink: 35 MHz at 2165-2200 MHz	96 NGSO	TDMA/ CDMA
MCHI Ellipso 2G	uplink: 35 MHz at 1990-2025 MHz downlink: 35 MHz at 2165-2200 MHz	26 NGSO	CDMA
TMI Cansat-M3	uplink: 35 MHz at 1990-2025 MHz downlink: 35 MHz at 2165-2200 MHz	1 GSO	TDMA/ CDMA

Table 6: Applicants to an FCC license at 2GHz

Following the results of the NPRM consultations, it is likely that in Europe, the CEPT will have to reconsider the overall situation of this band, with the objective of optimising the global usage of the band.

One Stop Shopping

The Licensing Directive (Directive 97/13/EC) is another key regulatory text of the European Union that foresees the establishment of a one-stop shopping mechanism. So far only a few Member States (France, Germany, Netherlands and UK) and Switzerland have entered into an agreement - predating the Directive - to apply a one-stop shopping procedure for the granting authorisations of certain satellite services such as VSAT and SNG. Already since summer 1997, the CEPT has been mandated by the Commission to develop satellite-specific rules for a one-stop shopping procedure for satellite authorisations across Europe, pursuant to Article 13 of the Licensing Directive.

Although such a mechanism would seem to be particularly appropriate for satellite systems which, by their very nature, provide regional if not global coverage, there has been quite some debate among European administrations about the added value of any such mechanism. While regulators generally are wary of foregoing ultimate control over the authorisation mechanism for their respective territories, the satellite industry has, at times, commented that a mere 'letter box' mechanism might not have the desired effect of simplification but could risk establishing yet another layer of procedure.

Work in CEPT is in progress and should cover all aspects related to satellite network operators, service providers and subscribers of small fixed earth stations or other mobile terminals. It is expected that the future mechanism would comprise a common database and provide common application forms for applicants. It will potentially cover fixed, mobile and broadcasting satellites, and could be in place before the year 2000.

Free Circulation of Mobile Terminals

Another issue to be considered is the carriage and use of terminals, particularly the carriage and use by foreign visitors. One of the key elements contributing to the success of a global mobile system will be the possibility of users to cross borders with their terminals without any regulatory formalities. Therefore measures must be taken to ensure the free circulation of GMPCS, and in a couple of years the IMT-2000 terminals. If the issue of terminal circulation is simply left to be sorted out by its own, widely divergent national practices may follow; some countries adopting liberal regimes and some other highly regulated systems.

To achieve the global circulation of GMPCS terminals, the first World Telecommunications Policy Forum in Geneva 1996 established the GMPCS MoU. Within the MoU, a set of arrangements have been developed on type approval, marking and licensing of GMPCS terminals, the access to GMPCS traffic data by administrations and the customs issues of GMPCS terminals.

Regarding type approval, it has to be clearly stated that the GMPCS arrangements do not provide for a global type approval. It is quite in accordance with the arrangements that each country requires its own type approval. Nevertheless, there is a general desire that those countries not having an existing national type approval regime would not start implementing one but would recognise the type approvals registered in the GMPCS registry maintained by ITU. Instead of being a global type approval, the type approval arrangement is part of the procedure, which through different notifications and registration in a common database leads to the possibility of circulating and using GMPCS terminals without the requirement of individual licences.

The Arrangement on the access to GMPCS traffic data was generated by the concern of several administrations, that large scale use of GMPCS in their respective countries would by-pass the national telecommunications network and cause a loss of revenues. Although it has been pointed out that increased use of GMPCS will cause an increase in the overall volume of calls over the country's own network, several administrations required this article as a safeguard clause. The GMPCS system operators are required to provide to any requesting administration data on the GMPCS traffic originating or routed to the administrations national territory. Operators are also required to assist in identifying unauthorised traffic flows.

The customs fees and duties may cause a particular threat against the global circulation of GMPCS terminals. The existing mobile satellite operators complain that there are countries, where the customs fee may be on the same level of money value as the price of the terminal. If this would be the case for GMPCS terminals, the global circulation would be seriously endangered. The GMPCS MoU was established in the ITU environment, which is not competent to prepare customs arrangements. Therefore the participating administrations merely agree to recommend to their national customs authorities that the duties on GMPCS terminals should be reduced or removed.

Within a short time the need for global circulation will also apply to IMT-2000 terminals. The satellite component of IMT-2000 is covered by the GMPCS arrangements, but how to deal with the terrestrial component? The solutions adopted by GMPCS MoU may not be suitable for terrestrial IMT-2000, because there are several fundamental differences between terrestrial and satellite communications. Nevertheless, the circulation of IMT-2000 terminals need to be arranged in a globally acceptable way.

In conclusion, it can be stated that the European Union has put in place a consistent and ambitious regulatory package, that takes into account the specific requirements of the satellite industry. Considering the complexity of a multi national environment, it can be stated that this package optimises the conditions of access to frequencies, the circulation of terminals, and the obtention of license. The European Commission is however trying to improve this regulatory package, where applicable. It has in particular launched a wide consultation on spectrum matter through a 'Spectrum Green Paper'. The results of the consultation will be published around mid July. In addition, it is planned to issue, by the end of the year, a document dealing with a review of the regulatory package, with the intention of identifying if modifications to the package are required. In that context, the European Commission is currently working with the space sector to identify the specific requirements of the satellite industry.

SYSTEMS AND STANDARDS

It appears from above figures and considerations that the market, in the medium to long term future, may be able to support a couple of S-UMTS systems. The situation regarding a potential development of such satellite based systems is however considerably different from that of terrestrial systems, especially in Europe. Europe has been able to develop a common vision for the evolution of the terrestrial mobile communication sector, and in particular to agree on an access standard for third generation UMTS/IMT 2000 systems, the 'UTRA' standard, which has generated wide support even from outside Europe. The same can not be said for S-UMTS. This implies that the S-PCS scenario, with system proponents following individual strategies, is likely to replicate itself, with the potential following drawbacks:

- different access technologies, with non inter-operable systems;
- consumer locked to a particular system and technology;
- further market fragmentation, leading to more expensive user terminals;
- absence of economies of scale and integration;
- less efficient integration of satellite/terrestrial access

In Europe, the definition work for a possible S-UMTS system has been extensively carried out by the European Space Agency (ESA),and by the SINUS project [8], a space industry led project of the ACTS programme, sponsored under the Fourth Framework programme (1994-1998) for R&D of the European Union. Following the ITU call for candidate RTT's to be submitted by June 98, ESA decided to submit the results of its work as a candidate RTT for S-IMT/2000, however with limited industrial support. Industry did on the other hand decide not to present the results of the SINUS project as a candidate

RTT. As a result, ITU received 6 proposals for a satellite access scheme, from the following sources:

- ESA SW-CDMA, Wideband CDMA
- ESA ST-CDMA, TD-CDMA
- TTA (Korea) Wideband CDMA, LEO;
- ICO TDMA, MEO system;
- Inmarsat TDMA, GEO system;
- Iridium, CDMA/TDMA LEO system;

As all these proposals have successfully passed the evaluation scheme for candidate RTT's elaborated by the ITU, the six proposals are now fully reflected in the ITU Recommendation IMT.RKEY [9] that specifies the characteristics of the access schemes.

Parameter	FMA2	ESA (SW-CDMA)	SINUS
Access	W-CDMA	W-CDMA	W-CDMA
Duplex Mode	FDD	FDD	FDD
Modulation	QPSK	*Forward*: QPSK; *Return:* BPSK with control channel on Q carrier	*Forward:* QPSK, *Return:* O-QPSK
Chip Rate (Mchip/s)	4.096	2.048 or 4.096	4.3008 or 4096
Bandwidth (Mhz)	4.4 - 5	2.5 or 5.0	4.8
Pulse Shape	RRC, 0.22	RRC, 0.22	RRC, 012
Frame length	10 ms	10 or 20 ms	10 ms
Interference Reduction	Multi-user detection supported	Linear MMSE supported, forward and return links	-
Spreading factor	4 to 256. Short codes	16 to 256, according to data rate	*Forward:* 384-PN codelength=8192 *Return:* 25 to 384-PN length=2E41
Multi rate concept	variable spreading and multi codes	Orthogonal, variable rates	*Forward:* multi code, WH 128 function; *Return:* variable spreading similar to FMA 2
Detection	Coherent	Coherent	*Forward:* coherent; *Return:* multi symbol differential
Handover	Mobile controled soft H/O Interfrequency supported	Mobile assisted, network initiated soft H/O; satellite diversity supported	Mobile asisted; Satellite diversity supported
Power control	16 PC groups per time slot	Reduced PC rate compared to FMA2	Close loop, one command per frame, reduced rate compared to FMA2

Table 7: Physical layer characteristics

For the systems studied in Europe, it can be stated that the approaches followed by the ESA and by the SINUS project are slightly different, but have lots of commonality. They both aim at maximising harmonisation with the terrestrial standards, thus allowing for the production of inexpensive multi-mode satellite/terrestrial terminals. SINUS has primarily concentrated on a W-CDMA solution, looking for maximum harmonisation with the FMA2 scheme, the terrestrial W-CDMA mode developed by the FRAMES ACTS projects [10], which has led to the UTRA standard adoption by ETSI. The proposed scheme aims at covering a whole range of orbits, from LEO (780 km) up to GEO, with use of variable parameters (e.g. for power control or interleaving depth). The ESA approach is similar, with a W-CDMA version also largely based on FMA2 [11], but optimised for orbits ranging from LEO to MEO. In addition, ESA has proposed to use a derivation of the FMA1 mode (TD-CDMA option for TDD use, the other mode of the European UTRA terrestrial standard), optimised to cover orbits ranging from HEO to GEO.

Beyond the need to harmonise the satellite access with terrestrial solutions, work under both ESA and SINUS project have concentrated on CDMA for following reasons:

- satellite diversity is an essential feature considering the required service availability rates;
- easier to access to different types of constellations;
- soft handover;
- simpler radio resource management, TDMA requires a more complex frequency planning between the satellite spots.
- CDMA offers high capacity solutions.

Table 7 summarise the main physical layer characteristics of the terrestrial FMA2 mode and of both SINUS and ESA W-CDMA satellite options. Tables 8 and 9 outline the service characteristics considered by the SINUS project and by ESA, respectively. Note that both proposals are designed to support a maximum bit rate of 144 kb/s.

As a follow up to this work, the possibility to further harmonise the TTA W-CDMA mode, the ESA SW-CDMA mode and the SINUS W-CDMA mode is investigated by the various parties, with the objective of refining the ITU RKEY satellite specifications.

Service Type	Bearer Class	Typical Bit Rate (kb/s)	B.E.R	Max end to end delay (ms)
Speech	RT	4 to 6	10E-3	400
Audio	RT	8 to 64	10E-5 to 10E-6	400
Data	NRT	1 to 144	10E-5 to 10E-6	500
Text	NRT	20	10E-6	500
Image	NRT	2.4 to 144	10E-6	400
Video	RT/NRT	8 to 144	10E-6	400

Table 8: SINUS Services characteristics. (RT=Real Time; NRT=Non Real Time)

Bearer Rate Kb/s	Max end to end delay (ms)	Services Qualities
1.2	400	10E-6
2.4	400	10E-3; 10E-5; 10E-6
4.8	400	10E-3; 10E-5; 10E-6
9.6	400	10E-3; 10E-5; 10E-6
16	400	10E-3; 10E-5; 10E-6
32	400	10E-3; 10E-5; 10E-6
64	400	10E-5; 10E-6

Table 8: ESA Services characteristics.

Even though the concept of a common standard (or of a family of standards) is appealing for the reasons mentioned above, it remains to be understood how far such concept can be supported by the satellite industry. In fact, there are currently four different system approaches competing with each other:

- **No a-priori defined standard.** This is in particular an approach prevailing among US players. A report released by TIA [12] with a view to studying and comparing the various satellite candidate RTT's concludes in that direction, outlining that optimisation of various service and system scenarios will inevitably call for various RTT's rather than for a single standardised version.
- **A-priory defined family of standards,** with maximum commonality with the terrestrial standards. This is in particular the approach followed by ESA. With some flexibility built in the specifications, this approach would imply that mobile satellite operators are no longer competing on technologies, but rather on services, as in the case with terrestrial systems and operators.
- **An intermediate approach,** developed in project SINUS, is to standardise only the upper layers, following the GRAN (Generic Radio Access Network) approach, whilst leaving the designers the flexibility of selecting the physical and MAC layers of their choice.

Some work has been dedicated to that approach in ETSI. It corresponds to clearly identifying the radio dependent features and radio independent features, with the objective of standardising the relevant interface only (Iu interface).

In Europe, the RAINBOW project has developed a network platform allowing to test this concept with a number of different radio accesses, including the satellite access developed by SINUS. In essence, the access becomes 'plug replaceable' visa a vis the core network, as conceptually outlined on figure 7.

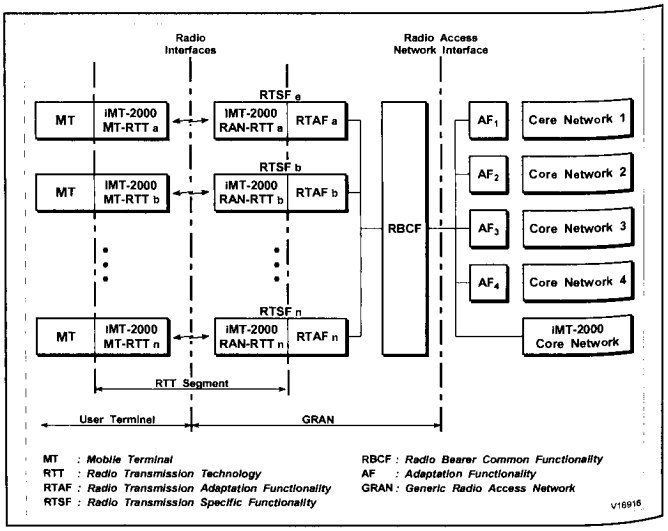


Figure 7: Iu INTERFACE CONCEPT

- **Convergence in the medium term.** Another approach, which has the potential to reconcile all previous approaches, is to let different standards develop, whilst looking for medium term convergence and interoperability through the development of software reconfigurable terminals (for both terrestrial and satellite accesses). The software radio approach is currently the subject of a number of exploratory projects in Europe, also currently sponsored by the 4th Framework Programme of the Union. In the field of satellite systems, the projects SORT [13] is currently working towards the implementation of a demonstrator with two critical functionality (chanelisation and sample rate adaption) and three types of air interfaces (terrestrial UTRA, GSM, Satellite W-CDMA of SINUS). The architecture of this demonstrator is outlined on figure 8.

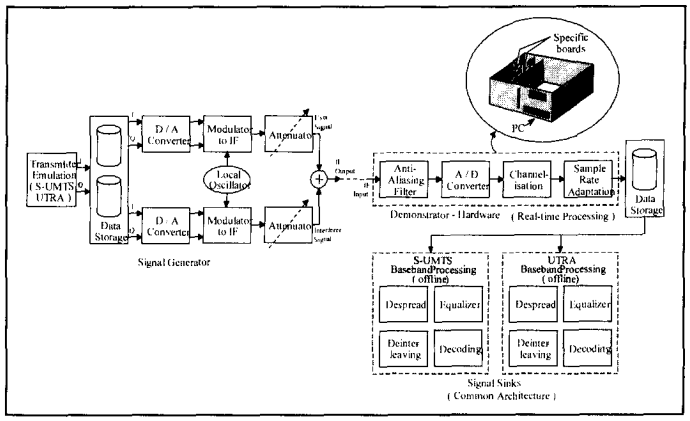


Figure 8: SORT Demonstrator Architecture.

TECHNOLOGY PERSPECTIVE

A number of key technological challenges remain open for S-UMTS, e.g. satellite technology issues such as availability of reliable On Board Processing, inter satellite links, large antenna for GEO systems, high gain/reconfigurable antenna, integrated on board technologies. In Europe, these issues are mainly addressed by Space Agencies such as the ESA.

Other technological issues are related to terrestrial technologies and systems. Among those, one can mention:

- integration with the terrestrial network for seamless service provision. ACTS projects SINUS and RAINBOW are addressing in particular inter-segment handover and demonstration of the 'radio independent' concept (figure 9).

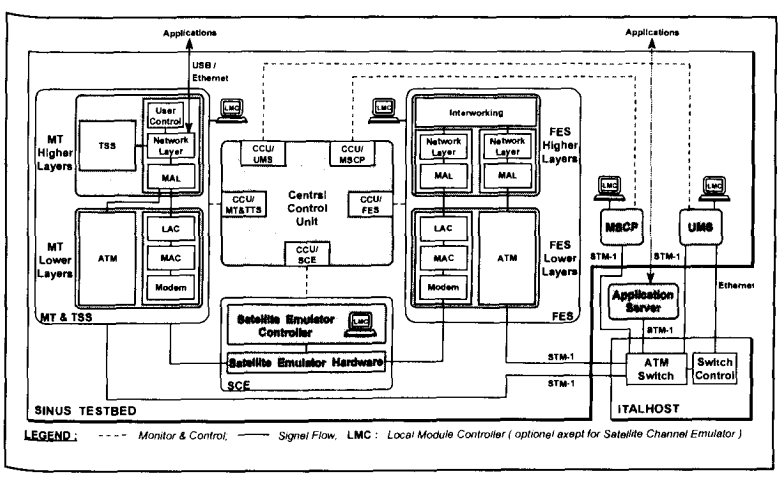


Figure 9 SINUS/RAINBOW integration Trial Architecture.

- Provision of terminal/network aware services. This issue will be more and more important, considering the heterogeneity of network platforms and terminals that will co-exist in the future. Development of API's capable of delivering a user with the service the underlying network and the terminal can support is thus a key issue, especially when roaming from a bandwidth rich environment such as a terrestrial network towards a bandwidth limited satellite network. In Europe, this issue is investigated by project SUMO [14] with a trial allowing to support service provision through two different satellite access schemes.
- Capability of the satellite systems to support services developed for terrestrial UMTS networks, such as MPEG-4 based services. In Europe, compatibility of MPEG 4 with satellite systems is extensively trialed through project TOMAS, using Inmarsat links and advanced modems.

The above gives a short overview of some technological issues currently investigated in Europe in the field of S-UMTS. Taking a wider perspective, and considering that eventually, S-UMTS will be fully integrated with terrestrial network and will provide similar services, the technological challenges facing the mobile satellite sector can be viewed as similar to those facing the mobile sector at large. Those are summarised below:

➢ **Network independence in a more and more heterogeneous environment**: the main challenge will be the possibility to seamlessly roam across a range of mobile or wireless access networks, providing different service support capabilities. GSM, DECT, D-AMPS, GPRS, UMTS, satellites etc, are typical examples, that are expected to be complemented by the emergence of future 4th generation mobile systems. If IP may be considered as a unifying 'glue' to reconcile such multiplicity of mobile environments, a number of key mobile related issues are still open: the ability of IP to support services seamlessly roaming across different environments; the ability to support different classes of services as a function of the limitations of the wireless access, the real time control and reconfiguration of the services depending on the access environment, the real time control needed to acquire and release lower layer network resources in the wireless access portion as a function of the roaming scenario. At the lower layer level, network independence is also tightly related to the advent of software re-configurable terminals, which will provide any terminal with the capability to access any mobile network through software download of the access method characteristics. Network management issues are critically related to these various aspects;

➢ **Transport optimisation**: Because of spectrum limitations, wireless/mobile access bandwidth will remain a scarce resource, and the irregularity and quantity of transmission, errors and link outages imposes major challenges on any system design. Future transport architectures (in all frequency bands) and protocols will have to take full advantage of available bandwidth, provide an optimised data delivery over the air link and offer error recovery and retransmission mechanisms that can quickly react to

link impairments. This is even more critical for satellite systems;

- **Terminal independence and terminal aware network:** a vast variety of mobile terminals can be expected in the future, with differing limitations and capabilities. This in turn necessitates the adaptation of services to terminal capabilities. One of the key challenges for future advanced mobile systems will lie in the capability negotiation and storage facilities provided by the networks. Terminal information such as screen resolution, colours, processing power... will need to be accessible to service providers for them to align their content accordingly.

- **Terminal features**: Mobile devices are still fairly restricted in terms of man-machine interface, on board memory, limited colour displays, processing capabilities. If operating systems for mobile devices are today providing a TCP/IP protocol stack, the processor still encounters severe difficulties to decode and encode multimedia content without specific hardware support. Popular mobile applications will have hundreds of features and research in man machine interface is still required to understand which features will be needed and how to build convenient and attractive mobile applications. The main machine independent programming of mobile applications is also a critical issue.

- **Generic mobile aware services support**: a key condition for the provision of services capable for roaming across heterogeneous mobile access platforms is the provision of a standardised set of mobile aware services, de-coupling the application from the underlying wireless access platform, and capable to support real time reconfiguration of the service profile as the terminal is roaming through different environments. This set of generic mobile aware services will have to include access to user location information, and also support access to timing information for timing related application (instantaneous access to value of shares, weather information or else). In addition, this set of services has to adapt the information content to the capabilities of the underlying network. This work has started under the WAP forum, but the standard will only support limited capabilities. It does not support monitoring and roaming features, nor automatic service adaptation. It is today not clear if WAP will be able to adapt to the evolution of powerful PDA's which support voice over IP. Related developments, which still need extensive validation in a wireless/mobile context is the new generation HTTP, HTTP-NG.

- **Global Service portability**: is an issue related to the previous one. It implies in particular the definition and standardisation of a globally accepted NNI interface to support transfer of service features from one operator environment to another.

- **Multiple Standards**: is an issue that may be expected to dominate the future evolution of mobile and wireless solutions at global level. This may raise the issue of the availability of multi-standards, with its associated service provisioning (e.g. roaming across distinct networks with different service capabilities and in particular with different authentication and user certification algorithms) and network management and billing problems.

These issues are related to network integration and seamless service provisioning and service portability. In addition, the shorter term issue of validating the W-CDMA solutions proposed for S-UMTS access remain open. A large number of W-CDMA demonstrators do today exist for terrestrial UMTS and extensive technology trials are currently being run by a large number of operators in the world. In comparison, the satellite W-CDMA remains largely untested.

THE FIFTH FRAMEWORK PROGRAMME

The 5th Framework Programme (5th FP) recently launched is a programme of R&D sponsored by the European Union addressing a multiplicity of research areas and covering the period 1999-2002. It is marked by a particular effort of selectivity and concentration on a limited number of areas and objectives. In terms of economic development and scientific and technological potential, the areas of work selected correspond to domains in which European firms can and must become more competitive, areas which are expanding and have good growth potential, and areas in which significant technological progress is anticipated. Employment, quality of life, health, positive social impact and a clean environment are obvious priorities and a serious concern of the EU citizens. A key aspect is that of subsidiarity namely that the areas of work are those where a "critical mass" in human and financial terms needs to be established and a mixture of complementary expertise found in the various countries is needed, and those that concern European problems and the development of the European area or involve aspects of standardisation. The 5th FP is designed seven programmes, four "thematic" and three "horizontal", all necessarily complementary and interrelated, with an overall budget envelope of 14.960 Bn Euro .

The "horizontal" programmes cover 1) co-operation in the field of R&D with third countries and international organisations with the objective of *confirming the international role of European research*, 2) the dissemination and optimisation of results through *innovation and participation of SMEs*, and 3) *improving human potential* through stimulation of training and mobility of researchers.

The "thematic" programmes are R&D programmes focusing on the following priorities: 1) *improving the quality of life and the management of resources of the living world*, 2) *creating a user-friendly information society*, 3) *promoting competitive and sustainable growth*, and 4) *preserving the ecosystem*. The work carried out in previous thematic programmes of the now ending 4th FP such as ACTS, Telematics and Esprit will be continued mainly in the scope of the Information Society programme.

The Information Society Technologies (IST) Programme

The most important aspect of the Information Society is its potential for renewed growth and the creation of new forms of work: several million jobs will result in Europe alone from the development of the Information Society, provided the challenges of technical excellence and preservation of cultural diversity can be met.

Achieving the full potential of the Information Society requires in turn continued R&D, side by side with a technology adoption effort. The aim of the Information Society Technologies (IST) programme is to have the advanced goods, services and methodologies made possible by the Information Society contribute to creating new jobs and strengthening the competitiveness of European companies, stimulating the development of new markets and services and strengthening the role of the public in society, improving the position of Europe as a place for investment, research and innovation, and strengthening the scientific and technological base of the EU with the goal of reinforcing global competitiveness. The need for accessibility and interoperability at all levels, from technologies and tools to systems and applications, is paramount.

The IST Programme has received a total budget envelope of 3.6 Bn Euro and is further broken down into 4 'Key Actions': *1) Systems and Services for the Citizen, 2) Electronic trade and new methods of work, 3) Multimedia content creation and access tools,* and *4) Essential technologies and infrastructures [15].* Mobile and Wireless Communications, including in particular satellite systems and services , are specifically being considered as part of Key Action 4.

Other topics within Key Action 4 are computing, communications and network technologies comprising architectures, protocols and methodologies, as well as their introduction and use; software and systems technologies and engineering; user interfaces and peripherals; and micro-electronics including circuit design and application development.

Mobile and Wireless Communications in the IST Programme

The impact of the technologies associated with the Information Society extends far beyond the industries directly involved: they have implications in terms of fundamental industrial (mass market perspective) and social (new activities, new relations) change.

The unprecedented growth of world-wide mobile/wireless markets, coupled with advances in communications technology and the accelerated development of services taking place in fixed networks, has made the introduction of a flexible and cost-effective Third Generation Mobile Communication System a priority.

The next phase, to be pursued in the scope of the IST programme, will add a different dimension. Concurrently with the much needed and anticipated integration of fixed and mobile, as well as terrestrial and satellite systems, the continued development of Mobile/Wireless Broadband Communication systems towards Mobile Multimedia, the systematic search for performance improvements through the (optimum) combination of technologies (from adaptive/smart antennas to modulation/coding schemes to compression algorithms to "data adaptation"), a dramatic paradigm shift is envisioned with the adoption of Soft Telecommunication concepts.

At the same time, research in enabling technologies will insure that the necessary developments are in place (from smart batteries to high temperature superconductivity to flat panel displays) to allow for a true mass market of telecommunication products. With those, a myriad of personalised telecommunication services will emerge, requiring the development of new service engineering tools and the extension to the telecommunications arena of Just-in-Time service definition and provisioning. To stress the point, a user-centric perspective will refocus the work towards the seamless offer of personalised, Just-What-the-Customer-Wants services, and provide the user with simple, user-friendly interfaces to the service providers, from profile definition to service selection to billing.

A particular aspect of the development of the Mobile and Wireless Communications sector is that relating to its satellite communications dimension. Today the European Union is sponsoring R&D projects in the field of S-UMTS. These projects are complementing the satellite technology developments carried out by Space Agencies and bring together the satellite and terrestrial communities with a view to developing solutions for inter-operable systems and services, with a global satellite/terrestrial perspective. It is intended to maintain that approach under the 5th FP, in collaboration with Space Agencies.
The IST Programme, will thus offer further opportunities to sponsor satellite communication projects, addressing either specific technology issues or broader system issues. Continuation of work on software radio for both satellite

and terrestrial environments, interoperability of terrestrial and satellite networks, global network management, advanced re-configurable ground segment for multimedia constellations, application trials for various user communities are some of the themes lending themselves to support.

It is expected that the work relating to Mobile and Wireless Communications will hence focus on the development and evolution of new generations of affordable terrestrial and satellite broadband wireless networks, systems and technologies, for both private and public environments, supporting advanced services and maximising spectral efficiency and optimising network performance. Particular attention will be paid to:

- Integrated seamless network that ensures global personal connectivity (through a multiplicity of radio systems deployed in a multi-layer, multi-dimensional cell architecture), and enables access to adaptive multimedia mobile services with performance, capabilities and quality comparable to those of fixed networks.
- Service and terminal mobility across wireless and wired networks.
- Advanced broadband communication technologies, systems, services and applications supporting interactive, symmetric and asymmetric, multicasting and narrowcasting, mobile and personal multimedia services, with regional or global coverage and integrated, where appropriate, with location services.
- Software re-configurable networks, systems and terminals (software distribution, software reconfiguration, adaptation to different air interfaces and network standards)
- Cheaper, lighter and more power efficient wireless and portable terminals.
- Take up measures.

As in the 4th Framework Programme, a key aspect will be technology assessment with validation and demonstration of broadband interactive mobile multimedia technologies and services.

CONCLUSIONS

This paper has reviewed the developments of S-UMTS, with a broad perspective. It has outlined that there are still many open issues before a broadband mobile system can be successfully deployed, and main conclusions are as follows:

- Market prospects indicate that there is a possible niche for mobile satellite systems to offer broadband services, in complement to terrestrial UMTS. The comparatively small number of users however requires to further validate these figures, and to devote particular attention to how best position any S-UMTS initiative in terms of time to market and technology. A possible optimum approach may be to 'piggyback' on terrestrial systems, by defining close partnerships allowing to minimise the overall infrastructure costs.
- Frequency spectrum remains a scarce resource. This is further complicated with the likely deployment of S-PCS systems in the S-UMTS frequency bands. Some form of global spectrum harmonisation is also likely to be required.
- Optimum regulation for the deployment of large regional or global systems remains a key issue. The European Union has taken the necessary measures to allow for co-ordinated introduction of S-PCS systems. Potential improvements, such as One Stop Shopping, are currently investigated with industry and Member States.
- Lack of standards for satellite access remain a critical issue, that will likely impose high terminal costs. Software radio may be the answer, but is a long term solution.
- Technology development is still required. In particular, technology allowing for optimum integration with terrestrial networks, seamless service provision and global management is needed.
- In the European Union, the 5th Framework Programme of R&D will offer opportunities to industry and research organisations to progress on these subjects, in collaboration with Space Agencies.

REFERENCES

[1] UMTS Forum Report 1, 'A Regulatory Framework for UMTS'; www.umts-forum.org.
[2] UMTS Forum Report 8, 'The Future Mobile Market'; www.umts-forum.org.
[3] ICO-TENS Report, 'The Potential for pan-European Multimedia Mobility'; EC TEN Contract 45557, Dec. 98.
[4]UMTS Forum Report 6, 'Assessing Global Requirements for the next Century'; www.umts-forum.org.
[5] Peter Scheele, CEPT TG1, 'Spectrum Issues for IMT 2000/UMTS', ITU Jersey Workshop, Nov 98.
[6] ITU TG 8/1, Draft CPM Report on WRC Agenda item 1.6.1, Doc 8-1/TEMP/164-E, March 98.
[7] FCC NPRM IB Docket No. 99-81RM-9328, 'The Establishment of Policies and Service Rules for the Mobile Satellite Service in the 2 GHz Band' March 99.
[8] ACTS Project AC212 SINUS, Deliverable 23, 'Satellite UMTS Radio Access Specification', June 97.
[9] ITU TG 8/1 Draft IMT.RKEY Recommendation, Doc 8-1/TEMP/168-E, March 99.
[10] ACTS Project AC 090 FRAMES.
[11] ESA Proposals for an RTT on S-UMTS, available at http//www.itu.int/imt/2-radio-dev/proposals/esa
[12] Report of TIA TR-34 Ad Hoc IMT-2000 Satellite RTT Evaluation Committee; June 98.
[13] ACTS Project AC 315 SORT, Deliverable 4, 'Implications on Future Software Radio', Oct 98.
[14] ACTS Project SUMO, Brussels Workshop Presentation, March 99.
[15] Information package of IST Programme, available at www.cordis.lu/ist/.

Universal Satellite Modulator

K.M.S. Murthy, S. Daigle and V. Allen, and M. Wlodyka
COM DEV Space Group, 155 Sheldon Drive,
Cambridge, Ontario, N1R 7H6, Canada
Ph. 519 622 2300 x 2681, Fax. 519 622 5971, Email: murthy@ieee.org

ABSTRACT

Developing [1] a very high speed, radiation hardened, satellite-on-board modulator which can be used for a wide range of applications is a challenge. The requirements of the modulator are derived from considering applications which include, mobile systems, multimedia systems, radio inter-satellite links (RISLs), remote sensing, digital video systems, military and other special mission systems. The primary focus of this paper is to present the design philosophy and design trades of a universal modulator which can support a wide range of applications. Major design challenges of such a universal modulator are in the areas of programmable FEC design, realization of multiple modulation constellations, and digital filter design for very high speed operation, provision of burst and beam-hopping controls, and realization of low power and phase imbalance between quadrature components.. This paper addresses some of these issues briefly.

INTRODUCTION

As advanced wide area mobile and multimedia applications and services emerge, the need for and value of digital regenerative satellite payload architectures increase. Services requiring packet level routing at the satellite node require demodulation and remodulation amongst other functions. The requirements of such a regenerative system will impose a considerable level of complexity on the on-board transmission system. The key elements of the transmission system which may differ for different applications include: data rate, encoding and modulation formats, standards compliance, filtering characteristics, control functions, mode of operation, power control etc. The modulator being the critical sub-system of the transmitter, should be smart and universal so as to comply with current standards, meet performance and possibly adapt to future changes. It is in this context that the development of a universal satellite modulator(USM) has been undertaken and the paper describes the salient features of the USM. The development of universal modulator for satellite applications is important to increase flexibility to address multiple applications and meet changing requirements.

[1] This paper was prepared under severe time constraints and the authors regret for errors and discontinuities if any.

APPLICATIONS

The modulator incorporates a variety of modulation and coding, and filtering options so as to optimize its performance for a wide range of applications. The applications considered include:

- mobile satellite systems
- multimedia satellite systems
- remote sensing satellite downlinks
- digital video broadcast and interactive systems
- military satcom systems
- special mission and systems

The type of satellite applications and system environment that USM will support are given in Table 1.

Table 1 Modulator Applications Environment

Satellite Systems	Satellite Orbits	Satellite Links	Data Rate (Mbps)	Frequency (GHz)
Broadband	GEO, LEO	-User - Gateway - ISL	16-OC3 OC1-OC3	Ku, Ka, V
DVB Sys.	GEO	- User - Gateway	30 - OC2	C, Ku, Ka
Remote Sensing	GEO	- User - Gateway	40 - 155	X, Ka
Mobile	GEO, LEO, MEO	-Gateway - Radio ISL	~ 10 OC1-OC3	Ku, Ka, V
Mil-Satcom	all	- User - Gateway	16 - OC3	
Special Missions	-	- User		

DESIGN PHILOSOPHY

The design philosophy is based on evolving a very high speed universal modulator to support a variety of applications. An all digital implementation approach which include a flexible digital filter (FIR) for wave shaping, near-baseband modulation, a variety of FEC encoders, flexible scrambler, interleaver and framer is adopted.

At the heart of the modulator is an ASIC which will have (1) self initializing and self testing capabilities, (2) factory programmable parameters, (3) on-board controllable parameters, (4) continuous, burst and beam-hop operational modes. The choice of modulation and encoding techniques are made by considering a number of factors such as, (1) power-bandwidth efficiency, (2) operating E_b/N_o domain, (3) channel impairments (propagation, interference, synchronization, etc.), (4) Standards compliance, (5) performance, and (6) implementation complexity.

OPERATING E_b/N_o DOMAIN

Operating E_b/N_o domain: The design of the high speed digital modulator focused on the environment in which it is to operate - i.e. LEO, GEO, ISL and to classify system bounds based on different application and service scenarios. Basic satellite and earth station parameters were obtained from the FCC filings on proposed satellite constellations as well as other sources. The first task was to establish the bounds of satellite parameters (EIRP, G/T, antenna, etc.). Given the satellite parameter bounds, then perform link design analysis to calculate the available E_b/N_o and required EIRP for a variety of ground terminals including, personal use, home business use, business use, and gateway terminals. Considering earth station parameters such as antenna diameter, LNA noise temperature and cost, the bound (range) for E_b/N_o was prepared. The list of FEC coding and modulation schemes considered would have their required E_b/N_o fall within this bound (see Figs 1 & 2).

Channel Impairments and interference model
Channel impairments applicable to both LEO and ISL systems such as Doppler were considered and accounted for in deriving the available E_b/N_o operating bounds for these systems. Interference models were developed for the both inter-satellite links (ISL) and LEO ground links taking into account adjacent satellite interference. The results were taken into account in establishing the E_b/N_o bounds.

Usually the two most important parameters in the design of a digital modulator are; (1) the power efficiency (measured in terms of the required E_b/N_o to achieve a given bit error rate (BER), and (2) the bandwidth efficiency (measured in terms of number of bits/sec per unit radio frequency bandwidth, i.e. bps/Hz). Our exercise of analyzing the EIRP and G/Ts of different satellite systems in conjunction with various receiver terminal G/Ts gave us ***"the boundaries"*** for the two critical parameters of the modem subsystem, viz.,

- required E_b/N_o for a specified bit error rate (BER)
- required bits/s per Hz efficiency

The typical E_b/N_o boundaries obtained are illustrated in **Figures 1 and 2** for GEO and LEO systems respectively The operating regions are classified and bounded by system which covers requirements of most of the proposed GEO. LEO systems.

POWER-BANDWIDTH TRADE

For bandwidth limited systems, low redundancy FEC codes can be used to increase the throughput efficiency of the system at a marginal E_b/N_o penalty to achieve a target BER performance. On the other hand, power limited systems can take advantage of lower rate coding to achieve required bit error rates at the expense of increased bandwidth (or alternatively, a lower acceptable data rate within fixed operating bandwidth constraints). In situations where fading due to rain attenuation occurs, an adaptable coding scheme would be effective, enabling near optimum operation and throughput to be achieved for given channel conditions. Many combinations of coding and modulation schemes are presently supported by the modulator signal processing block. Table 2 presents the power and bandwidth efficiency values for a variety of modulation and coding combinations.

Table 2 Power/Bandwidth Efficiency of Schemes

ID	Modulation/Coding	Min. Eb/No (dB)	Bandwidth Efficiency (40% roll-off) (bps/Hz)
MC1	QPSK, conv. 1/2, RS(204,188)	3.4	0.66
MC2	QPSK, conv. 2/3, RS(204,188)	3.9	0.88
MC3	QPSK, conv. 3/4, RS(204,188)	4.4	0.99
MC4	QPSK, conv. 5/6, RS(204,188)	4.9	1.10
MC5	QPSK, conv. 7/8, RS(204,188)	5.3	1.15
MC6	T8PSK, RS(204,188)	5.9	1.32
MC7	QPSK, conv. 1/2	6.7	0.71
MC8	T16QAM, RS(204,188)	6.9	1.97
MC9	QPSK, conv. 2/3	7.2	0.95
MC10	QPSK, conv. 3/4	8.0	1.07
MC11	QPSK, RS(71,53)	8.0	1.07
MC12	QPSK, conv. 7/8	8.7	1.25
MC13	T16QAM (3/4)	10.3	2.14
MC14	T8PSK (2/3)	10.4	1.43
MC15	QPSK	12.0	1.43
MC16	QPSK, conv. 3/4, RS(80,64)	TBD	1.14
MC17	QPSK, RS(236,212)	TBD	1.59
MC18	QPSK conv. 4/5, RS(236,212)	TBD	1.99
MC19	QPSK conv. 3/4, RS(236,212)	TBD	2.12

FEC ENCODING

Data Scrambling
In digital communications, it is necessary to avoid transmission of periodic patterns of ones and zeros in order to obtain proper energy dispersal of the signal being transmitted. Spectral lines and a skewed spectrum can cause considerable problems with amplifiers located in the signal path leading to distortion and interference. The ASIC performs proper scrambling of the data by randomising the input data with a pseudo random number (PRN) generator. Usual implementations range from 8 to 15 cell shift registers and a modulo 2 adder as shown in **Figure 3**. Flexibility is achieved by making the polynomial and the seed of the PRN programmable. Two insertion points in the data path were chosen to support Standards such as DVB and CCSDS.

FEC Encoding

A set of FEC schemes has been selected by undertaking extensive design trades by employing simulation and analysis methods. Consideration also has been given to accommodate coding schemes specified in major Standards (e.g. DVB, CCSDS, etc.). The FEC encoding block consists of three modules, a Reed-Solomon encoder, an Interleaver, and a Convolutional encoder with each of these modules having a number of parameters programmable. Provision is also made to support punctured codes, trellis coded modulation (TCM) and pragmatic trellis coded modualtion(PTCM). With the set of encoder modules provided in the block, a number of block (RS), convolutional, concatenated and trellis coded schemes could be realised to suit the application under consideration. These encoding functions are built-in an ASIC which can be programmed at the factory. Note that any or all of these encoding modules can be bypassed in order to address a specific application need. The FEC encoder block includes:

Reed Solomon (R-S) Encoder:
The R-S encoder is parameterized as R-S (N,K,t), where N, K and t, correspond to the number of output symbols, number of input symbols, and the maximum number of correctable symbols respectively. Note that R-S(N-*i*, K-*i*,*t*) are shortened codes for $i > 0$. Two R-S code families are supported: The first family includes R-S(255-i, K-i,t), where i and K are programmable codes. Codes of this family comply with the DVB (R-S(204,188,8)) and the CCSDS (R-S(255,223,16)) Standards. The second family includes R-S(15-i,K-i,t), where i and K are programmable. This family provides codes with shorter block length for encoding packet headers on systems with end-to-end (embedded) coding schemes.

Interleaver
An interleaver is usually used to increase the error correction capability of the FEC by converting longer length burst errors into correctable block or random errors. An outer coder, such as the RS, can then recover the transmitted data and achieve very low bit error rate (BER). A byte-wise, programmable convolutional interleaver is implemented. A convolutional interleaver has a major advantage over a block interleaver since for approximately the same performance less memory is required for its implementation. **Figure 4** shows a functional block diagram of the convolutional interleaver. A programmable interface supports configuring interleavers with variable depth (I<=16) and unit delay (M<= 26). Standards supported include DVB, with I=12 and M=17.

Convolutional Encoder
A programmable convolutional encoder with constraint length 7 (K=7) and can support a variety of code rates, punctured codes, trellis and pragmatic trellis codes has been chosen. The encoder can be used as inner code in concatenated schemes. In addition for rate ½, 2/3, ¾, etc. codes are also useful for obtaining coding gains without the bandwidth expansion of the rate ½ code. A punctured code is the rate ½ code with coded bits systematically removed; these codes have almost the same performance as n/(n+1) codes and allows the same encoder and decoder to be used for a wide range of code rates. Again, by means of a programmable interface a variety of codes can be selected including rate-1/2 convolutional and rate-2/3, 3/4, 5/6, & 7/8 punctured, pragmatic trellis coded modulation (PTCM) (2 dimensional rate of 2/3 for 8-PSK and rate of ¾ for 16-QAM). **Figure 5** shows a rate ½ convolutional encoder functional block diagram.

DIGITAL FIR FILTER DESIGN

Square root of raised cosine (SRRC) filter is the most desirable choice for band limited digital satellite communication channels. The combination of matched transmit and receive SRRC filters maximize the signal to noise ratio (SNR) and enables transmission without inter-symbol interference (ISI). Typically FOR satellite communication applications, the SRRC filters in are implemented with a roll-off factor ranging from 20 to 60 percent.

Advantages of Digital filters: One of the major characteristics of digital filters is their ability to adapt to any bit rate (corresponding cut-off frequency) without modifications to the filter structure and the coefficients. Programmable digital finite impulse response (FIR) filters can also easily accommodate various roll-off factors and can be optimized to meet requirements in terms of pass band ripple and stop band attenuation. Furthermore, this flexibility can be used to compensate for (sin x)/x magnitude distortion caused by digital-to-analog conversion. Other advantages include perfect linear phase, immunity to ageing effects, and the ability to compensate for other impairments.

Filter Design Requirements: Table 3 shows the pulse shaping filter spectral requirements which were used to design the FIR filters and to perform the analysis. Note that the spectral performance requirements of Table 3 exceed the requirements of the DVB spectral mask. Improved stop band characteristics gives the flexibility to adapt to other transmission scenarios. An example of such a system would be a (Frequency Division Multiplexing (FDM) system using closely spaced carriers. In such a case, having better stop band attenuation could improve the performance in the presence of adjacent channel interference or improve channel spacing.

Table 3 Pulse shaping spectral requirements

Parameter	Requirements
Maximum Overshoot (Passband Ripple)	0.25 dB
Minimum Stop Band Attenuation (Stop Band Ripple)	40 dB

Filter Design Optimization: It should be noted that the performance of the pulse shaping filter depends primarily

on the following parameters: implementation structure, word length (coefficient quantization), filter length (truncation), excess bandwidth parameter (roll-off) and undersampling or aliasing (number of samples per symbol (SPS)). In order to achieve an optimum design in terms of performance, speed, power, and complexity (minimum number of gates), it is important to investigate the impact of each of these parameters. We have chosen the integral of the mean square error to optimise the design and understand the behaviour of some of these parameters. Several iterations were done to obtain an optimum design. **Figure 6** shows an example of the analysis that was performed on the number of taps versus the MSE for several SPS. It can be seen that the effect of truncating the impulse response is more significant than the performance gain by over-sampling. In other words, it is better to represent more symbols (truncate less) in comparison to have more samples per symbol to represent the waveform.

Filter Architecture Selection: Two major implementation architectures were examined:

Direct (look-up tables) LUT: In this structure, waveforms are created by using sequences of modulated symbols of length (L) as address to memory locations that contain samples of the filtered waveform for that sequence [2]. The direct LUT RAM requirements: RAM Size = Nsps x l^L x Word Size. Table 4 shows the memory requirements in terms of RAM size and estimated number of gates for the modulation schemes required to be supported.

Table 4: Memory size and gates for different schemes

Mod. Scheme	# of levels	Aperture Symbols	Word Size	SPS (N_{SPS})	RAM Size	# of Kbytes *15
QPSK	2	8(16 taps)	8 bits	2	1/2Kbyte	7.5K
8-PSK	4	8(16 taps)	8 bits	2	128Kbyte	1920K
16-QAM	4	8(16 taps)	8 bits	2	128Kbyte	1920K

Direct Form Structure: This structure is a derivation from the non-recursive difference equation:

$$y(n) = \sum_{k=0}^{M-1} h(k)x(n-k)$$

The structure has a complexity of M multiplications and M-1 additions per output point [1]. Thus, this structure requires lower memory size than direct LUT but requires more processing.

The selected filter architecture consists of a mixed design between the table look-up and the direct form structure. Various modes of operation were included to meet diverse applications requirements: Normal, Double Span, 4-SPS and Complex Mode. The digital filter can support baseband as well as passband pulse generation to a very low IF (digital upconversion).

BURST AND BEAM HOP CONTROL

The high speed modulator is designed to function in a continuos or burst mode of operation in conjunction with global beam systems as well as fixed and hopping multibeam systems. Hopping beam systems offer increased spectrum utilization and adopt to variations in beam traffic. While designing a modulator for beam-hopping mode operation, one need to consider several important issues including :

- Data latency in the modulator due to functional elements such as interleaver, convolutional encoder, FIR filter etc.
- Timing required for inserting preamble and postamble
- Input buffer requirements
- Modulator amplifier 'turn-on' and 'turn-off' delays
- Beam switching time
- Burst lengths, and burst efficiency

Reduction in the overall delay of data through the modulator would also result in a corresponding reduction in the modulator buffering requirements. **Figure 7** shows an overview of the beam-hopping operation.

Delay Elements

Since the modulator has many interleaving, coding, modulation and data rate options, the delay through each functional block will vary according to configuration chosen. Synchronizing modulator data destined for the same beam and/or bursts is important and knowledge of the functional delay of each element is required for seamless execution. **Figure 8** shows the delay model used for assessing the net delay through the modulator. τ_D represents the delay directly associated with the data path. This will include latency of functional blocks such as the FEC encoder, interleaver, symbol mapper and the FIR filter. τ_A represents the time allowance for bit insertion (preamble, postamble, frame marker, etc.). Knowing and combining the values of both τ_D and τ_A for each modulation scheme will allow minimum transmission delays and minimum buffer as well.

Other issues for consideration

Having established the functionality of burst and beam-hopping modes, other detailed signal definitions are required. This includes the source and timing of various control signals which must support these modes. With our approach, safety features which protect spectral limits (e.g. no unmodulated carrier emission), and override features which maintain operability in the event of a functional or partial control block malfunction has also been considered.

ASIC ARCHITECTURE

ASIC Functional Blocks: The key elements of ASIC functional blocks is shown in **Figure 9**. The ASIC accepts high speed data (OC3) input from a digital switch (or router or data multiplexor) and performs scrambling, coding, interleaving, frame formatting, symbol mapping, pulse-shaping and outputs digital samples for the transmit waveforms. It also includes a control module which takes control signals from switching /multiplexing sub-systems as well as from telemetry and burst control sub-systems.

These commands will help program certain key parameters of the modulator. Also, a number of parameters pertaining to individual functional blocks can be programmed at the factory. Note that each of the functional blocks can be enabled or disabled or their order of insertion changed and thus, fully programmable to suit the applications environment.

CONCLUSIONS

The design philosophy and parameter trade for a universal satellite modulator has been described briefly. The key element of the modulator is an intelligent ASIC which incorporates a number of programmable features and functions. The key product discriminators of this ASIC include, high and variable data rate (up to OC3) operation, programmable concatenated FECs, programmable scrambler and interleaver, high performance digital filter for wave-shaping, smart mapper for multiple modulation generation, burst and beam-hop mode controller, and precision power controller. The modulator will support a wide range of power-efficient (e.g. BPSK), bandwidth-efficient (e.g. 16-QAM), and power-bandwidth efficient (e.g. TCM) schemes and fully support the requirements of DVB, CCSDS, and ATM Standards.

ACKNOWLEDGEMENT

This development program was partially supported under Advanced Satcom Program of Canadian Space Agency (Scientific Authority: Communications Research Center).

REFERENCES

[1] J. G. Proakis, D. G. Manolakis, Digital Signal Processing, Principles, Algorithms, and Applications, Upper Saddle River, Prentice Hall, 1996, ch. 7,pp 503.

[2] M. Vanderaar, D. Mortensen, R.Bexten, N.Nguyen, A Low Complexity Digital Encoder-Modulator for High Data Rate Satellite BISDN Applications, 0-7803-3682-8/96, IEEE,1996, pp357-363.

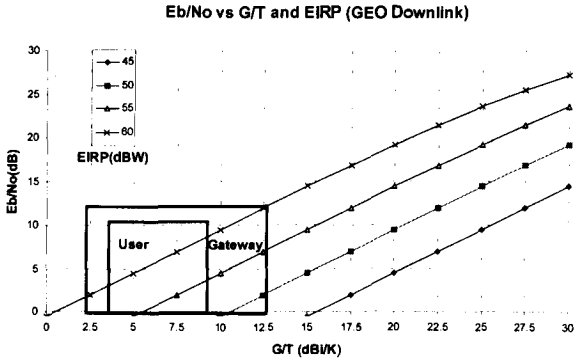


Figure 1 GEO Eb/No design bounds

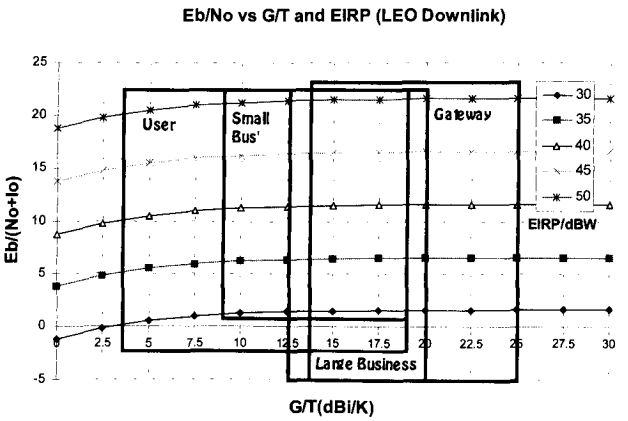


Figure 2 LEO Eb/No design bounds

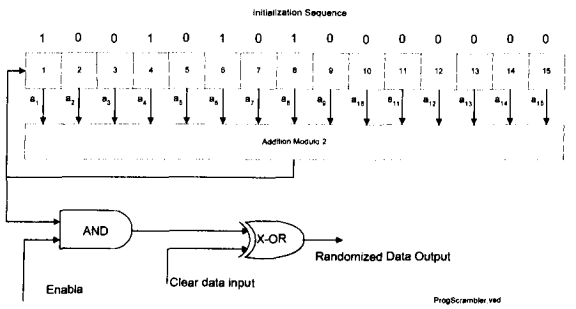


Figure 3 Programmable scrambler functional schematics

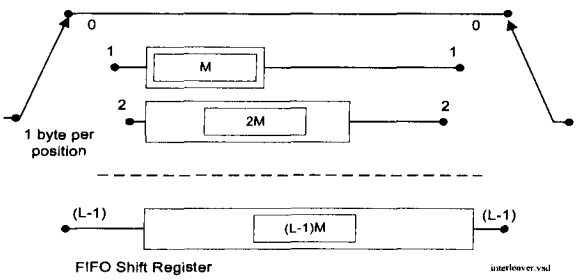


Figure 4 Programmable interleaver schematics

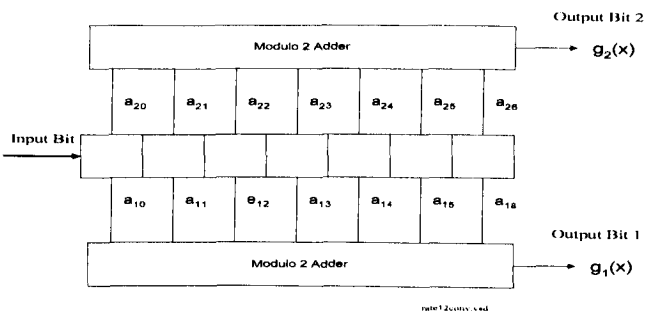


Figure 5 Rate ½ programmable convolutional encoder

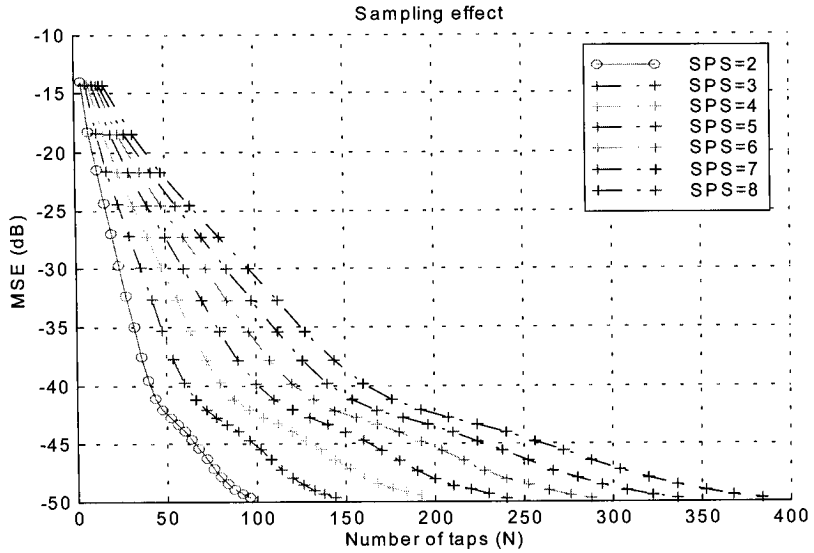


Figure 6 MSE analysis for number of taps for different samples per symbol (SPS)

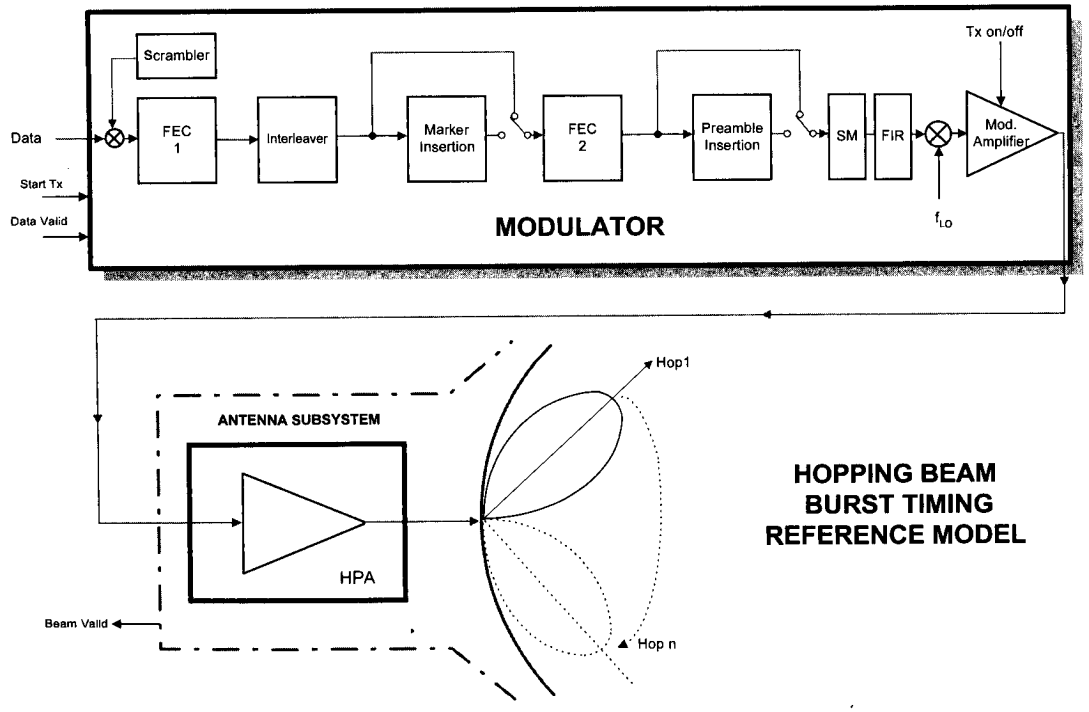


Figure 7 Hopping beam and burst timing reference model

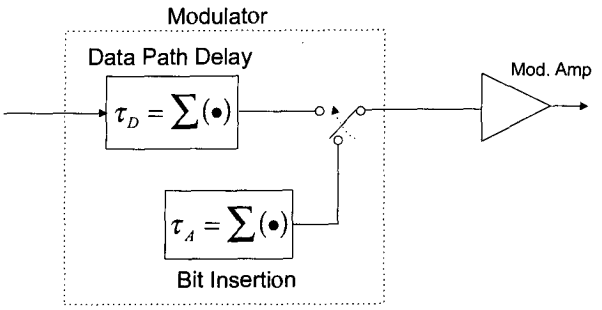


Figure 8 Delay model for various ASIC functional blocks

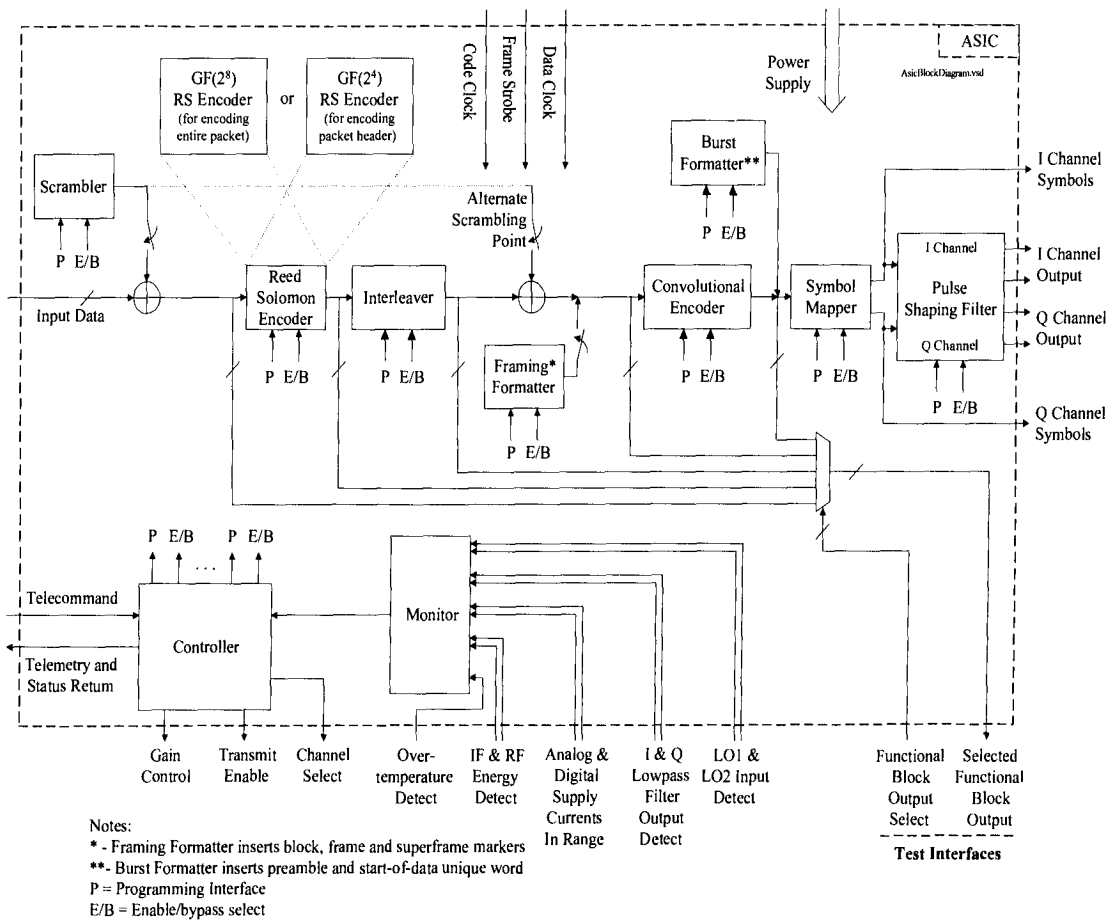


Figure 9 Functional schematics of modulator ASIC

Authors' Index

Date Due

29 MAY 02			

BRODART, CO. Cat. No. 23-233-003 Printed in U.S.A.