FINAL REPORT OF

A STUDY OF A JOINT STOCHASTIC PROCESS FOR AN INTEGRATED NETWORK

BY

R.H. KWONG

A. LEON-GARCIA

A.N. VENETSANOPOULOS

DEPARTMENT OF ELECTRICAL ENGINEERING
UNIVERSITY OF TORONTO
TORONTO, ONTARIO

31 March, 1980

Prepared for the Department of Communications, Canada, under Contract OSU79-00041

FINAL REPORT OF

A STUDY OF A JOINT STOCHASTIC PROCESS FOR AN INTEGRATED NETWORK

BY

R.H. KWONG

A. LEON-GARCIA

A.N. VENETSANOPOULOS

DEPARTMENT OF ELECTRICAL ENGINEERING
UNIVERSITY OF TORONTO
TORONTO, ONTARIO

31 March, 1980

Prepared for the Department of Communications, Canada, under Contract OSU79-00041

ABSTRACT

Stochastic processes arising in connection with a hybrid-switched integrated voice/data network are studied. Four cases are considered: exponentially distributed data packet lengths with finite or infinite buffer, and constant data packet lengths with finite or infinite buffer. Performance criteria are taken to be blocking probability for voice traffic, buffer overflow probability for data in the finite data buffer case, and the average data delay in the infinite data buffer case. An exact analysis is provided for each problem. Numerical difficulties associated with computing the exact solutions are then pointed out. This leads naturally to the consideration of the qualitative aspects of traffic behaviour. A simple and easy-to-compute fluid approximation is then given for evaluating the average data delay. It is shown to reflect the true qualitative traffic behaviour. It indicates, in particular, that large data queues will build up during heavy voice traffic conditions. A flow control procedure is then proposed for regulating the data queue. The performance of the hybrid switch with this flow control procedure is improved tremendously. The analysis of voice traffic with variable digitization rates is also given. Closed-form expressions for the voice traffic are obtained, which may have applications to flow control. Finally, the packet voice/data integration alternative is examined, and some methods for evaluating its performance are discussed.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

1.1 Transmission of Voice and Data in Conventional Networks

In conventional communication networks, voice and data traffic are handled differently because of their different characteristics and requirements. Voice traffic consists of long messages requiring continuous real time delivery with call duration of several minutes. This type of traffic, referred to as class I traffic, is transmitted mainly through circuit switching, in which an end-to-end circuit is established and maintained during the duration of a call. The cross-network delay is small and approximately constant. The last property is important for good quality of speech. If the network is busy, calls are then blocked. On the other hand, data traffic consists of either short messages requiring near real time delivery such as interactive data, referred to as class II traffic, or long messages requiring neither continuity nor real time delivery such as bulk data transfers, referred to as class III traffic. Data traffic is generally accepted for transmission, but it may experience delay. Interactive data is bursty in nature, where typically short messages alternate with long "think times". Such data traffic is now transmitted mainly through packet switching. In packet switching, data is grouped into packets, and each packet is routed independently through the network. The packets experience a variable crossnetwork delay with the packets arriving at their destinations possibly out of order. Long data messages can be handled using circuit or packet switching as it has no real time delivery requirements. Since long data messages will not play an important role in the evaluation of integration alternatives, we shall henceforth use the term data to refer to class II traffic only.

It is clear from the above description of traffic characteristics that it would be inefficient to transmit data by circuit switching. On the other hand, the transmission of voice traffic by packet switching is currently being considered. The motivation is that voice traffic, though less than interactive data, is still bursty in nature [1]. Hence it would be logical to consider using packet switching for voice also. There are, however, several difficulties that need to be overcome in packetized voice [1]. First, since voice traffic requires small cross-network delay, it is necessary that the delay incurred in routing the voice packets be small. If the network is congested, this may not be possible. In addition, the length of the packet must be designed carefully. If the packet is too short, a large percentage of the bits will be overhead bits, and inefficiency results. If the packet is too long, the packetization delay will also be long. Finally, since the packets may arrive at their destinations out of order, they must be reassembled to yield good speech quality. It is not yet clear how all these problems may be solved although packetized voice experiments seem to produce encouraging results [1]. For more detailed discussion on the various aspects of switching, see, for example, [2], [3].

1.2 The Case for an Integrated Voice and Data Network

To plan for future evolution of communication technology in which the variety and volume of traffic are expected to increase sharply, future networks must be able to provide efficient use of trunk capacity for a wide variety of transmission characteristics. Since the primary cost of a network is in the transmission segment, the integration of voice and data transmission in a common rather than two separate systems promises significant savings. Another advantage to integrating voice and data is the capability of providing interconnection between the broadest possible community of subscriber terminals. These factors have led to a growing interest in integrated networks. Its possibility has been studied in a number of papers dealing with the modelling and analysis of such networks, and some experimental integrated networks have been constructed [4]-[20].

1.3 Switching Alternatives in Integrated Networks

We have seen in Section 1.1 that circuit switching is inefficient for data. Thus we should not use solely circuit switching in integrated networks. This conclusion has been amply supported by cost and efficiency studies comparing circuit and packet switching schemes [18], [21]. Two approaches have so far emerged for designing integrated switching: the so-called hybrid switching scheme, and the voice/data packet switching scheme. In the hybrid switching scheme, both circuit and packet switching are provided by the network through a special time division multiplexing format whereby a frame of constant duration b is divided into two subframes, one dedicated to circuit switched traffic and the other to packet switched traffic [5], [7], [9]. The frame duration b is the same throughout the network in order to provide a nearly synchronous virtual path to circuit-switched traffic. The motivation for this switching scheme is to match the switching method to the traffic type. Thus voice traffic is circuit-switched while data is packet-switched. Within the hybrid switching scheme, there are also two methods of designing the subframes. In the so-called fixed boundary scheme, various traffic classes are not allowed to use idle bandwidth from the other subframe. In this case, the inefficiency of circuit switching remains with the circuit-switched portion of the frame. In the so-called movable boundary scheme, data traffic is allowed to use idle time slots in the circuit-switched subframe. Although this increases the complexity of the multiplexer, it is hoped that utilization of the channel will be enhanced. The performance of the voice traffic is analyzed using the probability of blocking criterion while that of data traffic is analyzed using the average delay criterion. One of the advantages of hybrid switching is that the user can decide on which service he would like to use. In the voice/data packet switching scheme, both voice and data are transmitted via packet switching. The motivation for this scheme is that, in addition to being able to utilize the burstiness of voice, voice and data can be handled in a unified manner. However, as voice requires a small cross-network delay, voice packets are given priority over data packets. In addition, it is necessary to use a "smoothing" buffer in order to deliver voice packets to their destination at a fixed delay and at a constant rate. Packets with crossnetwork delays larger than this fixed delay is considered lost. Past studies have shown that if 5% of the packets are lost, this causes noticeable but still acceptable degradation in the speech quality. This suggests that a suitable performance criterion for voice traffic is the percentage loss of packets [22]. For data traffic, the average delay criterion can be used.

There has not been a detailed comparative study of hybrid switching and packet switching in integrated networks, with perhaps the exception of [18]. Both seem to be viable schemes for integration, and each has its advantages and disadvantages. However, for the immediate future, it seems likely that the hybrid switching scheme may be more compatible with existing conventional networks [2].

1.4 Outline of Report

Our work is a contribution to the basic understanding of the interaction of voice and data traffic flow. We have focussed on some of the fundamental theoretical issues. Specifically, we analyze some queueing problems which arise naturally at the link level in integrated networks. Previous studies in this area include [5]-[7],[9],[10],[17],[19] dealing with the so-called hybrid-switched scheme, and [20], [22], [23] dealing with the voice/data packet-switched scheme. Although we have surveyed some of the work in voice/data packet-switched networks, we have concentrated our research on the hybrid-switched case, since this seems at this moment to be the most natural path of evolution towards integrated networks [2], [14]. It turns out that in the hybrid-switched case, earlier works contain inaccuracies [19]. Our contributions consist of more complete analyses of the stochastic processes involved, and the analyses of some simple yet useful flow control models. The organization of the report is given below.

In Chapter 2, we begin the presentation of our contributions. We first discuss in greater detail the mathematical models used in modelling the integrated switch. We pose the various analysis problems associated with stochastic processes arising in integrated networks. In particular, the queueing problems which naturally arise are pointed out. We then analyze in detail two queueing problems connected with hybrid switching. First we study the problem where data packet lengths are assumed to be exponentially distributed. With this assumption, the integrated switch can be analyzed using a two-dimensional birth and death process. We characterize the steady state joint distribution of voice and data traffic for the finite data buffer case as well as the infinite data buffer case. In the finite buffer case, we find that the steady state joint distribution can be obtained by solving a set of linear equations. However, for even a moderate number of voice and data channels and buffer size, the dimension of the system of equations is large. We also examine the performance criteria associated with this system: the blocking probabilities for voice and data (a data packet is blocked when it cannot be assigned a buffer space). The implications towards buffer memory managment are also discussed. In the infinite data buffer case, the steady state distribution for voice customers and the steady state generating function for data as a function of voice customers are characterized. Their complete determination, however, requires the computation of roots of certain polynomials inside the unit interval. The numerical difficulties involved prompt us to study approximate, rather than exact, analysis of the queueing problem. We survey some results available in the literature and propose also a simple fluid approximation for the problem. When we apply these results to the computation of the performance criterion:

the average delay for data packets, we find that the fluid approximation gives very reasonable results. These results are also compared with the simulation results reported in [19]. In general, the data queue builds up whenever the voice calls seize a large enough number of channels, leaving an insufficient number of channels for the transmission of data packets. This suggests that some flow control scheme is necessary to alleviate the congestion in the data queue. One such scheme will later be discussed.

The analysis of the queueing problem when the data packets have constant lengths is substantially different from the one for exponential packets. The Markovian nature of the process is no longer present and we have to take into account explicitly the frame structure. We use the same model as that considered by Fischer and Harris [9]. However, as was pointed out in [19], the analysis in [9] contains an error which invalidates the conclusions given. We here give an analysis of this case by considering jointly the voice and data traffic. We first study the case where the data buffer is finite. We show that the solution for the steady state distribution is again governed by the solution of a matrix equation. However, the number of nonzero entries in the matrix is much larger than the case of exponentially distributed packets, although the dimension of the matrix is reduced. In fact, in general, this matrix is essentially full. Hence numerical problems again would arise even for moderately sized problems. The case where the buffer size is infinite has been a long-standing problem in integrated networks. Here we present the correct solution to this problem by characterizing the steady state generating function for data for a given number of voice customers. Using the generating function, again one can compute the average delay for data. Analytical solution is available only for the simplest example. In the general case, the complete solution requires the determination of the roots of certain analytic functions inside the unit circle. Just as in the case of exponentially distributed packets, approximations are necessary for gaining insight into the nature of the solution. The results for the constant packet case are also compared to those for exponentially distributed packets, and the similarities and differences between the two cases are pointed out.

Similar to the movable boundary hybrid-switched system is the variable frame scheme. Here, instead of having a fixed frame, the frame size varies according to the traffic conditions. However, a maximum value is imposed for the frame size. Circuit-switched terminals are assumed to be permanently connected to the network and they are transmitted through the circuit-switched portion of the frame, where the number of voice slots is assumed to be random. Similarly, the number of slots in the packet-switched portion of the frame is also assumed to be random. This model has been suggested by Miyahara and Hasegawa in [15]. Here we briefly examine one possible method of analysis for this system.

The analysis of the joint voice and data process shows clearly that during periods with heavy voice traffic, the data queue builds up rapidly, giving rise to long delays. To reduce the data delay, some flow control mechanism is necessary. In Chapter 3, we examine a very simple flow control procedure involving the regulation of data flow into the link based on the number of voice customers present in the system. By assuming the packet lengths to be exponentially distributed, we are able to obtain expressions for the various performance parameters using two-dimensional birth-death analysis.

Since future communication requirements will likely involve voice traffic with different digitization rates [2], we analyze the queueing problem for this situation in Chapter 4. The blocking probability is found using a multidimensional birth-death analysis. The implications of these results to the control of voice digitization rates are then discussed. We can also consider a flow control procedure combining both direct data flow control and the control of digitization rates.

While we have not made any in-depth study of the packetized voice/data integration scheme, we have surveyed some of the literature. In Chapter 5, we report some of these results and compare the models and assumptions used in all-packet systems to those in hybrid circuit-packet systems. In particular, the differences in the performance criteria and their implications are examined.

Finally, in Chapter 6, we make some general conclusions on the current status of research in the integration of voice and data in communication networks. Our work is by no means comprehensive and there are a large number of extensions we can pursue. Some of the future research directions are also sketched.

CHAPTER 2

ANALYSIS OF THE STOCHASTIC PROCESSES ARISING IN HYBRID-SWITCHED NETWORKS

2.1 Mathematical Model for the Integrated Switch

In this chapter, we form our attention on hybrid-switched systems. We first describe briefly the multiplex structure used in hybrid switching. Time slices of fixed size, termed a frame, are allocated to the transmission of digitized voice and data packets. Each frame is divided into two portions, one allocated to voice traffic, the other to data. The voice traffic subframe is divided further into slots. Each active call has one or more slots reserved for its transmission. The number of slots siezed by a call is proportional to its voice digitization rate. For a more detailed description of the frame structure, see for example [5], [9].

Voice traffic is either accepted or rejected, with small connection delays and no error control. It represents a loss system in which if a connection cannot be established between the source and the destination, the call is blocked. The size of the voice subframe is designed so that the blocking probability for voice traffic is small enough to meet the performance requirements. Data traffic, on the other hand, is always accepted, if an infinite data buffer is assumed available, and is transmitted in packet form during the remainder of the frame. Data packets will experience a crossnetwork delay. Since reasonably short delays are generally required for data, data traffic performance is evaluated on the basis of average delay incurred. If the data buffer is finite, the data performance parameter is the probability of buffer overflow. To enchance link utilization, we assume the movable boundary scheme is adopted, in which data traffic is allowed to use any residual voice capacity available due to statistical variations in the voice traffic.

There are basically two different models for the voice and data traffic that we consider. In both of these models, the voice and data arrivals processes are assumed to be Poisson with parameters λ_1 and λ_2 respectively, and the holding time distribution for voice is assumed to be exponential with mean $\frac{1}{\mu_1}$. In the first model, the service time for data is assumed to be exponential with mean $\frac{1}{\mu_2}$. This effectively means that the packet lengths are assumed to be exponentially distributed with mean length $\frac{1}{\mu_2}$. In the second model, the serive time for data is assume to have a deterministic distribution $F(x)$, where $F(x)=0$ for $x<b$ and $F(x)=1$ for $x\geq b$. The number of voice and data channels in a frame is assumed to be N_V and N_D, respectively. We assume that all parameters are expressed in consistent units.

The multiplex structure described at the beginning of this section can be modelled as a queueing system with two types of arrivals, voice and data, and an operating rule that allows these customers access to the system (referred to as opening the gate). An arriving voice customer waits in a buffer until the next opening of the gate. If the number of free voice channels is greater than the number of voice customers ahead of him, he receives service. If not, he is lost and leaves the system. For the data traffic, arrivals are buffered, and at the opening of the gate, placed on the available data channels on a first come, first served basis. We assume the voice buffer to have infinite capacity. The data buffer may be assumed to have finite capacity, so that overflow can occur, or it may be assumed to have infinite capacity.

The queueing problems to be addressed may now be formulated as follows: given the data buffer capacity, find the blocking probability for voice traffic, and the average delay for data traffic in the case of infinite data buffer, and the buffer overflow probability in the case of a finite buffer. In the remaining sections of this chapter, we shall analyze the various queueing problems that arise.

2.2 Analysis of Traffic Behaviour when Data Packet Lengths are Exponentially Distributed with Finite Data Buffers

We first consider the case where data lengths are exponentially distributed with mean length $\frac{1}{\mu_2}$, and the data buffer contains M spaces. We also assume, for simplicity, that there is a basic slot size, and that N_V slots are available for voice and N_D slots for data, so that the total capacity is $N=N_V+N_D$ slots. If we ignore the time quantization introduced by the frame structure, we can then model the system as a two-dimensional Markov chain. We state the results obtained precisely as follows.

Let the voice and data arrival processes be Poisson with parameters λ_1 and λ_2 respectively, and that the service distribution be exponential with rates μ_1 and μ_2 respectively. The service and arrival distributions are assumed to be independent. Let Q_V and Q_D be the number of voice and data customers in the steady state, respectively. Define $P_{i,j}=P_r\{Q_V=i, Q_D=j\}$. We assume that voice calls have priority; that is, if there are less than N_V voice calls occupying the channel, any additional voice call that cannot find a free channel will preempt a data customer. The preempted data customer returns to the data buffer. We assume that there are M spaces in the data buffer. Then the steady state equations for π_{ij} are given by:

For $i = 0,1,\ldots,N_V-1$,

$$(\lambda_1+\lambda_2+i\mu_1+j\mu_2)P_{i,j} = \lambda_1 P_{i-1,j}+(i+1)\mu P_{i+1,j}+\lambda_2 P_{i,j-1}+(j+1)\mu_2 P_{i,j+1}$$
$$j=0,1,\ldots,N-i-1 \tag{2.2.1}$$

$$[\lambda_1+\lambda_2+i\mu_1+(N-i)\mu_2]P_{i,j} = \lambda_1 P_{i-1,j}+(i+1)\mu_1 P_{i+1,j}+\lambda_2 P_{i,j-1}$$
$$+ (N-i)\mu_2 P_{i,j+1} \qquad N-i\le j\le M+N-i-1 \tag{2.2.2}$$

$$[\lambda_1+i\mu_1+(N-i)]P_{i,j} = \lambda_1 P_{i-1,j}+\lambda_2 P_{i,j-1}+\lambda_1 P_{i-1,j+1}$$
$$j=M+N-i \tag{2.2.3}$$

For $i = N_V$, we get

$$(\lambda_2+N_V\mu_1+j\mu_2)P_{N_V,j} = \lambda_1 P_{N_V-1,j}+\lambda_2 P_{N_V,j-1}+(j+1)\mu_2 P_{N_V,j+1}$$
$$j=0,1,\ldots,N_D-1 \tag{2.2.4}$$

$$(\lambda_2+N_V\mu_1+N_D\mu_2)P_{N_V,j} = \lambda_1 P_{N_V-1,j}+\lambda_2 P_{N_V,j-1}+N_D\mu_2 P_{N_V,j+1}$$
$$N_D\le j\le M+N_D-1 \tag{2.2.5}$$

and

$$(N_V\mu_1+N_D\mu_2)P_{N_V,j} = \lambda_1 P_{N_V-1,j}+\lambda_2 P_{N_V,j-1}+\lambda_1 P_{N_V-1,j+1}$$
$$j=M+N_D \tag{2.2.6}$$

In Eq. (2.2.1)-(2.2.6), we define $P_{-1,j}=P_{N_V+1,j}=P_{i,-1}=P_{i,M+N+1-i}=0$.

These equations, together with the normalization condition $\sum_{i,j} P_{i,j}=1$, uniquely determine the steady state probabilities $P_{i,j}$.

As an example, we solve for the $\pi_{i,j}$'s in the case where $N_V=1$, $N_D=0$, M=1. Equations (2.2.1)-(2.2.6) reduce to

$$(\lambda_1+\lambda_2)P_{0,0} = \mu_1 P_{1,0} + \mu_2 P_{0,1}$$

$$(\lambda_1+\lambda_2+\mu_2)P_{0,1} = \mu_1 P_{1,1} + \lambda_2 P_{0,0} + \mu_2 P_{0,2}$$

$$(\lambda_1+\mu_2)P_{0,2} = \lambda_2 P_{0,1}$$

$$(\lambda_2+\mu_1)P_{1,0} = \lambda_1 P_{0,0}$$

$$\mu_1 P_{1,1} = \lambda_1 P_{0,1} + \lambda_2 P_{1,0} + \lambda_1 P_{0,2}$$

with normalization condition $\sum_{i,j} P_{i,j}=1$.

The solution of these equations is given by

$$P_{0,0} = \left[1 + \frac{\lambda_1\lambda_2+\lambda_2^2+\lambda_2\mu_1}{(\lambda_2+\mu_1)\mu_2} + \frac{\lambda_1\lambda_2^2+\lambda_2^3+\lambda_2^2\mu_1}{(\lambda_1+\mu_2)(\lambda_2+\mu_1)\mu_2} + \frac{\lambda_1}{\lambda_2+\mu_1} + \frac{\lambda_1^2\lambda_2+\lambda_1\lambda_2^2+\lambda_1\lambda_2\mu_1}{\mu_1(\lambda_2+\mu_1)\mu_2} + \frac{\lambda_1\lambda_2}{\mu_1(\lambda_2+\mu_1)}\right]^{-1} \tag{2.2.7}$$

$$P_{0,1} = \frac{\lambda_1\lambda_2+\lambda_2^2+\lambda_2\mu_1}{(\lambda_2+\mu_1)\mu_2} P_{0,0} \tag{2.2.8}$$

$$P_{0,2} = \frac{\lambda_2}{\lambda_1+\mu_2} P_{0,1} \tag{2.2.9}$$

$$P_{1,0} = \frac{\lambda_1}{\lambda_2+\mu_1} P_{0,0} \tag{2.2.10}$$

$$P_{1,1} = \frac{\lambda_1}{\mu_1} P_{0,1} + \frac{\lambda_2}{\mu_1} P_{1,0} \tag{2.2.11}$$

For $\lambda_1=0.01$, $\lambda_2=40$, $\mu_1=0.01$, $\mu_2=100$, we get

$$P_{0,0} = 0.337812281$$
$$P_{0,1} = 0.135158685$$
$$P_{1,0} = 8.44319622\times 10^{-5}$$
$$P_{1,1} = 0.472886534$$
$$P_{0,2} = 0.0540580682$$

If we count the total number of equations defined by (2.2.1)-(2.2.6), we find that there are $\frac{(N_V+1)}{2}(2M+2N_D+N_V+2)$ equations. For $N_V=10$, $N_D=5$, M=10, this gives 231 equations. Thus the total number of equations increases quite rapidly with increasing N_V, N_D and M. However, because of the special form of the equations, efficient iterative methods can be used for their solution [24].

The above model has previously been considered by Weinstein et al. in [19]. However, their results appear to contain errors. The correct steady state equations should be equations (2.2.1)-(2.2.6).

The steady state blocking probability P_L for voice is given simply by the Erlang B formula

$$P_L = \frac{\left(\frac{\lambda_1}{\mu_1}\right)^{N_V}}{N_V!} \left[\sum_{k=0}^{N_V} \frac{\left(\frac{\lambda_1}{\mu_1}\right)^{k}}{k!}\right]^{-1} \tag{2.2.12}$$

The buffer overflow probability P_B is given by

$$P_B = \sum_{i=0}^{N_V} P_{i,M+N-i} \tag{2.2.13}$$

where the $P_{i,j}$'s are the steady state probabilities given by (2.2.1)-(2.2.6). For the example where $N_V=1$, $N_D=0$, $M=1$ discussed above

$$P_L = \frac{1}{2}$$

and $$P_B = 0.526944602$$

These results may be used to design the size of the data buffer to achieve a certain buffer overflow probability. The model may also be modified to handle priority traffic. However, there does not seem to be a simple way of relating the buffer size to the overflow probability. For each set of traffic parameters, numerical solution of the equations (2.2.1)-(2.2.6) would be needed to determine the overflow probability. Some buffer management schemes are discussed in [17]. We shall not go into those details here.

The case where voice traffic cannot preempt data can be treated in the same framework. The resulting equations are simple modifications of (2.2.1)-(2.2.6), and the details are omitted.

In the above, we have treated the queueing process when the data buffer is finite. If the storage memory is assumed to be large compared to queue lengths, approximating it as an infinite size buffer would be valid. In the next section, we shall treat the infinite data buffer problem.

2.3 Analysis of Traffic Behaviour when Data Packet Lengths are Exponentially Distributed with Infinite Data Buffer

In this section, we study the stochastic process connected with hybrid switching for exponentially distributed data packets with infinite data buffer. This problem has been considered previously in [17] and [19]. In particular, [17] provides a rather detailed discussion of the queueing problem. Our inclusion of this material here is for completeness as well as for ease of comparison with the constant packet case discussed later.

We use the same notation and mathematical model as in Section 2.2 except that here, an infinite data buffer is assumed available. Then the steady state balance equations are given by:

For $i=0,1,\dots,N_v-1$, we have

$$(\lambda_1+i\mu_1+\lambda_2+j\mu_2)P_{i,j} = \lambda_1 P_{i-1,j}+(i+1)\mu_1 P_{i+1,j}+\lambda_2 P_{i,j-1}+(j+1)\mu_2 P_{i,j+1} \qquad 0\le j\le N-i-1 \tag{2.3.1}$$

$$[\lambda_1+i\mu_1+\lambda_2+(N-i)\mu_2]P_{i,j} = \lambda_1 P_{i-1,j}+(i+1)\mu_1 P_{i+1,j} + \lambda_2 P_{i,j-1}+(N-i)\mu_2 P_{i,j+1} \qquad j\ge N-i \tag{2.3.2}$$

For $i=N_v$, we have

$$(i\mu_1+\lambda_2+j\mu_2)P_{i,j} = \lambda_1 P_{i-1,j}+\lambda_2 P_{i,j-1}+(j+1)\mu_2 P_{i,j+1} \qquad 0\le j\le N-N_v-1 \tag{2.3.3}$$

$$[i\mu_1+\lambda_2+(N-i)\mu_2]P_{i,j} = \lambda_1 P_{i-1,j}+\lambda_2 P_{i,j-1}+(N-i)\mu_2 P_{i,j+1} \qquad j\ge N-N_v \tag{2.3.4}$$

In these equations, we have taken $P_{-1,j}=P_{N_v+1,j}=P_{i,-1}=0$.

Let $P_i=\sum_{j=0}^{\infty} P_{i,j}$. By summing over j in (2.3.1)-(2.3.4), it is straightforward to show that

$$P_i = \frac{\left(\frac{\lambda_1}{\mu_1}\right)^i}{i!}\left[\sum_{k=0}^{N_v}\left(\frac{\lambda_1}{\mu_1}\right)^k \frac{1}{k!}\right]^{-1} \tag{2.3.5}$$

Thus the steady state distribution for voice traffic is the same as that of an $M/M/N_v/N_v$ system, and is independent of the data traffic.

To analyze the data traffic, define the generating functions

$$\pi_i(z) = \sum_{j=0}^{\infty} P_{i,j}z^j \qquad i=0,1,\ldots,N_v$$

By multiplying (2.3.1) and (2.3.2) by z^j and summing over j, we get, after some calculations

$$\begin{aligned}
&[\lambda_1+i\mu_1+\lambda_2+(N-i)\mu_2]\pi_i(z)\\
&=\lambda_1\pi_{i-1}(z)+(i+1)\mu_1\pi_{i+1}(z)+[\lambda_2 z+\tfrac{1}{z}(N-i)\mu_2]\pi_i(z)\\
&+\sum_{j=0}^{N-i-1}[i+j+1-N]\mu_2 P_{i,j+1}z^j-\sum_{j=0}^{N-i-1}(i+j-N)\mu_2 P_{i,j}z^j\\
&-\frac{1}{z}P_{i,0}(N-i)\mu_2 \qquad i=0,1,\ldots,N_v-1
\end{aligned} \tag{2.3.6}$$

Similarly for $i=N_v$, we get

$$[N_V\mu_1+\lambda_2+(N-N_V)\mu_2]\pi_{N_V}(z) = \lambda_1\pi_{N_V-1}(z)+[\lambda_2 z+\frac{1}{z}(N-N_V)\mu_2]\pi_{N_V}(z)$$

$$+\sum_{j=0}^{N-N_V-1}(N_V+j+1-N)\mu_2 P_{N_V,j+1}z^j - \sum_{j=0}^{N-N_V-1}(N_V+j-N)\mu_2 P_{N_V,j}z^j$$

$$-\frac{1}{z}P_{N_V,0}(N-N_V)\mu_2 \qquad (2.3.7)$$

Equations (2.3.6) and (2.3.7) may be combined into a matrix equation of the form

$$A(z)\pi(z) = b(z) \qquad (2.3.8)$$

where $\pi(z)'=[\pi_1(z)\ \pi_2(z)\ \ldots\ \pi_{N_V}(z)]$, $A(z)$ is an $N_V \times N_V$ matrix consisting of the coefficients of the $\pi_i(z)$'s in equations (2.3.6) and (2.3.7), and $b(z)$ is a vector containing the unknowns $P_{i,j}$, $i=0,\ldots,N_V$, $j=0,\ldots,N-i-1$. Note that $A(z)$ is a tridiagonal matrix. To solve (2.3.8), we need to determine the unknowns defining the vector $b(z)$. Equations (2.3.1) and (2.3.2) may be used recursively to express the unknown $P_{i,j}$'s for $i=1,\ldots,N_V$, $j=0,1,\ldots,N-i-1$, in terms of $P_{0,j}$'s for $j=0,1,\ldots,N$. Thus we need $N+1$ equations to determine the $P_{0,j}$'s. We can obtain $N-N_V$ equations from (2.3.3). One equation can be found by the condition that in the steady state, the carried load must equal the expected number of busy servers (or equivalently, the condition that $\sum_{i,j} P_{i,j}=1$). The remaining N_V equations are obtained by first finding the N_V unique roots of det $A(z)$ in (0,1) and then requiring det $R_i(z)$ to also vanish at these points, where $R_i(z)$ is the matrix obtained by replacing the ith column in $A(z)$ with $b(z)$. The above method of solution is described in greater detail in [17] for a slightly different formulation, where, in particular, the uniqueness of the roots is det(A(z)) in (0,1) is proved. It is similar to the approach described in [25].

After $\pi_i(z)$ has been found, the expected number of data customers is given by

$$E(Q_D) = \sum_{i=0}^{N_V} \pi_i'(1) \tag{2.3.9}$$

By Little's Theorem [26], the average data delay is given by

$$E(W_D) = \frac{1}{\lambda_2} E(Q_D) \tag{2.3.10}$$

The above results in principle enables us to solve the hybrid switch queueing problem for exponentially distributed data lengths with infinite data buffer. However, it is difficult to apply even for moderate values of N_V and N_D because of the numerical difficulties involved. In particular, since the parameters (λ_1,μ_1) and (λ_2,μ_2) generally differ from each other by several orders of magnitude, roundoff errors in the computations become very significant. Chang has illustrated the numerical difficulties with examples in [17]. Even if extended precision is used, it is difficult to relate the system performance to the system parameters owing to the complicated calculations involved. In later sections, we discuss some approximation methods which are simpler to apply, and which give the same qualitative traffic behaviour as the exact solution.

It is possible, however, to solve (2.3.8) explicitly in the very simple special case of N_V=1, N_D=0. The solution is given in [19], which we quote here:

$$E(W_D) = \frac{\rho_1(1+\rho_1)^2+\rho_1\lambda_2/\mu_1}{\lambda_2(1+\rho_1)(1-\rho_2-\rho_1\rho_2)} \tag{2.3.11}$$

where $\rho_1 = \dfrac{\lambda_1}{\mu_1}$ and $\rho_2 = \dfrac{\lambda_2}{\mu_2}$.

As an example, if we take λ_1=0.01, μ_1=0.01, μ_2=100, and plot $E(W_D)$ against λ_2, we obtain the graph given in Figure 1. The large data delay is due mainly to the presence of the λ_2/μ_1 factor in the numerator. This factor can be interpreted as the expected number of data packet arrivals during an average voice holding time. Since λ_2 is generally several orders of magnitude larger than μ_1, λ_2/μ_1 can be quite large. These results will be compared later with those for constant data packet lengths, where similar data delay characteristics are also obtained.

2.4 Analysis of Traffic Behaviour when Data Packet Lengths are Constant with Finite Data Buffer

In the previous two sections, we have analyzed the voice and data traffic behaviour when data packet lengths are exponentially distributed. The situation for constant data packet lengths is mathematically substantially different. In this case, it is necessary to take into account explicitly the frame structure associated with the hybrid switch. Our model is similar to that given in [9] for the infinite data buffer problem. In this section, we treat the finite data buffer problem, while the infinite data buffer problem is treated in the next section.

Let us recall the hybrid switch model for the constant packet case. The voice and data arrivals are assumed to be Poisson with parameters λ_1 and θ (this is used instead of λ_2 to distinguish the constant packet case from the exponentially distributed case), respectively. The voice service time is assume to be exponential with mean $\frac{1}{\mu_1}$, and the data service time is assumed to be deterministic with distribution function $F(x)=0$ for $x<b$,

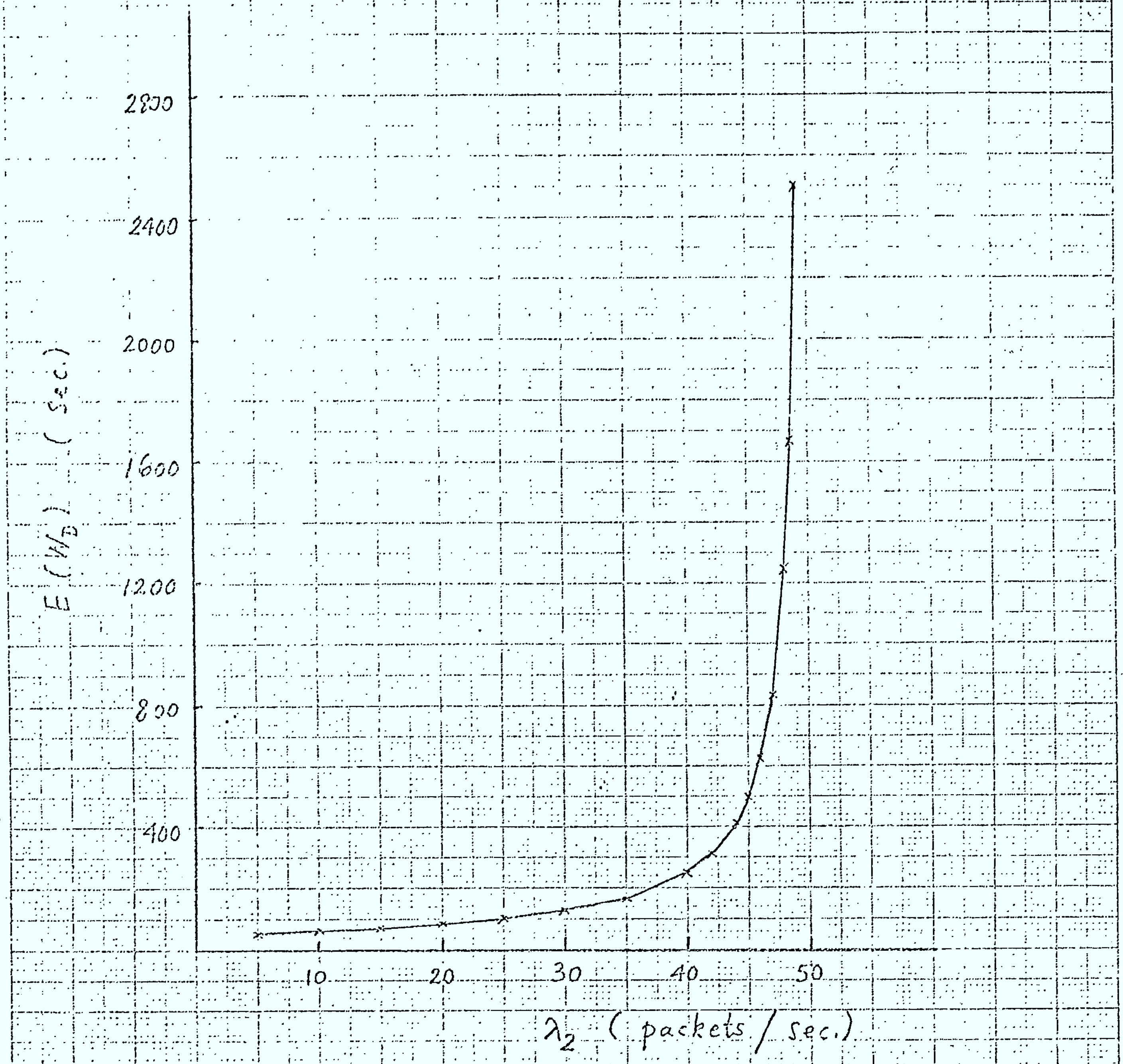


Figure 1. Average data delay as function of data load (exponential data packets): $N_V = 1$, $N_D = 0$,
$\lambda_1 = 0{\cdot}01$, $\mu_1 = 0{\cdot}01$, $\mu_2 = 100$.

and $F(x)=1$ for $x\geq b$. Every b seconds, the gate opens to allow voice and data traffic access to the channels.

Since the service time for data is deterministic, we can no longer model the joint voice and data process as a two-dimensional birth and death process. However, we may still model the joint stochastic process as a two-dimensional Markov chain. To do so, let us define p_i^V to be the probability of i voice arrivals in a length of time b, q_{kj}^V the probability of j busy voice channels just before the opening of the gate given that there were k present just after the last opening, and p_i^D the probability of i data customer arrivals in a length of time b. By our assumptions, we have

$$p_i^V = e^{-\lambda_1 b} \frac{(\lambda_1 b)^i}{i!} \tag{2.4.1}$$

$$q_{kj}^V = \binom{k}{j}\left[1-e^{-\mu_1 b}\right]^{k-j}\left[e^{-\mu_1 b}\right]^j \tag{2.4.2}$$

$$p_i^D = e^{-\theta b} \frac{(\theta b)^i}{i!} \tag{2.4.3}$$

It was shown in [9] that the queueing system is stable when $\theta b<N-E(Q_V)$ where Q_V is the number of voice customers in the system in the steady state. In [9], the steady state distribution of Q_V is obtained. Their results are as follows:

Let π_i^V be the steady state probability of having i busy voice channels just before the gate opens. Then π_i^V, $i=0,1,\ldots,N_V$, are determined by the equations

$$\pi_j^V = \sum_{i=0}^{N_V} \pi_i^V P_{ij}^V \qquad j=0,1,\ldots,N_V \tag{2.4.4}$$

together with the normalization condition

$$\sum_{j=0}^{N_v} \pi_j^v = 1 \tag{2.4.5}$$

where P_{ij}^v are the transition probabilities given by

$$\begin{aligned} P_{ij}^v &= \sum_{k=\max(i,j)}^{N_v-1} q_{kj}^v p_{k-i}^v + q_{N_v j}^v \sum_{k=N_v-i}^{\infty} p_k^v \quad \text{for } i=0,1,\ldots,N_v-1 \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j=0,1,\ldots,N_v \\ &= q_{N_v j}^v \qquad \text{for } i=N_v,\ j=0,1,\ldots,N_v \end{aligned} \tag{2.4.6}$$

The blocking probability P_L for voice is determined as follows. Let the number of voice customers between any two successive gate openings be denoted by ERS. Then we have

$$ERS = \sum_{j=0}^{N_v-2} \pi_j^v \left\{ \sum_{k=1}^{N_v-j-1} kp_k^v + (N_v-j)\left[1 - \sum_{k=0}^{N_v-j-1} p_k^v\right]\right\} + \pi_{N_v-1}^v (1-p_0^v) \tag{2.4.7}$$

And P_L is given by

$$P_L = 1 - \frac{ERS}{\lambda_1 b} \tag{2.4.8}$$

In Eq. (2.4.4), one equation from that system is in fact redundant. Thus (2.4.4) and (2.4.5) represent N_v+1 linear equations in N_v+1 unknowns. They may be solved using standard numerical methods for solving linear systems of equations.

It is also noted in [9] that the blocking probability P_L is well-approximated by the Erlang B formula (2.3.5) provided $\lambda_1 b$ is sufficiently small. In the examples that we examine, this will indeed be the case, and we shall usually take the Erlang B formula, which is much more readily evaluated then (2.4.8), to estimate P_L.

We now turn to the analysis of the data traffic behaviour. In this section, we assume that the data buffer is finite with M spaces, and we take the performance criteria to be the buffer overflow probability. To correctly analyze the data traffic, we need to consider the joint distribution of voice and data.

Let us define Q_t^V and Q_t^D to be the number of voice and data customers, respectively, in the channel immediately after the t^{th} opening of the gate. We can then write down the following conditional probabilities for Q_t^V:

$$P_r\{Q_t^V=m|Q_{t-1}^V=j,\ Q_{t-1}^D=i\} = \sum_{k=0}^{\min(j,m)} q_{jk}^V p_{m-k}^V \qquad m=0,1,\ldots,N_V-1 \tag{2.4.9}$$

$$P_r\{Q_t^V=N_V|Q_{t-1}^V=j,\ Q_{t-1}^D=i\} = \sum_{k=0}^{j} [q_{jk}^V \sum_{m=N_V}^{\infty} p_{m-k}^V] \tag{2.4.10}$$

Similarly, we obtain the following conditional probabilities for Q_t^D:

$$P_r\{Q_t^D=n|Q_t^V=m,\ Q_{t-1}^D=i,\ Q_{t-1}^V=j\}$$

$$= \frac{e^{-\theta b}(\theta b)^n}{n!} \qquad \text{for } i\le N-j-1 \text{ and } n<M$$

$$= \frac{e^{-\theta b}(\theta b)^{n-i+N-j}}{(n-i+N-j)!} \qquad \text{for } N-j\le i\le M \text{ and } i+j-N\le n<M$$

$$= \sum_{k=M}^{\infty} \frac{e^{-\theta b}(\theta b)^k}{k!} \qquad \text{for } i\le N-j-1,\ n=M$$

$$= \sum_{k=M}^{\infty} \frac{e^{-\theta b}(\theta b)^{k+N-i-j}}{(k+N-i-j)!} \qquad \text{for } N-j\le i\le M,\ i+j-N\le n=M$$

$$= 0 \qquad \text{otherwise} \tag{2.4.11}$$

In (2.4.11), if $i\le N-j-1$, then all the data customers received service in the $(t-1)^{st}$ gating period. Thus the number of data customers in the system (channel plus buffer) after the opening of the gate is the same as the number of data arrivals during the gating period. If $i\ge N-j$, then $i+j-N$ customers are still in the system at the t^{th} opening of the gate. To make up to n data customers, we therefore need $n+N-i-j$ arrivals. The same reasoning applies to the derivation for the case where $n=M$.

The joint transition probabilities are now given by

$$P_r\{Q_t^V=m,\ Q_t^D=n|Q_{t-1}^V=j,\ Q_{t-1}^D=i\}$$

$$= P_r\{Q_t^D=n|Q_t^V=m,\ Q_{t-1}^V=j,\ Q_{t-1}^D=i\}\ .$$

$$P_r\{Q_t^V=m|Q_{t-1}^V=j,\ Q_{t-1}^D=i\} \tag{2.4.12}$$

Note that the right hand side of (2.4.12) is independent of t. We may therefore denote the joint transition probabilities by $p_{ji,mn}$. They are given by

$$
\begin{aligned}
p_{ji,mn} &= \sum_{k=0}^{\min(j,m)} q^v_{jk} p^v_{m-k} \frac{e^{-\theta b}(\theta b)^n}{n!} && \text{for } \begin{array}{l} i\le N-j-1 \\ m<N_v \\ n<M \end{array} \\
&= \sum_{k=0}^{\min(j,m)} q^v_{jk} p^v_{m-k} \frac{e^{-\theta b}(\theta b)^{n+N-i-j}}{(n+N-i-j)!} && \text{for } \begin{array}{l} N-j\le i\le M \\ m<N_v \\ i+j-N\le n<M \end{array} \\
&= \sum_{k=0}^{\min(j,m)} \left[q^v_{jk} \sum_{r=N_v}^{\infty} p^v_{r-k} \right] \frac{e^{-\theta b}(\theta b)^n}{n!} && \text{for } \begin{array}{l} i\le N-j-1 \\ m=N_v \\ n<M \end{array} \\
&= \sum_{k=0}^{\min(j,m)} \left[q^v_{jk} \sum_{r=N_v}^{\infty} p^v_{r-k} \right] \frac{e^{-\theta b}(\theta b)^{n+N-i-j}}{(n+N-i-j)!} && \text{for } \begin{array}{l} N-j\le i\le M \\ m=N_v \\ i+j-N\le n<M \end{array} \\
&= \sum_{k=0}^{\min(j,m)} q^v_{jk} p^v_{m-k} \sum_{r=M}^{\infty} \frac{e^{-\theta b}(\theta b)^r}{r!} && \text{for } \begin{array}{l} i\le N-j-1 \\ m<N_v \\ n=M \end{array} \\
&= \sum_{k=0}^{\min(j,m)} q^v_{jk} p^v_{m-k} \sum_{r=M}^{\infty} \frac{e^{-\theta b}(\theta b)^{r+N-i-j}}{(r+N-i-j)!} && \text{for } \begin{array}{l} N-j\le i\le M \\ m<N_v \\ i+j-N\le n=M \end{array} \\
&= \sum_{k=0}^{\min(j,m)} \left[q^v_{jk} \sum_{r=N_v}^{\infty} p^v_{r-k} \right] \sum_{s=M}^{\infty} \frac{e^{-\theta b}(\theta b)^s}{s!} && \text{for } \begin{array}{l} i\le N-j-1 \\ m=N_v \\ n=M \end{array} \\
&= \sum_{k=0}^{\min(j,m)} \left[q^v_{jk} \sum_{r=N_v}^{\infty} p^v_{r-k} \right] \sum_{s=M}^{\infty} \frac{e^{-\theta b}(\theta b)^{s+N-i-j}}{(s+N-i-j)!} && \text{for } \begin{array}{l} N-j\le i\le M \\ m=N_v \\ i+j-N\le n=M \end{array} \\
&= 0 \qquad \text{otherwise .} && (2.4.13)
\end{aligned}
$$

Note that $p_{ji,mn}=0$ whenever $i>M$ or $n>M$. This is due to the fact that there can be at most M data customers waiting for the gate to open. This should be compared with the exponentially distributed data packet case analyzed in Section 2.2 where no such restriction holds.

Let π_{ij} denote the steady state joint probability $\lim_{t\to\infty} P_r\{Q_t^V=i,\ Q_t^D=j\}$. Then under the condition for stability of the queueing process, we have that the π_{ij}'s are uniquely determined by the equations

$$\pi_{m,n} = \sum_{i,j} \pi_{ij} p_{ij,mn} \tag{2.4.14}$$

together with the normalization condition

$$\sum_{i,j} \pi_{ij} = 1 \tag{2.4.15}$$

If we compare this solution with the solution of the exponentially distributed packet lengths, finite data buffer problem given in Section 2.2, we see that both solutions require solving a system of linear equations. However, while in the exponential packet case, the matrix involved is sparse, in the present situation the matrix involved is virtually full. This is because the transition matrix $(p_{ij,mn})$ is almost a full matrix. On the other hand, the dimension of the linear system of equations in this case is generally lower than that in the exponential packet case.

As an illustration, consider the very simple system where $N_V=1$, $N_D=0$, $M=1$, $\lambda_1=0.01$, $\mu_1=0.01$, $\theta=40$, $b=0.01$. Equation (2.4.14) gives rise to four equations, three of which are independent. Together with the normalization condition (2.4.15), the steady state probabilities are determined as follows:

$$\pi_{00} = 0.3351097591$$

$$\pi_{01} = 0.1648652414$$

$$\pi_{10} = 0.0001016314125$$

$$\pi_{11} = 0.4999233685$$

Comparing these results to the corresponding case with exponential data packets, we see that the numerical values of the steady state probabilities are quite close, even though the models used are different.

The total number of equations in (2.4.14) is $(N_V+1)(M+1)$. Thus there are $\frac{N_V+1}{2}(2N_D+N_V)$ fewer equations in this case than the exponential packet case. For example, for $N_V=3$, $N_D=1$, $M=2$, there are 12 states associated with the constant packet case, but there are 22 states associated with the exponential packet case. However, for a reasonably-sized data buffer, the dimension of the system will still be large. Furthermore, because of the fullness of the matrix, (2.4.14) will be numerically more difficult to solve. As in the exponential packet case, these results may be used to design the size of the data buffer to achieve a certain buffer overflow probability, but to apply them in practice may first require extensive parametric studies.

2.5 Analysis of Traffic Behaviour when Data Packet Lengths are Constant with Infinite Data Buffer

In this section, we analyze the queueing problem arising in hybrid-switched networks with constant data packet lengths with infinite data buffer. Here, the queueing model used is the same as the one in the previous section except that an infinite data buffer is assumed. The performance criteria are the blocking probability for voice and the average delay for data. This problem was first studied by Kummerle [7] and then by Fischer and Harris [9], and Occhiogrosso et al. [10]. Kummerle [7] basically gave an ad-hoc approximation of the traffic behaviour. While the voice traffic was correctly analyzed in [9], the analysis for data traffic in [9] and [10] contains errors, as was pointed out in [17] and [19]. We give here the correct results governing data traffic, thereby solving this problem of long standing theoretical interest.

We first note that the voice traffic behaviour is exactly the same as that described in the previous section. In particular, the blocking probability is given by (2.4.8). We shall now concentrate on the data traffic analysis.

As before, let us use the notation π_{ij} to denote the steady state probability of having i voice customers and j data customers in the system. Define the generating function $\pi_m(z)$, $m=0,1,\ldots,N_v$ by

$$\pi_m(z) = \sum_{n=0}^{\infty} \pi_{mn} z^n$$

In Appendix 1, we derive the following equations for $\pi_m(z)$, $m=0,1,\ldots,N_v$:

$$\pi_m(z) = \sum_{j=0}^{N_v} \sum_{i=0}^{N-j-1} \sum_{k=0}^{\min(j,m)} q^v_{jk} p^v_{m-k} \, e^{-\theta b(1-z)} \pi_{ji} (1-z^{i+j-N})$$
$$+ \sum_{j=0}^{N_v} \sum_{k=0}^{\min(j,m)} q^v_{jk} p^v_{m-k} \, e^{-\theta b(1-z)} \pi_j(z) z^{j-N} \qquad m=0,1,\ldots,N_v-1 \tag{2.5.1}$$

and

$$\pi_{N_v}(z) = \sum_{j=0}^{N_v} \sum_{i=0}^{N-j-1} \sum_{k=0}^{j} \left[q_{jk}^{v} \sum_{r=N_v}^{\infty} p_{r-k}^{v} \right] \pi_{ji} e^{-\theta b(1-z)} (1-z^{i+j-N})$$

$$+ \sum_{j=0}^{N_v} \sum_{k=0}^{j} \left[q_{jk}^{v} \sum_{r=N_v}^{\infty} p_{r-k}^{v} \right] \pi_j(z) e^{-\theta b(1-z)} z^{j-N} \tag{2.5.2}$$

Equations (2.5.1) and (2.5.2) may be written in the form

$$A_c(z)\pi(z) = b_c(z) \tag{2.5.3}$$

where $\pi(z)'=[\pi_1(z),\ldots,\pi_{N_v}(z)]$, $A_c(z)$ is an $N_v \times N_v$ matrix consisting of the coefficients of the $\pi_i(z)$'s, and $b_c(z)$ is a vector containing the unknowns π_{ji}, $j=0,\ldots,N_v$, $i=0,1,\ldots,N-j-1$. The total number of unknowns are therefore $N+(N-1)+\ldots+(N-N_v+1)+(N-N_v)=\frac{N_v+1}{2}(2N-N_v)$. To determine these unknowns, we find the roots of det $A_c(z)$ inside the unit circle. In Appendix 1, it is shown that there are $\frac{N_v+1}{2}(2N-N_v)$ unique roots of det $A_c(z)$ inside $|z|\leq 1$. One of these roots is $z=1$. If the remaining roots are denoted by z_i, then we obtain $\frac{N_v+1}{2}(2N-N_v)-1$ equations by requiring det $A_c^i(z)$ vanish at the z_i's where $A_c^i(z)$ is the matrix obtained by replacing the ith column of $A_c(z)$ by $b_c(z)$. An additional equation is provided by the normalization condition $\sum_{i,j} \pi_{ij}=1$. These equations together then uniquely determine $b_c(z)$ and hence $\pi(z)$. Once the $\pi_i(z)$'s are determined, the average data delay is given by

$$E(W_D) = \frac{b}{2} + \frac{1}{\theta} \prod_{i=0}^{N_v} \pi_i'(1) \tag{2.5.4}$$

where the first term in the right hand side of (2.5.4) represents the delay due to the gate, and the second term represents the delay once access has been gained to the system.

Let us compare the solution procedure for the constant packet problem with the one for exponential packets. In the exponential packet case, the main computations involve the determination of the roots of det $A(z)$. Since $A(z)$ is tridiagonal with entries which are polynomials in z, det $A(z)$ can be evaluated using recursive relations satisfied by its principal mirrors. The roots of det $A(z)$ can then be found using standard root finding techniques. In fact, by looking at the sign changes of the principal minors of $A(z)$, it is possible to find intervals which bracket the roots of det $A(z)$ [17], [25]. On the other hand, the matrix $A_c(z)$ is full with entries being analytic (transcendental) functions of z. There are no special properties of $A_c(z)$ that can be exploited. Determining the roots of det $A_c(z)$ will be a difficult numerical problem for even moderate values of N_V, even more so that in the case of $A(z)$. Thus, while the $\pi_i(z)$'s can in principle be found using the above procedure, it is very difficult to carry out in practice except when N_V is very small, say ≤ 4. The need for suitable approximation methods is even more evident in this case.

Just as in the exponentially distributed packet lengths case, analytical solution for the average data delay is available in the simplest possible case, $N_V=1$, $N_D=0$. The average number of data customers in this case is given by

$$EQ_D = \{2[q^V_{10}p^V_0+q^V_{11}p^V_0\theta b-\theta b]\pi_{00}\theta b(1-2q^V_{11}p^V_0)$$

$$-\pi_{00}(1-p^V_0q^V_{11})\theta b(2q^V_{10}p^V_0+3q^V_{11}p^V_0\theta b-\theta b-2)\}\cdot$$

$$[2(q^V_{11}p^V_0\theta b + q^V_{10}p^V_0 - \theta b)^2]^{-1} \qquad (2.5.5)$$

where $\pi_{00} = [q^V_{10}p^V_0+q^V_{11}p^V_0\theta b-\theta b][1-p^V_0+q^V_{10}p^V_0]^{-1}$

The average data delay is then given by

$$EW_D = \frac{b}{2} + \frac{1}{\theta} E(Q_D) \tag{2.5.6}$$

The derivation of (2.5.5) is given in Appendix 2.

For λ_1=0.01, μ_1=0.01, and b=0.01, the average delay is plotted against θ in Figure 2. Notice that from the approximation using the Erlang B formula, EQ_V=0.5 so that the condition for stability of the data queue is approximately $\theta b<0.5$. This is shown clearly in Figure 2. If we compare these results to those for the exponential case, we find they are very close. In fact, for θ=40, the average delay in both cases is approximately 250 sec. In [19], it is reported that the simulation results produce an estimate of the average delay in the constant packet case for θ=40 to be 176 sec. This seems to be somewhat low compared to the analytical results. We have also independently programmed the simulation model of [19], and we have obtained an estimate of 223 sec. This casts some doubt on the validity of some of the simulation results in [19].

In the above four sections, we have looked at the exact analyses of stochastic processes arising in hybrid-switched integrated networks. We have seen that due to the numerical difficulties involved, simple approximations which reflect the qualitative behaviour of the voice and data traffic are very much needed. In the next section, we shall discuss qualitatively the behaviour of the voice and data traffic and offer some explanation of this behaviour. In Section 2.6, we shall discuss some useful approximation methods.

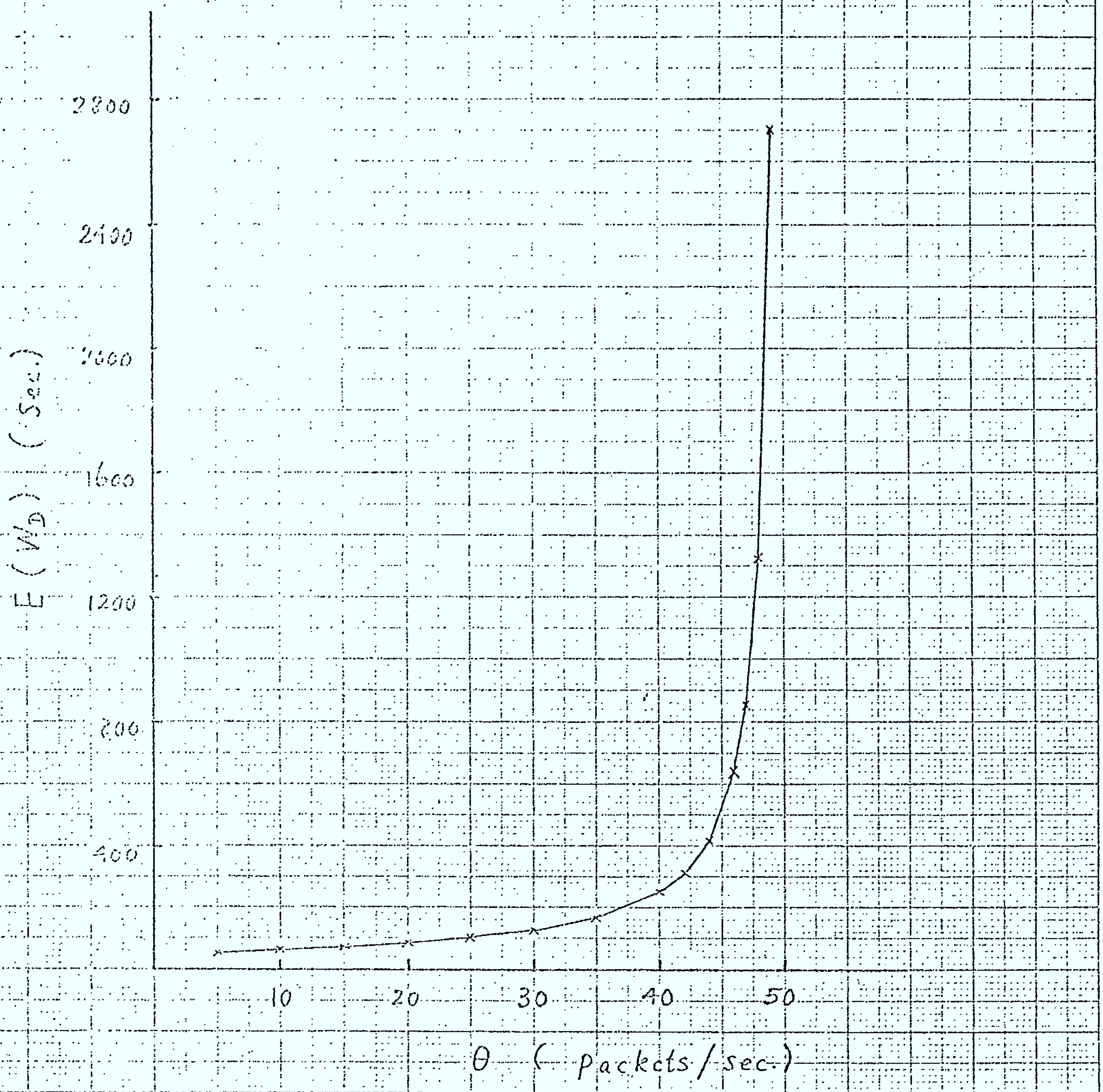


Figure 2: Average data delay as function of data load (constant data packets): $N_V = 1$, $N_D = 0$, $\lambda_1 = 0.01$, $\mu_1 = 0.01$, $b = 0.01$

2.6 Behaviour Modes of the Hybrid Switch

It is not evident from the theoretical analysis given in the previous sections how the hybrid switch behaves except in the simple case of $N_V=1$, $N_D=0$ where analytically solutions are possible. However, numerical solutions of some examples given in [17] show that $\pi_i'(1)$ increases rapidly with i, where $\pi_i(z)$ is the generating function for data with i voice customers, as defined in Section 2.3. Intuitively, if we have more voice customers in the channel, there are then fewer slots available for data transmission, and hence the data delay will increase. This is in contrast to the results in [9] and [10], where the conclusion is essentially that if $E(Q_V)$ is the average number of busy voice slots, then the performance of the data will be that of a single server queue of capacity $N-E(Q_V)$. In particular, the rapid increase of $\pi_i'(1)$ does not occur in [9], [10]. However, extensive simulations supporting our above comment have been given in [19]. These simulation results show how the data queue builds up during periods when there are a large number of voice calls, and how the data queue dissipates when the voice calls drop. We now give some additional theoretical basis for the explanation of the data traffic behaviour, using the work of Yechiali and Naor [27].

In [27], a single server queueing system is considered in which the capacity of the server alternates between two levels. The period that the server remains at a given level has an exponential distribution and customers are assumed to have exponential service times. Thus their model corresponds to a system in which data traffic is serviced at one of two different rates.

Let $\frac{1}{\lambda_1}$ be the mean holding time in state 1 and let $\frac{1}{\mu_1}$ be the mean holding time in state 2. Then the steady state probability of being in state 1 is $\pi_1=\frac{\mu_1}{\lambda_1+\mu_1}$ and the probability of being in state 2 is $\pi_2=\frac{\lambda_1}{\lambda_1+\mu_1}$.

Assume that data messages have arrival rate λ_2 in both states and average service time $\frac{1}{\mu_2(1)}$ when the server is in state 1 and $\frac{1}{\mu_2(2)}$ when the server is in state 2, (see Figure 3). The average capacity of the system is $\overline{\mu}_2=\pi_1\mu_2(1)+\pi_2\mu_2(2)$ messages per second and λ_2 must be less than $\overline{\mu}_2$ in order for the queue to be stable.

Yechiali and Naor [27] found expressions for the mean number of customers in the system in terms of the roots of a third order polynomial. By looking at several extreme cases they were able to characterize several modes of system behaviour.

Case A: If either λ_1 or μ_1 vanishes, then the system reduces to an M/M/1 queue with capacity $\mu_2(1)$ or $\mu_2(2)$ when λ_1 or μ_1 vanishes respectively.

Case B: If $\mu_1\to\infty$ while λ_1 remains finite then the system reduces to an M/M/1 queue with capacity $\mu_2(1)$.

Case C: Suppose that very rapid oscillations occur between levels 1 and 2. In particular suppose that λ_1 and μ_1 tend to infinity with $\lambda_1=C\mu_1$ where C is a positive finite constant. The system then reduces to an M/M/1 queue with capacity $\overline{\mu}_2$. That is, when the capacity varies very rapidly, then the average data message "sees" a single server with capacity equal to the average capacity of the system, $\overline{\mu}_2$.

Case D: Now suppose that the transitions between the two levels of service are very sluggish: λ_1 and μ_1 tend to zero with $\lambda_1=C\mu_1$ where C is a finite constant. There are two subcases.

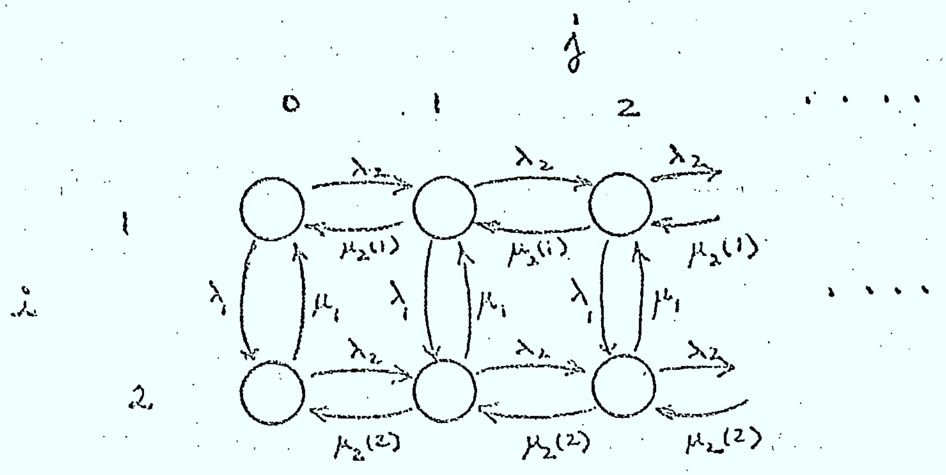


FIGURE 3. Queueing System with Two Service Levels

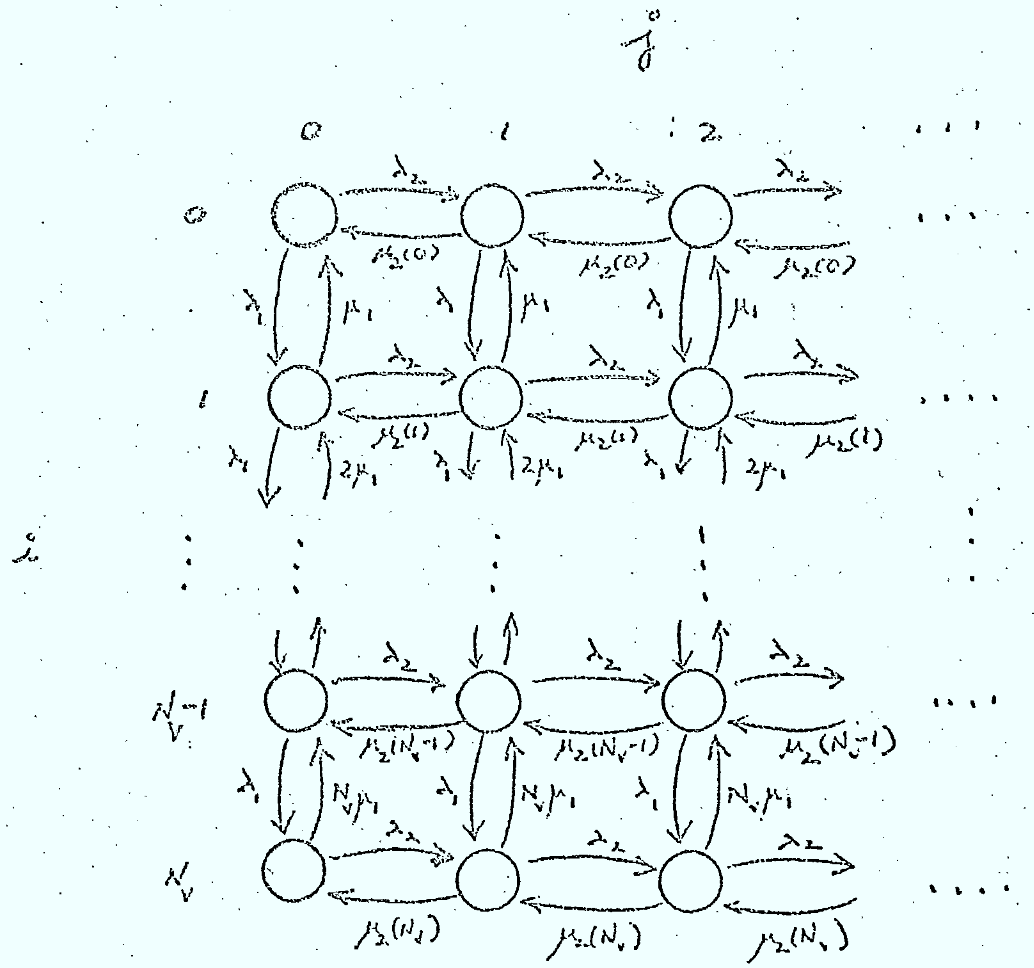


FIGURE 4. Queueing System with N_v+1 Service Levels

Case D1: In this case we have that the data arrival rate λ_2 is always less than the capacity at both levels of service, that is, $\lambda_2<\mu_2(1)$ and $\lambda_2<\mu_2(2)$. In this case we find that the system exhibits two quasi-stationary modes corresponding to an M/M/1 queue with capacity $\mu_2(1)$ when the system is in state 1 and to an M/M/1 queue with capacity $\mu_2(2)$ when the system is in state 2.

Case D2: In this case we have that $\lambda_2<\mu_2(1)$ but $\lambda_2>\mu_2(2)$. (Of course we require that $\lambda_2<\overline{\mu}_2$ in order for the overall system to be stable.) The system exhibits a quasi-stationary mode when the system is in state 1, and a nonstationary mode when it is in state 2. The number of customers in the system increases steadily while the system is in state 2 and is "flushed out" when the system returns to state 1. Yechiali and Naor showed that the expected number of customers in the system can be made arbitrarily large by making appropriate choices of λ_1 and μ_1.

The Yechiali-Naor model can be extended to the case where the level of service can take on N_v possible values. The associated transition diagram is shown in Figure 4. The generating function approach again yields expressions for the mean number of customers in the system in terms of the roots of a polynomial. The same modes of system behaviour can be expected. In particular, cases C and D explain the discrepancy between the results found in [9] and those found in [19] and Sections 2.3 and 2.5. Suppose that the level of service changes many times during the service time of a single data message. Then the capacity available to a data message is approximately equal to the average capacity $\overline{\mu}_2$ and is thus effectively independent of the process regulating the instantaneous capacity. The studies in [9] and [10] assumed that the number of data messages in the system and the instantaneous capacity are independent. This assumption is valid only if the process $Q_v(t)$, the number of voice calls, varies rapidly during the service time of a single message. The rate of change in the level of service when the number of voice customers is $Q_v(t)$ is $\lambda_1 + Q_v(t)\mu_1$. In general the voice parameters λ_1 and μ_1 are much smaller than the corresponding parameters for the data traffic. Thus the above assumption is valid only if $Q_v(t)$ is a very large number. Thus the results in [9] and [10] will hold only for very large scale systems in which a very large number of voice calls are handled simultaneously.

The hybrid switching systems investigated in [9] and simulated in [19] involved a small number of simultaneous voice calls. Consequently the systems exhibit type D behaviour. When the offered data load is less than the instantaneous capacity the system settles into a quasi-stationary mode. When the offered load exceeds the instantaneous capacity, the system is temporarily unstable and the data queues build up until the system returns to a quasi-stationary mode.

Through the above discussion, some insight has been gained into the qualitative behaviour of the hybrid switch. In the next section, we examine some approximate methods of analysis which give the same qualitative performance characteristics as the exact analysis discussed in the previous sections.

2.7 Approximate Methods of Analysis for Hybrid-Switched Integrated Networks

Some approximate methods of analysis for hybrid-switched integrated networks have been previously presented in [7] and [17]. In [7], approximate formulas are given for the average data delay which distinguish the cases where voice traffic is heavy and where it is light. It is shown that large data buffer queues can build up during overload periods. The approximations given in [7] are rather crude and refinements have been studied in [17]. In [17], a conditional mean approximation and a diffusion approximation for the exponentially distributed data packet case are analyzed. Both of these approximations require solutions by iteration since they involve solutions of nonlinear systems of equations. Some numerical results from these approximations are then compared to the exact solution. They indeed show the same qualitative behaviour as the exact solution. Hence, they can be used to evaluate the system performance. We shall not go into the details of these methods. The interested reader is referred to [17] for a full discussion.

Here, we present yet another approximate method for estimating the performance of the hybrid switch. We believe that our method is simpler to use than those in [17] since it does not involve the iterative solution of a nonlinear system of equations. We also present some numerical results comparing our approximation to the simulation results presented in [19]. We have not, however, attempted to compare our approximation to those reported in [17].

We have already seen in the previous section that the voice and data random processes are not independent. We assume that the voice process changes slowly relative to the data process so that the behaviour in case D of the previous section prevails. For the case where the offered data load never exceeds the instantaneous capacity we use a weighted average of the submode performance to estimate the average performance. For the case where the offered data load occasionally exceeds the instantaneous capacity we develop a fluid approximation model that gives simple formulas for the average performance.

Assume that messages have Poisson arrivals with rate λ_2 messages/second and average message length $\frac{1}{\mu_2}$ data units/message. Let $\theta=\lambda_2/\mu_2$ be the resulting average number of data unit arrivals per second. The average number of arrivals per frame is then θb, where b is the frame duration. For convenience we will assume that a data unit corresponds to a single slot. The average data load is then given by

$$\rho_D = \frac{\theta b}{N-E(Q_V)}$$

where $N-E(Q_V)$ is the average number of slots per frame available to data traffic. The instantaneous data load is then

$$\rho_D(t) = \frac{\theta b}{N-Q_V(t)} \quad .$$

If the instantaneous load is always less than one, and if the voice process varies much more slowly than the data process, then case D1 holds and the system can be viewed as consisting of several quasi-stationary modes. In this case we approximate the average number of messages in the system by

$$E[Q_D] = \sum_{i=0}^{N_V} E[Q_D \mid Q_V=i]\pi_i$$

where

$$\pi_i = \frac{\rho_1^i/i!}{\left(\sum_{j=0}^{N_V} \rho_1^j/j!\right)} \quad ,$$

is the steady state probability of the voice process being in state i, and $E[Q_D|Q_V=i]$ is the average number of messages in a queueing system with load $\theta b/(N-i)$. For example if the messages are assumed to be exponentially distributed we obtain:

$$E[Q_D] = \sum_{i=0}^{N_V} \frac{\theta b/(N-i)}{1-\theta b/N-i} \pi_i$$

$$= \sum_{i=0}^{N_V} \frac{\theta b}{N-i-\theta b} \pi_i \quad .$$

The function $\theta b/(N-\theta b-i)$ is a convex U function of i; so applying Jensen's inequality we obtain

$$E[Q_D] = E\left[\frac{\theta b}{N-\theta b-Q_V}\right]$$

$$\geq \frac{\theta b}{N-\theta b-E(Q_V)}$$

$$= \frac{\theta b/(N-E(Q_V))}{1-\theta b/(N-E(Q_V))},$$

with equality if and only if $\pi_i=1$ for some i. That is, the performance of this system is always worse than that of a system with capacity $N-E(Q_V)$.

Now consider the case in which the instantaneous load occasionally exceeds one. In this case some of the modes are unstable: data buffers build up until the instantaneous load becomes less than one. The evolution of the system can be viewed as consisting of alternating periods of quasi-stationary behaviour during which queues are stable and of overload periods during which queues build up. A net amount of work arrives during an overload period. The interarrival times of these "work arrivals" correspond to the time between overloads. Thus a correspondence can be established between the buffer build up during overload periods and the unfinished work in a queueing system. This correspondence was developed by Hsu in [28]. He considered a queueing system in which the server is available for some random time β and then not available for some random time α. He assumes that data arrive in a continuous flow of d units/second. During transmit periods, the server processes data at a rate of b' units/second. The net departure rate from the system is therefore $c=b'-d$ units/second during a transmit period. The net arrival rate into the system during a non-transmit period is d units/second.

Let $X=d\alpha$ and $Y=c\beta$. Then X represents the amount of work accumulated during an overload period, and Y represents the amount of work that can be processed in between overload periods. Consider $Q_D(t)$ the number of messages in the system of time t. The onset of an overload period corresponds to a valley in the graph of $Q_D(t)$ versus time, and the ending of an overload period corresponds to a peak in the graph. Hsu shows in his paper that the mean peak in the graph of $Q_D(t)$ corresponds to the average total response time (waiting time+service time) in a GI/G/I queueing system in which arrivals have the same distribution as Y and the service times have the same distribution as X. He also showed that the valleys in $Q_D'(t)$ correspond to the average waiting time in the same GI/G/I system. Thus the results for these queueing system can be used to estimate the buffer contents in queueing systems with interrupted service periods. In order to apply Hsu's model it is necessary to neglect the randomness of the message arrivals and service times and instead assume constant flows of data into and out of the system. In effect Hsu's approach leads to a fluid approximation.

Now consider the hybrid switch. The capacity available to data at time t is $(N-Q_v(t))b$ units/frame. The average capacity available to data during a period of overload is

$$(N-\overline{Q}_v)b \text{ units/frame}$$

where

$$\overline{Q}_v = E[Q_v|\text{overload}]$$

$$= \frac{\sum_{i=\lceil\theta b\rceil}^{N_v} iP_i}{\sum_{i'=\lceil\theta b\rceil}^{N_v} P_{i'}} .$$

where $\lceil x\rceil$ denotes the smallest integer $\geq x$.

The average capacity during a nonoverload period is

$$(N-\underline{Q}_v)b \text{ units/frame}$$

where

$$\underline{Q}_v = E[Q_v | \text{no overload}]$$

$$= \frac{\sum_{i=0}^{\lceil \theta b \rceil - 1} i P_i}{\sum_{i'=0}^{\lceil \theta b \rceil - 1} P_{i'}} .$$

The net flow into the system during an overload period is therefore

$$d = \theta b - (N - \overline{Q}_v) \text{ units/frame}$$

and the net possible flow out of the system during stable periods is

$$c = (N - \underline{Q}_v) - \theta b .$$

In Appendix 3 we present a method for computing the average and variance of the overload and stable periods, T, $\sigma^2(T)$, S, $\sigma^2(S)$, respectively. The M/M/1 and the M/G/1 results then give:

M/M/1 $\quad E(\text{buffer peaks}) = \frac{1}{\mu - \lambda}$

$$E(\text{buffer valleys}) = \frac{\rho}{\mu - \lambda}$$

M/G/1 $\quad E(\text{buffer peaks}) = \frac{1}{\mu} + \frac{1}{\mu}\left[\left(\frac{\rho}{1-\rho}\right)\left(\frac{1+c_b^2}{2}\right)\right]$

$$E(\text{buffer valleys}) = \frac{1}{\mu}\left[\left(\frac{\rho}{1-\rho}\right)\left(\frac{1+c_b^2}{2}\right)\right]$$

where

$$\frac{1}{\lambda} = \frac{cS}{b} \quad , \qquad \frac{1}{\mu} = \frac{dT}{b}$$

and

$$c_b^2 = \frac{\sigma_x^2}{m_x^2} = \mu^2 \sigma^2(T) \frac{d^2}{b^2}$$

Figures 5 and 6 show the average data delay estimates using the M/M/1 and the M/G/1 models respectively, for the case where N_V=10, N_D=5, λ=0.05, and μ=0.01. The upper curve in each figure corresponds to the average buffer peaks and the lower curves to the average buffer valleys. Note that since we are using a fluid approximation, the curves should underestimate the actual performance. This is generally true in the curves except at very high load ($\theta b \simeq 9.75$). The simulation method used in [19] however tends to underestimate the average delay so it does not follow that the fluid approximation is inaccurate for these values.

Note that the scale in Figures 5 and 6 is in <u>seconds</u>. Thus the performance of the hybrid switch is orders of magnitude greater than what was predicted in [9]. In Chapter 3 we present a flow control procedure that results in a significant performance improvement.

2.8 Integrated Switching with Variable Frame

In this section, we briefly discuss an integrated switching scheme which is similar to the hybrid switching scheme studied so far. In hybrid switching, the time frames are taken to be fixed. This is required for the synchronous mode of transmission for circuit-switched traffic. However,

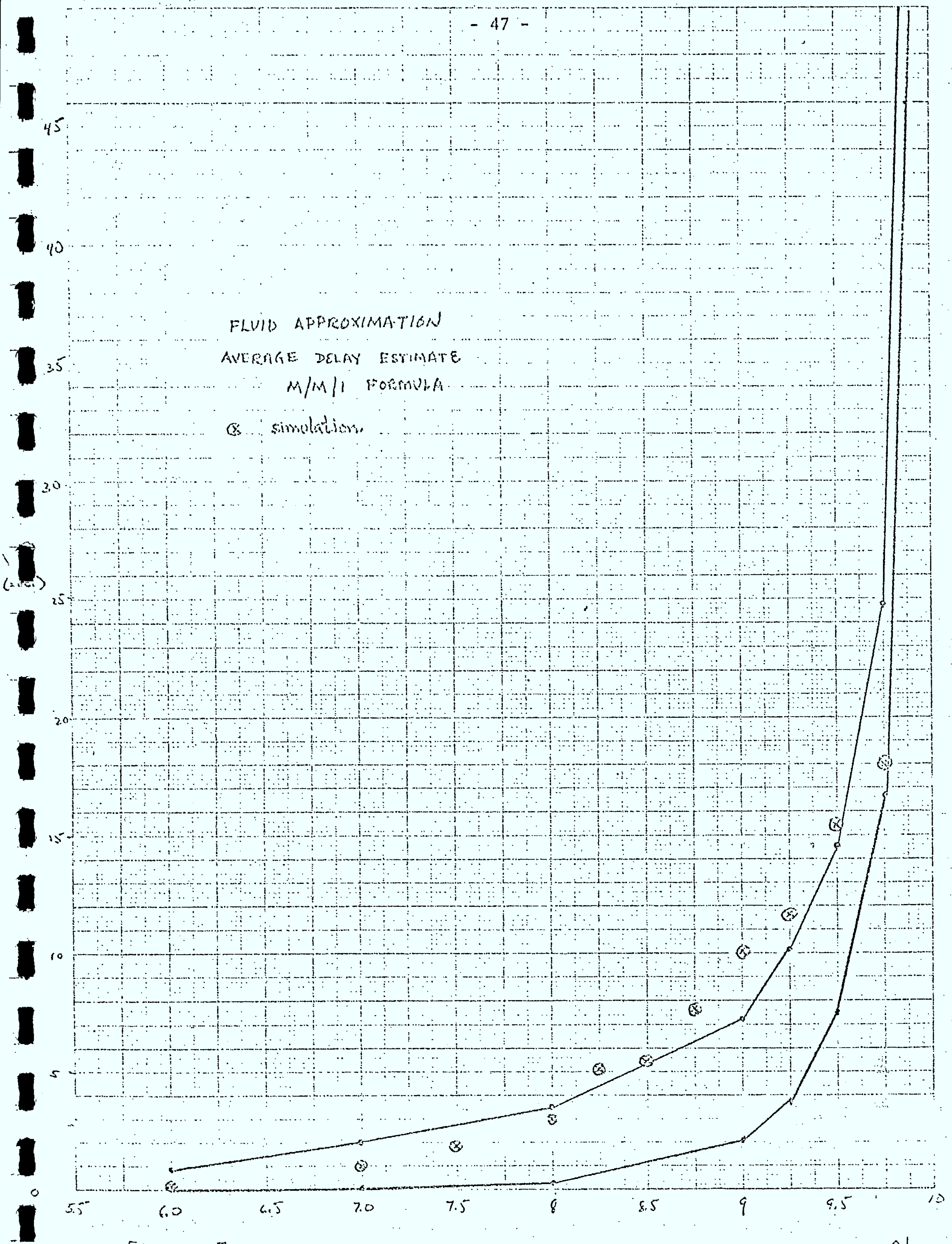

FLUID APPROXIMATION
AVERAGE DELAY ESTIMATE
M/M/1 FORMULA
⊗ simulation
45
40
35
30
25
20
15
10
5
0
5.5
6.0
6.5
7.0
7.5
8
8.5
9
9.5
10
θb

Figure 5.

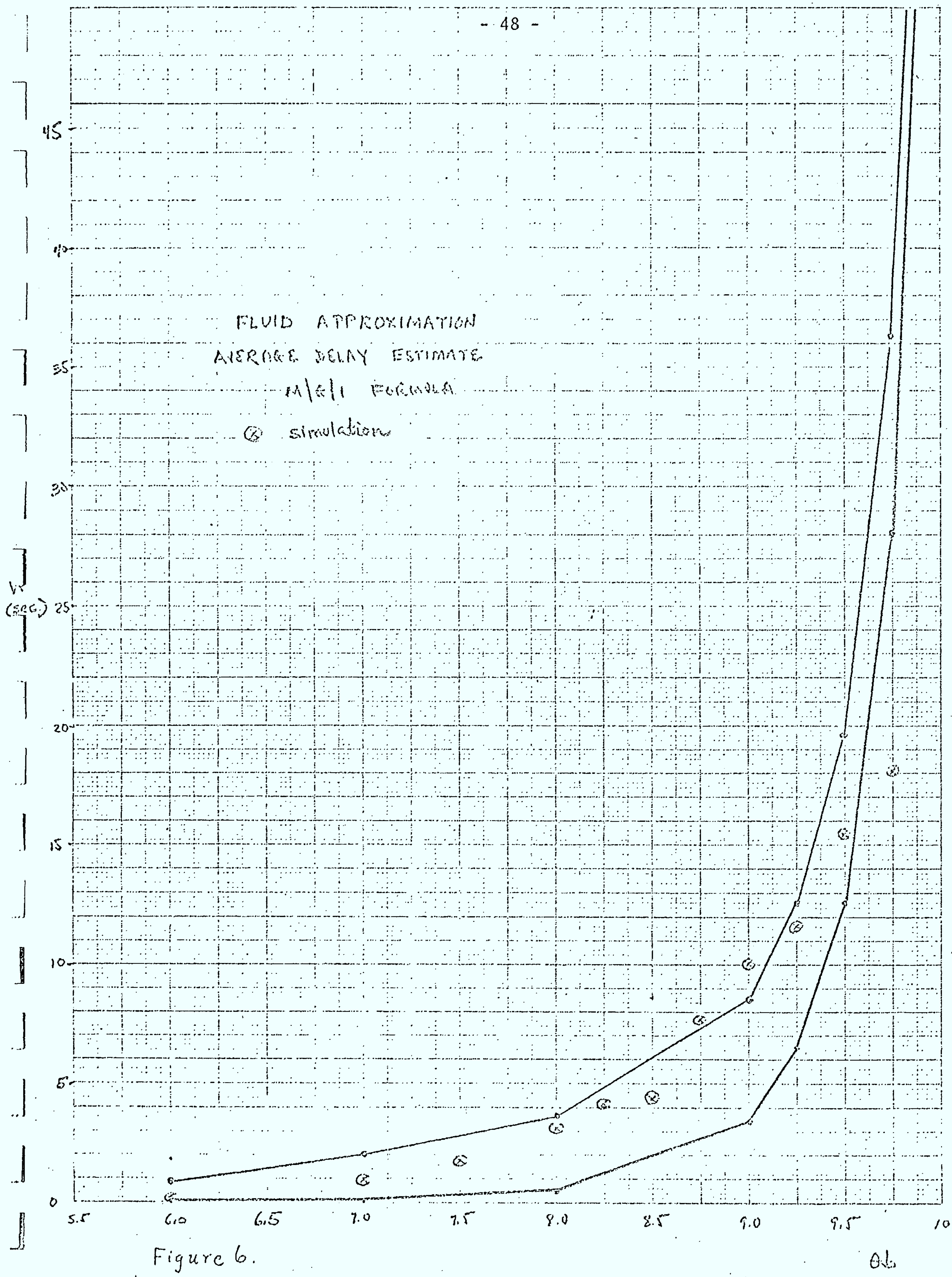

FLUID APPROXIMATION
AVERAGE DELAY ESTIMATE
M|G|1 FORMULA
⊗ simulation
W (SEC.)
45
40
35
30
25
20
15
10
5
0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
θb

Figure 6.

this may result in portions of the frame being unused. One way of eliminating the unused slots is to allow the frame size to vary in accordance with the variations of the traffic. Synchronization of circuit-switched customers now poses a problem resulting in nontransparent transmission. It is, however possible to bound the framesize to a certain maximum value admissible for transparency of circuit switching and with the help of some buffering, realize integrated circuit- and packet-switched systems with variable frame length. Such systems provide a saving in bandwidth, but at the cost of additional hardware and software complexity.

In the scheme examined here we assume that the integrated frame is divided in a circuit-switched and a packet-switched portion. The packet traffic has intermittent access to the system along with the circuit-switched sessions. The frame size may vary from a very small delimiter value, in case of no traffic present, to a maximum value corresponding to highly congested traffic conditions. The maximum frame length must not violate the line switching transparency condition.

Since circuit switching is preferred for lengthy sessions, we assume that circuit-switched customers are permanently connected to the multiplexer and may be either in the transmission mode or in the idle mode. Packets are stored in a buffer and served in a first-in-first-out mode (FIFO). The frame accommodates all active circuit-switched sessions and up to a maximum number of packets waiting in queue.

One possible method of analysis for this system assumes a round robin service algorithm [29] for the circuit-switched traffic. The average delay for circuit-switched traffic can then be evaluated. The packet-switched data traffic is more difficult to analyze. If we assume the number of occupied voice slots to be independent from frame to frame, the problem simplifies, and we find the average delay for data. However, we have seen in Section 2.6 that the independence assumption is only valid under certain conditions. Thus, further investigations are needed before the performance of this system can be adequately analyzed. We have not pursued the details.

CHAPTER 3

A DATA FLOW CONTROL PROCEDURE FOR THE HYBRID SWITCH

3.1 The Need for Flow Control

In Chapter 2, we have analyzed queueing problems in connection with the hybrid-switched integrated network. We have seen that if we attempt to achieve high channel utilization by using a small number of data channels N_D and allowing data traffic to exploit the excess capacity allotted to voice traffic, large data queues will build up, leading to either buffer overflow or excessive delays. It is clear then that some flow control mechanism is needed to regulate the data flow into the switch during high voice channel occupancy periods and to prevent congestion that leads to excessive delays.

The investigation of flow control procedures has been undertaken in [17] and [19]. In [17], flow control is approached from the point of view of buffer management. It considers how incoming data packets should be assigned buffer space in a finite buffer. Various assignment procedures are analyzed which take into account the different sizes and priorities of data packets. Since packets are blocked when buffer spaces are fully occupied, in effect the average delay encountered by customers admitted into the system is reduced. Messages not admitted into the system are presumably retransmitted at a later time. Such data packets will suffer extra delay. In [19], simulation results using various data flow control and voice rate control schemes for the hybrid switch are obtained. The best results obtained involve the scheme which combines fixing a limit on the data buffer and using data-queue-dependent voice rate control.

However, no mathematical model is constructed for a systematic study of flow control procedures.

In this chapter we present a flow control procedure which consists of regulating the data flow into the switch so that it matches the instantaneous capacity of the switch. In particular the flow control scheme operates so that the instantaneous load (ratio of offered data load to available capacity) is always constant. Thus the data flow into the system is increased when a voice call ends and is decreased when a voice call is set up. It turns out that closed form solutions can be found for the performance of this system and that the performance is the same as that of a single server queue of capacity equal to the average capacity available to data, i.e., $N\text{-}E(Q_v)$.

3.2 Description and Analysis of a Data Flow Control Procedure

In order to motivate the flow control procedure consider first the single channel case. Voice calls arrive with exponentially distributed interarrival times of mean $\frac{1}{\lambda_1}$ and exponentially distributed service times $\frac{1}{\mu_1}$. A voice call arriving when there is another voice call already in the system is blocked and cleared. Data is transmitted whenever there are no voice calls in the system and data transmissions are preempted upon arrival of a voice call. We assume that the data traffic has Poisson arrivals of rate λ_2 messages per second and that the messages have exponentially distributed lengths of $\frac{1}{\mu_2}$ transmission seconds/message. The transition diagram for the system state (i,j) where i=number of voice calls, and j=number of data customers is given below.

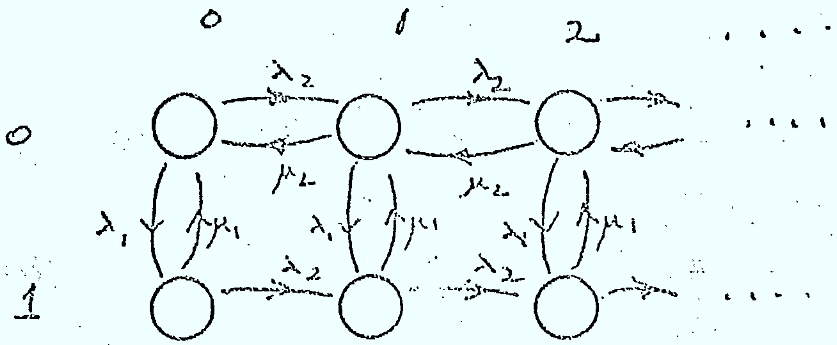


Figure 7.

Clearly when i=1 the data buffer steadily builds up. Weinstein et al. [19] has found the average number of customers in the system to be

$$E[Q_D] = \frac{\rho_2(1+\rho_1)^2+\rho_1\ \lambda_2/\mu_1}{(1+\rho_1)(1-\rho_2-\rho_1\rho_2)}$$

where

$$\rho_1 = \frac{\lambda_1}{\mu_1} \quad \text{and} \quad \rho_2 = \frac{\lambda_2}{\mu_2} \ .$$

Note the effect of the term $\frac{\lambda_2}{\mu_1}$. This term represents the average data buffer buildup during the service of a single voice customer. Clearly as the ratio of λ_2 to μ_1 is increased the average data contents can be made arbitrarily large. Now consider a flow control procedure in which the data arrival rate λ_2' is proportional to the instantaneous capacity. In this case the instantaneous capacity is zero when a voice call is present, so the flow control reduces the data flow into the switch to zero. The leads to the following transition diagram:

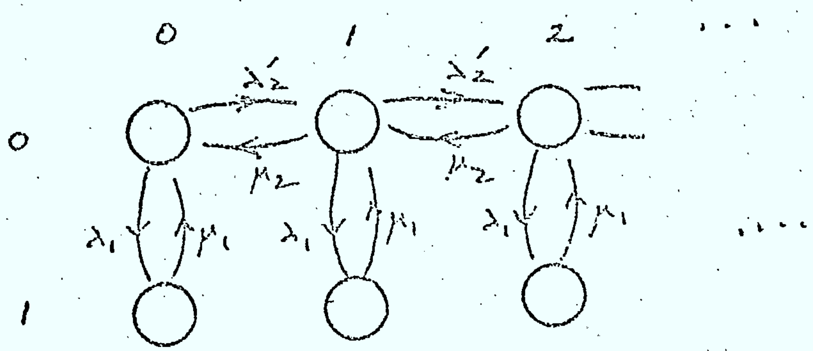


Figure 8.

The steady state probabilities for this two dimensional Markov chain is given by the following expressions:

$$P_{0j} = \frac{1}{1+\rho_1}(1-\rho_2')\rho_2'^{j} \qquad j=0,1,\ldots$$

$$P_{1j} = \frac{\rho_1}{1+\rho_1}(1-\rho_2')\rho_2'^{j} \qquad j=0,1,\ldots$$

where $\rho_2'=\lambda_2'/\mu_2$. The mean number of customers in the system and the average arrival rate are then found to be

$$E_{fc}[Q_D] = \frac{\rho_2'}{1-\rho_2'}$$

$$\overline{\lambda}_D = \lambda_2' P[Q_V=0] = \frac{\lambda_2'}{1+\rho_1} \quad .$$

In order to compare the performance of the single channel hybrid switch with and without flow control it is necessary to compare them at equal average arrival rates, that is, at $\lambda_2=\overline{\lambda}_D$. Letting $\rho_D=\overline{\lambda}_D/\mu_2$ and rearranging the expressions for $E[Q_D]$ and $E_{fc}[Q_D]$ we obtain:

$$E_{fc}[Q_D] = \frac{(\overline{\lambda}_D/\mu_2)(1+\rho_1)}{1-(\overline{\lambda}_D/\mu_2)(1+\rho_1)}$$

$$= \frac{\rho_D(1+\rho_1)}{1-\rho_D-\rho_1\rho_D}$$

and

$$E[Q_D] = \frac{\rho_D(1+\rho_1)}{1-\rho_D-\rho_1\rho_D} + \frac{\rho_1\lambda_D/\mu_1}{(1+\rho_1)(1-\rho_D-\rho_1\rho_D)} \quad .$$

Two observations are in order: First, the performance of the system with flow control is the same as that of an M/M/1 system with utilization $\frac{\bar{\lambda}_D}{\mu_2}(1+\rho_1)$. Secondly, the performance of the system without flow control is given by that of the system with flow control plus the second term in the equation. These additional buffer contents are clearly due to the buffer buildup during voice calls.

In the multichannel case, the state space is given by the set $\{(i,j):0\le i\le N_v,\ 0\le j<\infty\}$. The departure rate of data traffic during state (i,j) is $\mu_2(i,j)=(N-i)\mu_2$, if we assume that the hybrid switch operates in a single server fashion. The flow control procedure would then regulate the input data flow to be $\lambda_2(i,j)=(1-i/N)\lambda_2$. The resulting instantaneous load is then given by

$$\frac{\lambda_2(i,j)}{\mu_2(i,j)} = \frac{\lambda_2(1-\frac{i}{N})}{N(1-\frac{i}{N})\mu_2} = \frac{\lambda_2}{N\mu_2}$$

which is a constant independent of the state (i,j). This flow control scheme has the effect of making the voice and data processes independent. In Appendix 4, we show that the steady state probabilities of the system are given by

$$P_{ij} = P_i P_j \qquad 0\le i\le N_v, \quad 0\le j<\infty$$

where

$$P_i = \frac{\left(\frac{\lambda_1}{\mu_1}\right)^i}{i!}\left[\sum_{i'=0}^{N_v}\left(\frac{\lambda_1}{\mu_1}\right)^{i'}\right]^{-1}$$

and

$$P_j = \left(1 - \frac{\lambda_2}{N\mu_2}\right)\left(\frac{\lambda_2}{N\mu_2}\right)^j .$$

The distribution of voice customers in the system is given by the Erlang B formula for the $M/M/N_V/N_V$ system, and the distribution of messages in the system is given by the geometric distribution corresponding to the M/M/1 system. The call blocking probability and the average number of voice customers in the system is therefore given by

$$P_B = \frac{1}{N_V!}\left(\frac{\lambda_1}{\mu_1}\right)^{N_V}\left[\sum_{i'=0}^{N_V}\left(\frac{\lambda_1}{\mu_1}\right)^{i'} / i'!\right]^{-1}$$

and

$$E[Q_V] = \left(\frac{\lambda_1}{\mu_1}\right)(1-P_B) .$$

The average data arrival rate is given by

$$\begin{aligned}\overline{\lambda}_D &= E[\lambda_2(Q_V, Q_D)] \\ &= E\left[\lambda_2\left(1-\frac{Q_V}{N}\right)\right] \\ &= \lambda_2\left(1-\frac{1}{N}E(Q_V)\right)\end{aligned}$$

The average number of data customers in the system is therefore given by

$$E[Q_D] = \frac{\dfrac{\lambda_2}{N\mu_2}}{1-\dfrac{\lambda_2}{N\mu_2}} = \frac{\dfrac{\overline{\lambda}_D}{N\mu_2}}{\left(1-\frac{1}{N}E(Q_V)\right) - \dfrac{\overline{\lambda}_D}{N\mu_2}}$$

Note that the expected number of data customers in the system is the same as that of an M/M/1 queue with capacity $(1-E(Q_V)/N)$.

As a numerical example consider the hybrid switch with the following parameters: $N=15$, $N_V=10$, $\lambda_1=1/20$, $1/\mu_1=100$, $\lambda_1/\mu_1=5$ and $\overline{\lambda}_D=9$ packets/frame. The blocking probability is then $P_B=.0184$ and the average number of voice customers in the system is $E(Q_V)=4.91$. The average capacity available to data traffic is $15-E(Q_V)=10.09$ packets/frame. The simulation results for the system without flow control gives $E(Q_D)=9000$ packets. The performance of the system using flow control gives $E(Q_D)=8.25$ packets. The performance has improved by a factor of 1000.

CHAPTER 4

ANALYSIS OF VOICE TRAFFIC HAVING VARIABLE DIGITIZATION RATES

4.1 Introduction

The circuit-switched portion of the hybrid switch can handle a variety of voice-digitization rates. By requiring that all allowable rates be a multiple of the basic-size slot it is possible to formulate the problem in terms of a multi-channel (multiple server) system. In the next section we will present the solution for the case where there are two types of calls, one requiring one slot per frame, and the other requiring two slots per frame. For example, this problem could arise in a T1 transmission system in which calls of rates 64 kb/s and 32 kb/s are to be handled. The approach is readily extended to cases with more traffic types. Then in Section 4.3, we discuss the implications of these results to flow control.

4.2 Blocking Probabilities for Voice Traffic with Variable Digitization Rates

In the following we focus our attention on the circuit-switched portion of the frame. As in previous sections, we assume that this subframe consists of N_V basic-sized slots. Since the duration of voice call is much greater than the duration of a frame, we can view this system as consisting of N_V channels (servers) as in the classic telephone trunking problem. The difference from the classic problem is that we now have two classes of traffic, each with a different digitization rate. This is equivalent to the situation where one class seizes a single channel and holds it for a random length of time, and the other which seizes two channels and releases them simultaneously after some random holding time. The problem is to determine the blocking probabilities for the two classes of traffic. For simplicity, we shall refer to this problem as the multicapacity problem. This problem is not new and in fact appears as a home work problem in Kleinrock [26].

The problem is solved using a two-dimensional Markov chain. Let $N_V=2m$, and assume that class 1 traffic seizes two channels, has Poisson arrivals of rate λ_1 and mean holding time $1/\mu_1$, and assume that class two traffic seizes a single channel, has Poisson arrivals of rate λ_2 and has mean holding time $1/\mu_2$. In Appendix 5, we show that the class 1, class 2, and total blocking probabilities are given respectively by

$$P_{B1} = \sum_{j=0}^{m-1} \{P_{j,2(m-j)-1} + P_{j,2(m-j)}\} + P_{m,0}$$

$$P_{B2} = \sum_{j=0}^{m} \{P_{j,2(m-j)}\}$$

$$P_B = \frac{\lambda_1}{\lambda_1+\lambda_2} P_{B1} + \frac{\lambda_2}{\lambda_1+\lambda_2} P_{B2}$$

where

$$\rho_1 = \lambda_1/\mu_1 , \quad \rho_2 = \lambda_2/\mu_2 \text{ and}$$

$$P_{j,k} \triangleq \Pr[j \text{ type 1 customer and } k \text{ type 2 customers in system}]$$

$$= \frac{\frac{1}{j!}\rho_1^j \frac{1}{k!}\rho_2^k}{\left[\sum_{j'=0}^{m} \sum_{k'=0}^{2(m-j)} \frac{1}{j'!}\rho_1^{j'} \frac{1}{k'!}\rho_2^{k'}\right]}$$

It can be shown that the average number of type 1 and type 2 customers are given respectively by

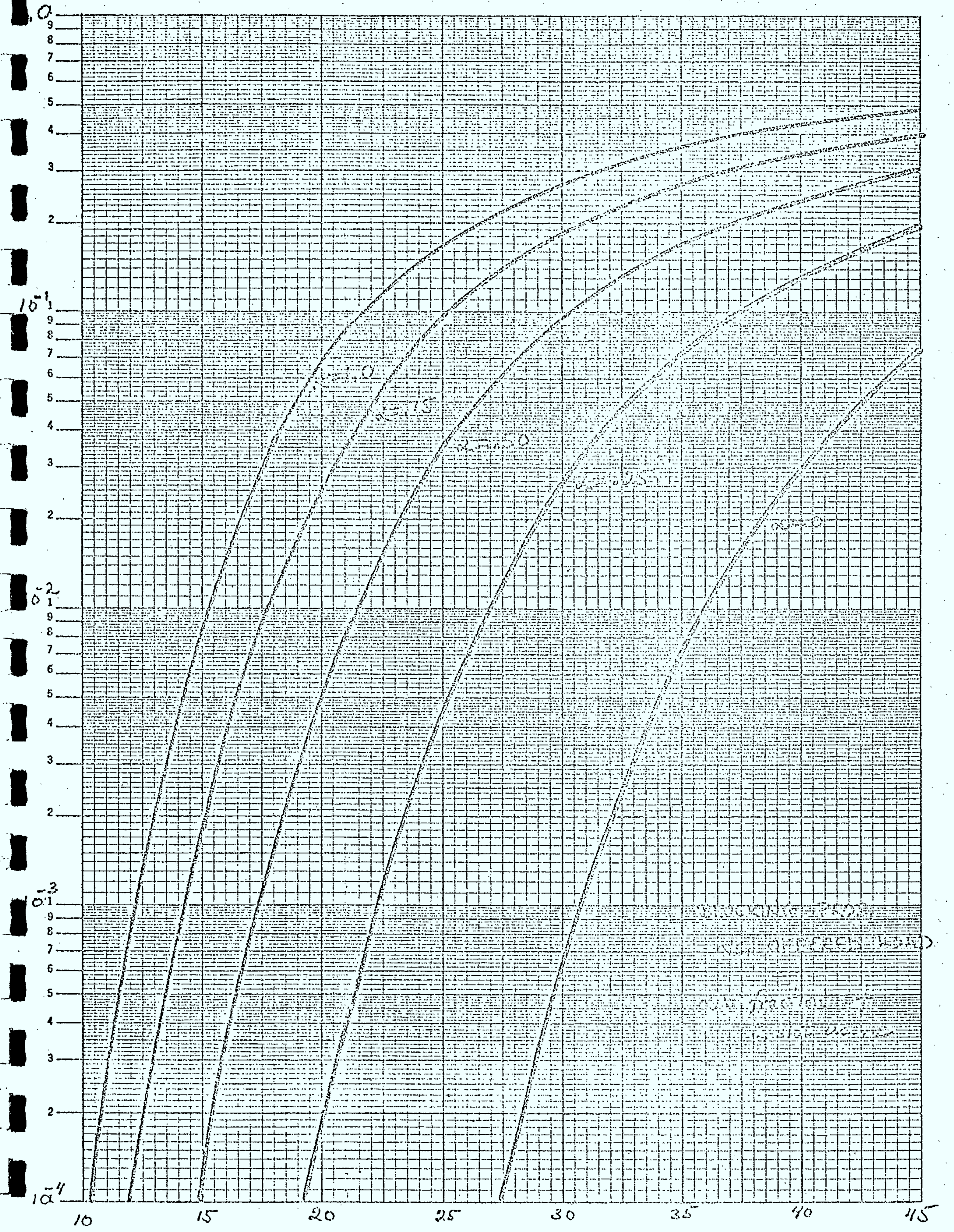
$$E[Q_1] = \rho_1[1-P_{B1}]$$

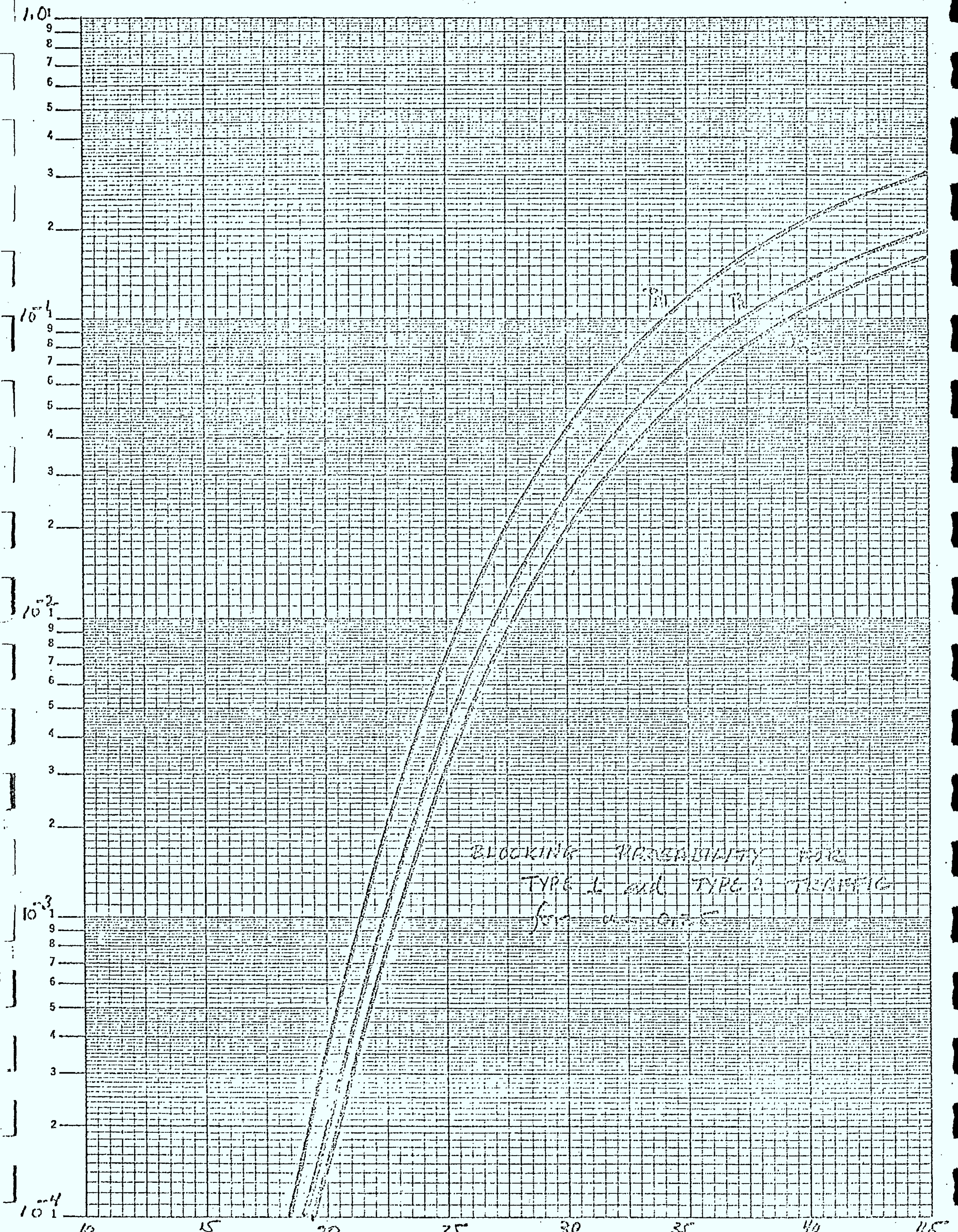
$$E[Q_2] = \rho_2[1-P_{B2}] \quad .$$

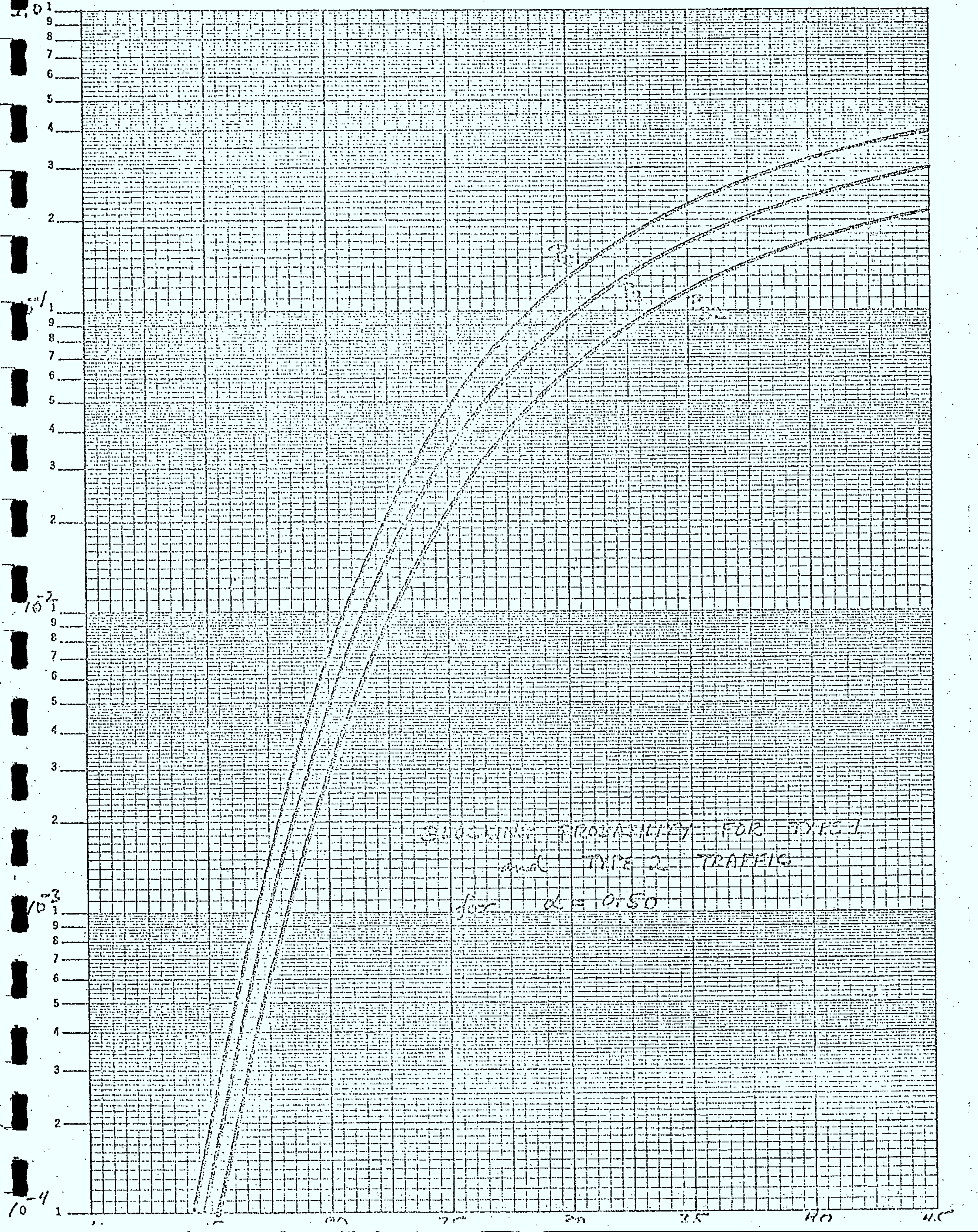
It then follows that the average number of busy channels in the system is

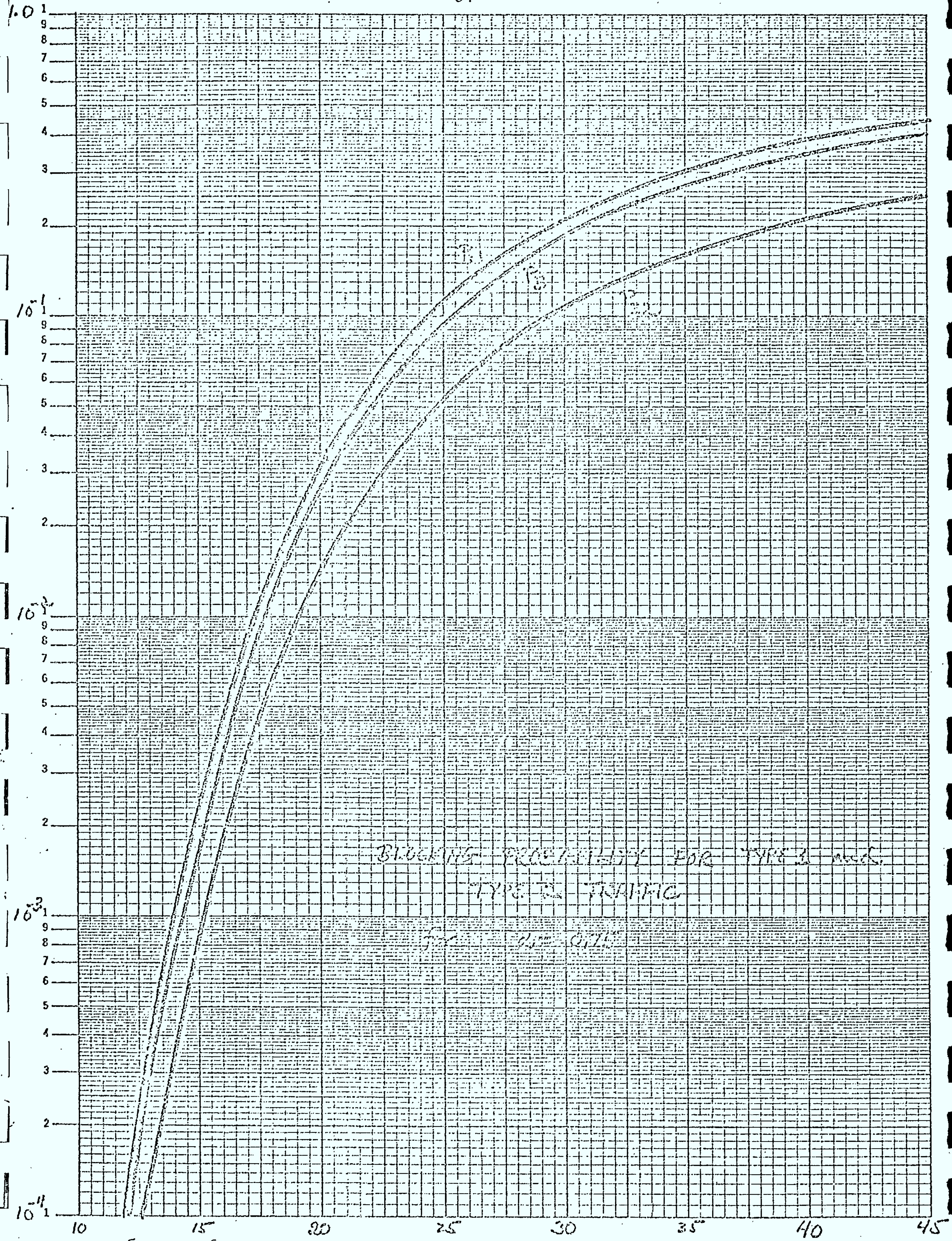
$$E[2Q_1+Q_2] = 2\rho_1[1-P_{B1}] + \rho_2[1-P_{B2}] \quad .$$

As an example we consider a T1 transmission system which is to handle voice calls with rates of 32 kb/s and 64 kb/s. The frame consists of 48 slots; type 1 traffic seizes two slots and type 2 traffice seizes one slot. We assume that both traffic types have mean holding times of 3 minutes and that the fraction of type 1 traffic is given by $\alpha=\frac{\lambda_1}{\lambda_1+\lambda_2}$. Figure 9 shows the blocking probability as a function of the offered voice traffic $\rho=\lambda/\mu$ with the traffic mix α as a parameter. The steepness of the curves demonstrate the usual sensitivity of P_B with respect to overloads in the offered traffic. The next 3 figures show the blocking probabilities of the various traffic types for traffic mixes of .25, .50, and .75. In a sense type 2 traffic receives priority over type 1 because type 1 traffic will be blocked whenever type 2 is, but not vice versa.


1.0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10
15
20
25
30
35
40
45


1.0
10^-1
10^-2
10^-3
10^-4
10
15
20
25
30
35
40
45
BLOCKING PROBABILITY FOR
TYPE 1 and TYPE 2 TRAFFIC


BLOCKING PROBABILITY FOR TYPE 1
and TYPE 2 TRAFFIC
for α = 0.50


1.0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
1
9
8
7
6
5
4
3
2
10
15
20
25
30
35
40
45
BLOCKING PROBABILITY FOR TYPE 1 and TYPE 2 TRAFFIC

4.3 Implications to Flow Control

The flow control procedure presented in Section 3.2 can be readily extended to the case where the voice traffic has several digitization rates. Let the frame consist of N slots, and assume that the number of voice slots is $N_v=2m$. As in Section 4.2 assume that class 1 voice traffic seizes two slots and that class 2 voice traffic seizes 1 slot. If the number of type 1 and type 2 calls in the system is i and j respectively, then the number of slots available to voice traffic at that instant is N-2i-j. In order to maintain the instantaneous offered data load constant, then the flow control procedure must be

$$\lambda_2(i,j) = (1 - \frac{2i+j}{N})\lambda_2 \quad .$$

This flow control procedure will have the effect of making the voice and data processes independent. As a result the data customer occupancy statistics will correspond to those of a single server queue with capacity $1-E(2Q_1+Q_2)/N$. The voice statistics will be the same as those in Section 4.2 since the voice process is independent of the data process.

One possibility which has been discussed by other researchers is flow control procedures in which the voice digitization rate of incoming calls is regulated according to the instantaneous loading of the hybrid switch. This approach involves trading off voice quality vs. call blocking probability vs. delay of data customers. It would be interesting to find out if simple solutions as the one presented above are possible for this case. We have not yet explored this possibility.

CHAPTER 5

STOCHASTIC PROCESSES ARISING IN PACKET VOICE/DATA INTEGRATED NETWORKS

5.1 Introduction

In the previous chapters, we have focused our attention on the hybrid-switched integration scheme. In this chapter, we survey some of the literature on the packet voice/data integration scheme. We examine some of the tools that can be used for analysis. We discuss the performance criteria used and their implications towards performance evaluation. Finally, we speculate briefly on the future development of packet voice/data integrated systems.

5.2 Various Packet Voice/Data Integration Schemes and Performance Criteria

There are basically two modes of packet transmission: datagram, and virtual circuit. In the datagram mode, each packet is routed independently through the network. The packets experience a variable crossnetwork delay and may arrive out of order. In data traffic, this may be of no concern. However, for packetized voice, some buffering and reassembly scheme is needed. In the virtual circuit mode, there is a call set-up procedure during which a virtual circuit is established. All packets associated with the call will then follow this path. This is analogous to circuit switching except that bandwidth is not reserved for the call. So the packets will still experience delay although the header overhead for packets will be reduced.

The basic analysis problem in the packet voice/data integration scheme is to study the queueing system generated by multiplexing two types of traffic, voice and data, with different characteristics through a common network.

As discussed in Chapter 1, voice traffic has stringent timing requirements. It will therefore be given priority over data. A number of priority queueing disciplines can be conceived. For example, we may have a nonpreemptive priority discipline in which the integrated switch will accept new packets for transmission only at the completion of transmitting a packet. It chooses a voice packet if one can be found, and when no more voice packets are ready for transmission, it multiplexes data packets into the remaining time slots. We may also have a preemptive priority discipline in which voice packets preempt data packets. The data packet preempted returns to the data buffer and waits for the next available time slot.

There is also a variety of performance criteria that can be used. The data performance is, as usual, measured by its delay characteristics. The voice performance criterion should be one that reflects timing requirements necessary for good quality speech. Two possibilities are blocking probability for voice, and the percentage loss of packets due to delay. The first possibility was examined by Arthurs and Stuck [20]. They examined the cases where the switch operates synchronously or asynchronously and analyzed the voice buffer size needed. They showed that the operation of the switch does not affect the blocking probability if the voice buffer is designed properly. The blocking probability itself is simply given by the Erlang B formula. The data delay is approximated by the delay seen by a low priority task in a two-level nonpreemptive priority queueing system with a single server. Analytical formulas are obtained for the calculation of an upper bound for the average data delay given the voice and data traffic statistics. The scheme proposed here is somewhat similar to the hybrid-switched systems analyzed in Chapter 2 where the number of data slots N_D per frame is zero. Its main virtue is that the formulas are easy to evaluate. However, there is no explicit consideration of the quality of speech; it is assumed that for a reasonably small blocking probability, good speech quality can be obtained. Also, the upper bound for the average data delay is generally pessimistic, so that the performance evaluation given is only a very crude one.

The percentage loss of packets criterion for voice is used in [22] and [13]. In this scheme, a "smoothing" buffer is assumed to be used to deliver packets which have arrived before a fixed maximum delay. The actual delay experienced during the call would be equivalent to this maximum delay. In this way, packets that have arrived on time will be delivered at a fixed delay. Packets which arrive later than the maximum allowable delay cannot be used in the voice output and is considered lost. Based on studies of the degradation of speech quality due to loss of packets, 5% or less of packet loss is acceptable. This criterion in turn translates into probability of packets experiencing a delay larger than the maximum value. The actual application of this criterion is difficult to carry out in practice, because it requires the evaluation of the voice delay distribution rather than the average voice delay. Approximations have been suggested in [22].

From the above discussion, we can see that the queueing system arising in packet voice/data integrated networks is of a well-known form. However, if we use conventional performance parameters such as those examined in [20], they may not be suitable as a true measure of the quality of speech. If we use a criterion appropriate for the evaluation of speech quality, as in [22], the queueing problem then becomes much more difficult to analyze. This is in contrast to the hybrid switching situation where the performance criteria used are standard and appropriate for the queueing problem.

5.3 Speculations on Future Developments of Packet Voice/Data Integrated Systems

It seems to us that future developments of packet voice/data integrated networks depend crucially on the development of speech processing algorithms for reassembling voice packets into the correct order, and maintaining reasonable quality of speech in the face of packet losses due to delay. As such, it seems that at this point, the more urgent issues are in the refinement of packetized voice transmission and its evaluation. Once methods of packetized voice transmission have been standardized, the incorporation of data as a low priority task should not be too formidable an undertaking. Certainly flow control schemes will have to be more complex since they need to handle two rather different types of traffic. However, the nature of the queueing system involved is well-known and should lend itself readily to analysis.

Another factor that will play an important part is the cost involved in incorporating packetized voice into existing transmission facilities. This requires the development of low-cost speech processing devices. Recent advances in digital technology show that such devices may be realized in the not too distant future.

CHAPTER 6

CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

In this report, we have analyzed in detail some stochastic processes which arise in connection with hybrid-switched integrated networks, and we have given exact analyses for the performance parameters. Since the exact analyses are difficult to carry out numerically, we have developed useful approximate methods of performance analysis. We have analyzed the data queue build-up phenomenon and we have proposed a simple flow control method that will reduce the size of the data queue. We have also provided the expressions for the blocking probabilities for voice traffic with different digitization rates. The packet voice/data switching alternative has also been briefly discussed.

We feel that our main contributions are the much more complete analyses we have given for the hybrid switch, and the flow control scheme which leads to vastly improved system performance. As a result, a solid understanding has been developed for the behaviour of the hybrid switch. The main conclusion is that hybrid switching with flow control is definitely a viable scheme for integrating voice and data in a common communication network. It offers on the one hand the transparency of circuit-switched voice, and on the other the efficiency of packet switching for data. As long as some type of flow control procedure which regulates the data queue is used, the system can be designed, at least at the link level, to satisfy the performance criteria specified.

It is difficult to compare at this stage the performance of hybrid-switching with that of packet switching. It seems that hybrid switching is better understood at this point, and perhaps more readily adapted to existing networks. Packetized voice transmission is still in an early stage of its development, but conceptually it is well-suited to integration. A natural evolutionary path would seem to be to first develop hybrid-switched integrated systems which can be readily implemented in existing networks, then develop packet voice/data systems after we have gained better understanding into the transmission of packetized voice.

In this study, we have concentrated on the fundamental queueing problems arising at the link level in the hybrid switching scheme. In an integrated network, many other system aspects will have to be analyzed. As future research directions then, we mention some problems which have been suggested directly by our work, and some others concerning the network performance as a whole.

(1) We have proposed in Section 2.7 a fluid approximation for the analysis of the hybrid switch. We believe it is simpler to use than previous approximation schemes proposed in the literature. It would be very useful to compare the various approximation schemes from the point of view of accuracy, and sensitivity to parameter changes and model assumptions. The aim is to develop a simple and yet robust approximation scheme that can be used effectively for network design.

(2) In Chapter 3, we have analyzed a very simple flow control procedure based on regulating the incoming data traffic rate. This suggests very naturally control problems based on the control of arrival and service rates

associated with the integrated network traffic. Such problems would be interesting generalizations of control problems associated with a single queue since effective control of the data queue in our case depends on the voice traffic.

(3) We have indicated in Chapter 4 that we may combine voice digitization rate control and our data flow control scheme into a single flow control procedure. It remains to analyze the performance of such a hybrid-switched system in terms of the blocking probabilities for voice and the delay for data. Other flow control procedures, such as limiting the size of the data queue when voice traffic is heavy, may also be considered. It would be important to have a performance comparison of the flow control methods, using perhaps the mathematical models we have constructed in Chapters 3 and 4.

(4) We have not performed an in-depth analysis ourselves on the packet voice/data switching scheme. However, our discussion in Chapter 5 shows that better understanding of packetized voice transmission is needed. It would be very useful to construct analysis models which are mathematically tractable, but which also explicitly incorporate speech quality considerations. Once this has been successfully accomplished, we can then investigate the integrated packet switching problem.

(5) The results we have obtained at the link level should be extended to analyze network performance. Performance parameters of interest are the throughput, the end-to-end blocking probability, the average end-to-end delay for data, the average circuit set-up and disconnect times, etc. Other system aspects not present at the link level, such as routing, will have to

APPENDIX 1

CHARACTERIZATION OF THE GENERATING FUNCTIONS FOR DATA WITH CONSTANT PACKET LENGTHS

We derive here the equations characterizing the generating functions for data with constant packet lengths. We define

$$p^t_{mn} = \Pr\{Q^V_t = m,\ Q^D_t = n\}$$

As in Section 2.4, we can calculate the transition probabilities for the joint Q^V_t, Q^D_t process as follows. We have the conditional probabilities,

$$\Pr\{Q^V_t = m \mid Q^V_{t-1} = j,\ Q^D_{t-1} = i\}$$

$$= \sum_{k=0}^{\min(j,m)} q^V_{jk}\, p^V_{m-k} \qquad m=0,1,\ldots,N_V-1 \qquad \text{(A.1.1)}$$

$$\Pr\{Q^V_t = N_V \mid Q^V_{t-1} = j,\ Q^D_{t-1} = i\}$$

$$= \sum_{k=0}^{j} [q^V_{jk} \sum_{r=N_V}^{\infty} p_{r-k}] \qquad \text{(A.1.2)}$$

$$\Pr\{Q^D_t = n \mid Q^V_t = m,\ Q^V_{t-1} = j,\ Q^D_{t-1} = i\}$$

$$= \frac{e^{-\theta b}(\theta b)^n}{n!} \qquad \text{if } i \le N-j-1$$

$$= \frac{e^{-\theta b}(\theta b)^{n+N-i-j}}{(n+N-i-j)!} \qquad \text{if } N-j \le i,\ i+j-N \le n \qquad \text{(A.1.3)}$$

be addressed. This is a largely unexplored area from the theoretical point of view and in a sense our detailed link analysis is to develop the tools for the network study. It is a difficult and challenging problem on which much research needs to be done.

Using $\quad \Pr\{Q_t^V=m,\ Q_t^D=n \mid Q_{t-1}^V=j,\ Q_{t-1}^D=i\}$

$$= \Pr\{Q_t^D=n \mid Q_t^V=m,\ Q_{t-1}^V=j,\ Q_{t-1}^D=i\}\ .$$

$$\Pr\{Q_t^V=m \mid Q_{t-1}^V=j,\ Q_{t-1}^D=i\}$$

we get the transition probabilities

$$
\begin{aligned}
p_{ji,mn} &= \sum_{k=0}^{\min(j,m)} q_{jk}^V\, p_{m-k}^V \frac{e^{-\theta b}(\theta b)^n}{n!} && \text{for } \begin{array}{l} i \le N-j-1 \\ m < N_V \end{array} \\
&= \sum_{k=0}^{\min(j,m)} q_{jk}^V\, p_{m-k}^V \frac{e^{-\theta b}(\theta b)^{n+N-i-j}}{(n+N-i-j)!} && \text{for } \begin{array}{l} N-j \le i \\ m < N_V \\ i+j-N \le n \end{array} \\
&= \sum_{k=0}^{j} [q_{jk}^V \sum_{r=N_V}^{\infty} p_{r-k}^V] \frac{e^{-\theta b}(\theta b)^n}{n!} && \text{for } \begin{array}{l} i \le N-j-1 \\ m = N_V \end{array} \\
&= \sum_{k=0}^{j} [q_{jk}^V \sum_{r=N_V}^{\infty} p_{r-k}^V] \frac{e^{-\theta b}(\theta b)^{n+N-i-j}}{(n+N-i-j)!} && \text{for } \begin{array}{l} N-j \le i \\ m = N_V \\ i+j-N \le n \end{array}
\end{aligned}
\tag{A.1.4}
$$

From (A.1.4), we find that

$$\sum_{n=0}^{\infty} p_{ji,mn} \, z^n = \sum_{k=0}^{\min(j,m)} q^v_{jk} \, p^v_{m-k} \, e^{-\theta b(1-z)} \qquad \text{for} \quad \begin{array}{l} i \le N-j-i \\ m < N_v \end{array}$$

$$= \sum_{k=0}^{\min(j,m)} q^v_{jk} \, p^v_{m-k} \, e^{-\theta b(1-z)} z^{i+j-N} \qquad \text{for} \quad \begin{array}{l} N-j \le i \\ m < N_v \end{array}$$

$$= \sum_{k=0}^{j} [q^v_{jk} \sum_{r=N_v}^{\infty} p^v_{r-k}] e^{-\theta b(1-z)} \qquad \text{for} \quad \begin{array}{l} i \le N-j-1 \\ m = N_v \end{array}$$

$$= \sum_{k=0}^{j} [q^v_{jk} \sum_{r=N_v}^{\infty} p^v_{r-k}] e^{-\theta b(1-z)} z^{i+j-N} \qquad \text{for} \quad \begin{array}{l} N-j \le i \\ m = N_v \end{array} \tag{A.1.5}$$

Let $p^t_m(z) = \sum_{n=0}^{\infty} p^t_{mn} z^n$. Since

$$\sum_{n=0}^{\infty} p^t_{mn} z^n = \sum_{i,j} \sum_{n=0}^{\infty} p_{ji,mn} \, p^{t-1}_{ji} z^n$$

we find, after some computations, that

$$p^t_m(z) = \sum_{j=0}^{N_v} \sum_{i=0}^{N-j-1} \sum_{k=0}^{\min(j,m)} q^v_{jk} \, p^v_{m-k} \, e^{-\theta b(1-z)} p^{t-1}_{ji} (1 - z^{i+j-N})$$

$$+ \sum_{j=0}^{N_v} \sum_{k=0}^{\min(j,m)} q^v_{jk} \, p^v_{m-k} \, e^{-\theta b(1-z)} p^{t-1}_{j}(z) z^{j-N} \tag{A.1.6}$$

$$\text{for} \quad m = 0, 1, \ldots, N_v - 1$$

Assuming that the stability condition for the queueing process holds, we have, on letting $t \to \infty$ in (A.1.6),

$$\pi_m(z) = \sum_{j=0}^{N_v} \sum_{i=0}^{N-j-1} \sum_{k=0}^{\min(j,m)} q_{jk}^v \, p_{m-k}^v \, e^{-\theta b(1-z)} \pi_{ji}(1-z^{i+j-N})$$

$$+ \sum_{j=0}^{N_v} \sum_{k=0}^{\min(j,m)} q_{jk}^v \, p_{m-k}^v \, e^{-\theta b(1-z)} \pi_j(z) z^{j-N} \qquad \text{(A.1.7)}$$

$$\text{for } m=0,1,\ldots,N_v-1$$

Equation (A.1.7) is precisely Eq. (2.5.1). In a similar way, we can derive the equation for $\pi_{N_v}(z)$, which is precisely Eq. (2.5.2).

Equations (2.5.1)-(2.5.2) may be rewritten as

$$A_c(z)\pi(z) = b_c(z) \qquad \text{(A.1.8)}$$

To determine $b_c(z)$, we need $K=\frac{N_v+1}{2}(2N-N_v)$ equations. Now the matrix $A_c(z)$ is of the form

$$A_c(z)=\begin{bmatrix} 1-c_{00}e^{-\theta b(1-z)}z^{-N} & -c_{01}e^{-\theta b(1-z)}z^{1-N} & \cdots & -c_{0,N_v}e^{-\theta b(1-z)}z^{N_v-N} \\ -c_{10}e^{-\theta b(1-z)}z^{-N} & 1-c_{11}e^{-\theta b(1-z)}z^{1-N} & \cdots & -c_{1,N_v}e^{-\theta b(1-z)}z^{N_v-N} \\ \vdots & \vdots & & \vdots \\ -c_{N_v,0}e^{-\theta b(1-z)}z^{-N} & & & 1-c_{N_v,N_v}e^{-\theta b(1-z)}z^{N_v-N} \end{bmatrix}$$

Then det $A_c(z)$ will be of the form

$$\det A_c(z) = \frac{1}{z^K}[z^K + \sum_{j=0}^{K-1} c_j \, z^j \, e^{-\theta b(1-z)k_j}]$$

for some set of integers k_j, $j=0,\ldots,K-1$ and real numbers c_j, $j=0,1,\ldots,K-1$ with $|c_j|\le 1$. Applying Rouche's theorem, we find that there are K distinct roots of det $A_c(z)$ inside $|z|\le 1$. The solution of (A.1.8) may now proceed in the way discussed in Section 2.5

APPENDIX 2

ANALYTICAL EXPRESSIONS FOR THE AVERAGE DATA DELAY IN A SINGLE CHANNEL HYBRID SWITCH

We specialize the results of Section 2.5 to the case where $N_V=1$ and $N_D=0$. Equation (2.5.3) in this case becomes

$$\begin{bmatrix} 1-p_0^v e^{-\theta b(1-z)} z^{-1} & -q_{10}^v p_0^v e^{-\theta b(1-z)} \\ -(1-p_0^v) e^{-\theta b(1-z)} z^{-1} & 1-(1-q_{10}^v p_0^v) e^{-\theta b(1-z)} \end{bmatrix} \begin{bmatrix} \pi_0(z) \\ \pi_1(z) \end{bmatrix} = \begin{bmatrix} p_0^v e^{-\theta b(1-z)} \pi_{00}(1-z^{-1}) \\ (1-p_0)\pi_{00} e^{-\theta b(1-z)}(1-z^{-1}) \end{bmatrix} \tag{A.2.1}$$

There is only 1 root inside $|z|\le 1$, and it is $z=1$. We need only the normalization condition

$$\sum_i \pi_i(1) = \sum_{i,j} \pi_{ij} = 1 \tag{A.2.2}$$

to determine the unknown quantity π_{00} in (A.2.1). Solving for $\pi_0(z)$ and $\pi_1(z)$, we obtain, after some computations,

$$\pi_0(z) = p_0^v e^{-\theta b(1-z)} \pi_{00}(z-1)[1-q_{11}^v e^{-\theta b(1-z)}]\{z[1-e^{-\theta b(1-z)}] - p_0^v e^{-\theta b(1-z)} + zq_{10}^v p_0^v e^{-\theta b(1-z)} + q_{11}^v p_0^v e^{-2\theta b(1-z)}\}^{-1} \tag{A.2.3}$$

$$\pi_1(z) = \pi_{00} e^{-\theta b(1-z)}(z-1)(1-p_0^v)\{z[1-e^{-\theta b(1-z)}] - p_0^v e^{-\theta b(1-z)} + zq_{10}^v p_0^v e^{-\theta b(1-z)} + q_{11}^v p_0^v e^{-2\theta b(1-z)}\}^{-1} \tag{A.2.4}$$

Since the numerators and denominators of $\pi_0(z)$ and $\pi_1(z)$ all vanish at z=1, we need to apply L'Hospital's rule to evaluate $\pi_0(1)$ and $\pi_1(1)$. We find

$$\pi_0(1) = \frac{p_0^v q_{10}^v \pi_{00}}{q_{10}^v p_0^v + q_{11}^v p_0^v \theta b - \theta b} \tag{A.2.5}$$

$$\pi_1(1) = \frac{\pi_{00}(1-p_0^v)}{q_{10}^v p_0^v + q_{11}^v p_0^v \theta b - \theta b} \tag{A.2.6}$$

Using the normalization condition (A.2.2), we get

$$\pi_{00} = \frac{q_{10}^v p_0^v + q_{11}^v p_0^v \theta b - \theta b}{1 - p_0^v + q_{10}^v p_0^v} \tag{A.2.7}$$

Equations (A.2.3), (A.2.4) and (A.2.7) completely determine $\pi_0(z)$ and $\pi_1(z)$. Now the generating function for the steady state data customers Q_D is given by

$$G_D(z) = \sum_i \pi_i(z)$$

so that in this case

$$G_D(z) = \pi_{00} e^{-\theta b(1-z)} (z-1)[1-p_0^v q_{11}^v e^{-\theta b(1-z)}] \cdot \{z[1-e^{-\theta b(1-z)}]$$

$$- p_0^v e^{-\theta b(1-z)} + z q_{10}^v p_0^v e^{-\theta b(1-z)} + q_{11}^v p_0^v e^{-2\theta b(1-z)}\}^{-1} \tag{A.2.8}$$

The average number of data customers is then given by

$$E(Q_D) = \frac{d}{dz} G_D(z)\bigg|_{z=1} \tag{A.2.9}$$

Let $u(z)$ and $v(z)$ be the numerator and denominator of $G_D(z)$, respectively. $G_D'(z)$ will be indeterminate at $z=1$ since its numerator as well as denominator vanish at $z=1$. Applying L'Hospital's rule to $G_D'(z)$ gives

$$G_D'(1) = \frac{v'(1)u''(1)-u'(1)v''(1)}{2[v'(1)]^2} \tag{A.2.10}$$

Computation then yields

$$\begin{aligned} EQ_D = {} & [2(q_{10}^v p_0^v + q_{11}^v p_0^v \theta b - \theta b)\pi_{00}\theta b(1-2q_{11}^v p_0^v) \\ & - \pi_{00}(1-p_0^v q_{11}^v)\theta b(2q_{10}^v p_0^v + 3q_{11}^v p_0^v \theta b - \theta b - 2)][2(q_{11}^v p_0^v \theta b + q_{10}^v p_0^v - \theta b)^2]^{-1} \end{aligned} \tag{A.2.11}$$

which is precisely Eq. (2.5.5).

APPENDIX 3

COMPUTING THE OVERLOAD PERIOD STATISTICS

The voice process $Q_V(t)$ is modelled by a continuous-time, finite state Markov process with state transition diagram given below.

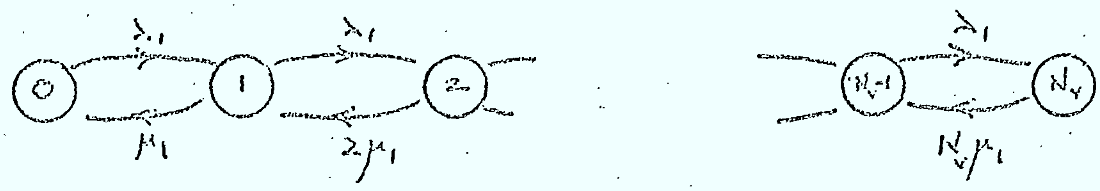


Figure 13.

If the process is in state i then the next state will be i+1 with probability $\alpha_i = \frac{\lambda_1}{\lambda_1 + i\mu_1} = \frac{\rho_V}{\rho_V + i}$ or state i-1 with probability $1-\alpha_i$. The mean occupancy time in state i is

$$\tau_i = \begin{cases} \frac{1}{\lambda_1 + i\mu_1} & i=0,\ldots,N_V-1 \\ \frac{1}{N_V\mu_1} & i=N_V \end{cases}$$

Suppose that the data arrival rate is θb, then the system is in the overload state if $i \geq i^* = N - \lfloor \theta b \rfloor$, where $\lfloor \theta b \rfloor$ is the largest integer $\leq \theta b$. We are interested in computing the mean and variance of the elapsed time between the instant where $Q_V(t)$ first increases to i^* to the instant when $Q_V(t)$ finally becomes less than i^*. The problem has a simple recursive solution. Consider first the case where $i^*=N_V$. Let T_{i^*} be the mean overload period, then $T_{i*}=\tau_{N_v}$. Now suppose that $i*=N_v-1$. From the diagram below it is clear that T_{i*} is now given by the sum of a random number of random times:

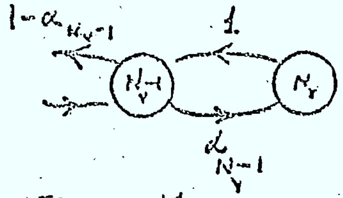


Figure 14.

$$T_{N_v-1} = \begin{cases} \tau_{N_v-1}(0) & \text{with probability } 1-\alpha_{N_v-1} \\ \tau_{N_v-1}(0)+\sum_{k=1}^{K}(\tau_{N_v-1}(k)+\tau_{N_v}(k)) & \text{for } K\geq 1 \text{ with probability} \\ & (1-\alpha_{N_v-1})\alpha_{N_v-1}^{K} \end{cases}$$

where $\tau_i(k)$ are independent, identically exponentially distributed random variables with mean τ_i. The random variable K has a geometric distribution with parameter α_{N_v-1}. The mean and variance are found as follows.

$$E[T_{N_v-1}] = E_K[E[T_{N_v-1}|K]]$$

$$= \sum_{k=0}^{\infty} E[T_{N_v-1}|K](1-\alpha_{N_v-1})\alpha_{N_v-1}^{k}$$

$$= \tau_{N_v-1} + (1-\alpha_{N_v-1}) \sum_{k=1}^{\infty} k(\tau_{N_v-1}+\tau_{N_v})\alpha_{N_v-1}^{k}$$

$$= \tau_{N_v-1} + \frac{\alpha_{N_v-1}}{1-\alpha_{N_v-1}}(\tau_{N_v-1}+E[T_{N_v}])$$

$$VAR[T_{N_v-1}] = E[(T_{N_v-1}-E(T_{N_v-1}))^2]$$

$$= \tau^2_{N_v-1} + \frac{\alpha_{N_v-1}}{1-\alpha_{N_v-1}} (\tau^2_{N_v-1}+VAR[T_{N_v}])$$

$$+ \frac{\alpha_{N_v-1}}{(1-\alpha_{N_v-1})^2} (\tau_{N_v-1}+E[T_{N_v}])^2$$

We have intentionally written the above expressions so as to point out the dependence of $E(T_{N_v-1})$ and $VAR(T_{N_v-1})$ on $E(T_{N_v})$ and $VAR(T_{N_v})$. Now suppose that T_{i*} has been computed and that T_{i*-1} is to be computed. Then from the figure below and from the above development it is clear that the following recursive formulas can be used.

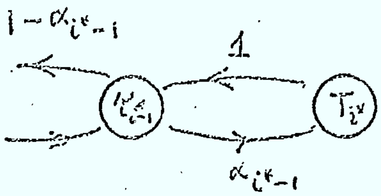


Figure 15.

$$E[T_{i*-1}] = \tau_{i*-1} + \left(\frac{\alpha_{i*-1}}{1-\alpha_{i*-1}}\right)[\tau_{i*-1}+E(T_{i*})]$$

$$VAR[T_{i*-1}] = \tau^2_{i*-1} + \left(\frac{\alpha_{i*-1}}{1-\alpha_{i*-1}}\right)[\tau^2_{i*-1}+VAR[T_{i*}]]$$

$$+ \frac{\alpha_{i*-1}}{(1-\alpha_{i*-1})^2} [\tau_{i*-1}+E[T_{i*}]]^2$$

In addition to the mean and variance of the overload period we are also interested in the mean and variance of the stable periods. These are simply the elapsed time between the instant when $Q_V(t)$ first becomes less than i*-1 to the instant when it first returns to a value greater than i*-1. Clearly the method developed above for the mean and variance of the overload periods can also be used to find the same parameters for the stable periods.

APPENDIX 4

STEADY STATE PROBABILITIES OF HYBRID SWITCH WITH FLOW CONTROL

The transition diagram for the hybrid switch with flow control is given in Figure 8 . The difference equations for the steady state probabilities are given by:

For $0<i<N_v$, $0<j<\infty$

$$(\lambda_1+i\mu_1+\lambda_2(1-\frac{i}{N})+(N-i)\mu_2)P_{ij}$$

$$=\lambda_1 P_{i-1,j}+(i+1)\mu_1 P_{i+1,j}+\lambda_2(1-\frac{i}{N})P_{i,j-1}+(N-i)\mu_2 P_{i,j+1};$$

for $i=0$, $0<j<\infty$

$$(\lambda_1+\lambda_2+N\mu_2)P_{0j}=\mu_1 P_{1j}+\lambda_2 P_{0,j-1}+N\mu_2 P_{0,j+1};$$

for $i=N_v$, $0<j<\infty$

$$(N_v\mu_1+\lambda_2(1-\frac{N_v}{N})+(N-N_v)\mu_2)P_{N_v,j}=\lambda_1 P_{N_v-1,j}+\lambda_2(1-\frac{N_v}{N})P_{N_v,j-1}$$

$$+(N-N_v)\mu_2 P_{N_v,j+1};$$

for $j=0$, $0<i<N_v$

$$(\lambda_1+i\mu_1+\lambda_2(1-\frac{i}{N}))P_{i0}=\lambda_1 P_{i-1,0}+(i+1)\mu_1 P_{i+1,0}+(N-i)\mu_2 P_{i,1};$$

for $j=0$, $i=0$

$$(\lambda_1+\lambda_2)P_{00}=\mu_1 P_{10}+N\mu_2 P_{01};$$

for $j=0$, $i=N_v$

$$(N_v\mu_1+\lambda_2(1-\frac{N_v}{N}))P_{N_v,0} = \lambda_1 P_{N_v-1,0} + (N-N_v)\mu_2 P_{N_v,1} .$$

Note that every equation above can be decomposed into a set of local balance equations of the form

$$P_{i,j} = \frac{1}{i}\left(\frac{\lambda_1}{\mu_1}\right)P_{i-1,j}$$

$$P_{i,j} = \left(\frac{\lambda_2}{N\mu_2}\right)P_{i,j-1}$$

If we can find a P_{ij} satisfying the local equations, then the steady state equations will also be satisfied. Clearly the solution below does this.

$$P_{ij} = \frac{\frac{1}{i!}\left(\frac{\lambda_1}{\mu_1}\right)^i}{\sum_{i'=0}^{N_v}\left(\frac{\lambda_1}{\mu_1}\right)^{i'}\frac{1}{i'!}}\,(1-\frac{\lambda_2}{N\mu_2})(\frac{\lambda_2}{N\mu_2})^j$$

$$= P_i P_j$$

where P_i is the steady state probability for an $M/M/N_v/N_v$ queueing system and P_j is the steady state probability for an M/M/1 system.

APPENDIX 5

STEADY STATE PROBABILITIES FOR THE MULTICAPACITY PROBLEM

The state transition diagram for the multicapacity problem is given in Figure 16. The difference equations for the steady state probabilities are given by:

For $0<2j+k<2m-1$,

$$(\lambda_1+\lambda_2+k\mu_2+j\mu_1)P_{jk}=\lambda_1 P_{j-1,k}+\lambda_2 P_{j,k-1}+(k+1)\mu_2 P_{j,k+1}+(j+1)\mu_1 P_{j+1,k};$$

for $j=0$, $0<k\leq 2m-2$

$$(\lambda_1+\lambda_2+k\mu_2)P_{0k}=\lambda_2 P_{0,k-1}+(k+1)\mu_2 P_{0,k+1}+\mu_1 P_{1,k} \quad ;$$

for $k=0$, $0<j\leq m-1$

$$(\lambda_1+\lambda_2+j\mu_1)P_{j0}=\lambda_1 P_{j-1,0}+\mu_2 P_{j,1}+(j+1)\mu_1 P_{j+1,k} \quad ;$$

for $j=0$, $k=0$

$$(\lambda_1+\lambda_2)P_{00}=\mu_1 P_{10}+\mu_2 P_{01} \quad ;$$

for $j=0$, $k=2m-1$

$$(\lambda_2+(2m-1)\mu_2)P_{0,2m-1}=\lambda_2 P_{0,2m-2}+2m\mu_2 P_{0,2m} \quad ;$$

for $j=0$, $k=2m$

$$2m\mu_2 P_{0,2m}=\lambda_2 P_{0,2m-1} \quad ;$$

for $j=0$, $i=N_v$

$$(N_v\mu_1+\lambda_2(1-\frac{N_v}{N}))P_{N_v,0} = \lambda_1 P_{N_v-1,0} + (N-N_v)\mu_2 P_{N_v,1} .$$

Note that every equation above can be decomposed into a set of local balance equations of the form

$$P_{i,j} = \frac{1}{i}\left(\frac{\lambda_1}{\mu_1}\right)P_{i-1,j}$$

$$P_{i,j} = \left(\frac{\lambda_2}{N\mu_2}\right)P_{i,j-1}$$

If we can find a P_{ij} satisfying the local equations, then the steady state equations will also be satisfied. Clearly the solution below does this.

$$P_{ij} = \frac{\frac{1}{i!}\left(\frac{\lambda_1}{\mu_1}\right)^i}{\sum_{i'=0}^{N_v}\left(\frac{\lambda_1}{\mu_1}\right)^{i'}\frac{1}{i'!}}(1-\frac{\lambda_2}{N\mu_2})(\frac{\lambda_2}{N\mu_2})^j$$

$$= P_i P_j$$

where P_i is the steady state probability for an $M/M/N_v/N_v$ queueing system and P_j is the steady state probability for an M/M/1 system.

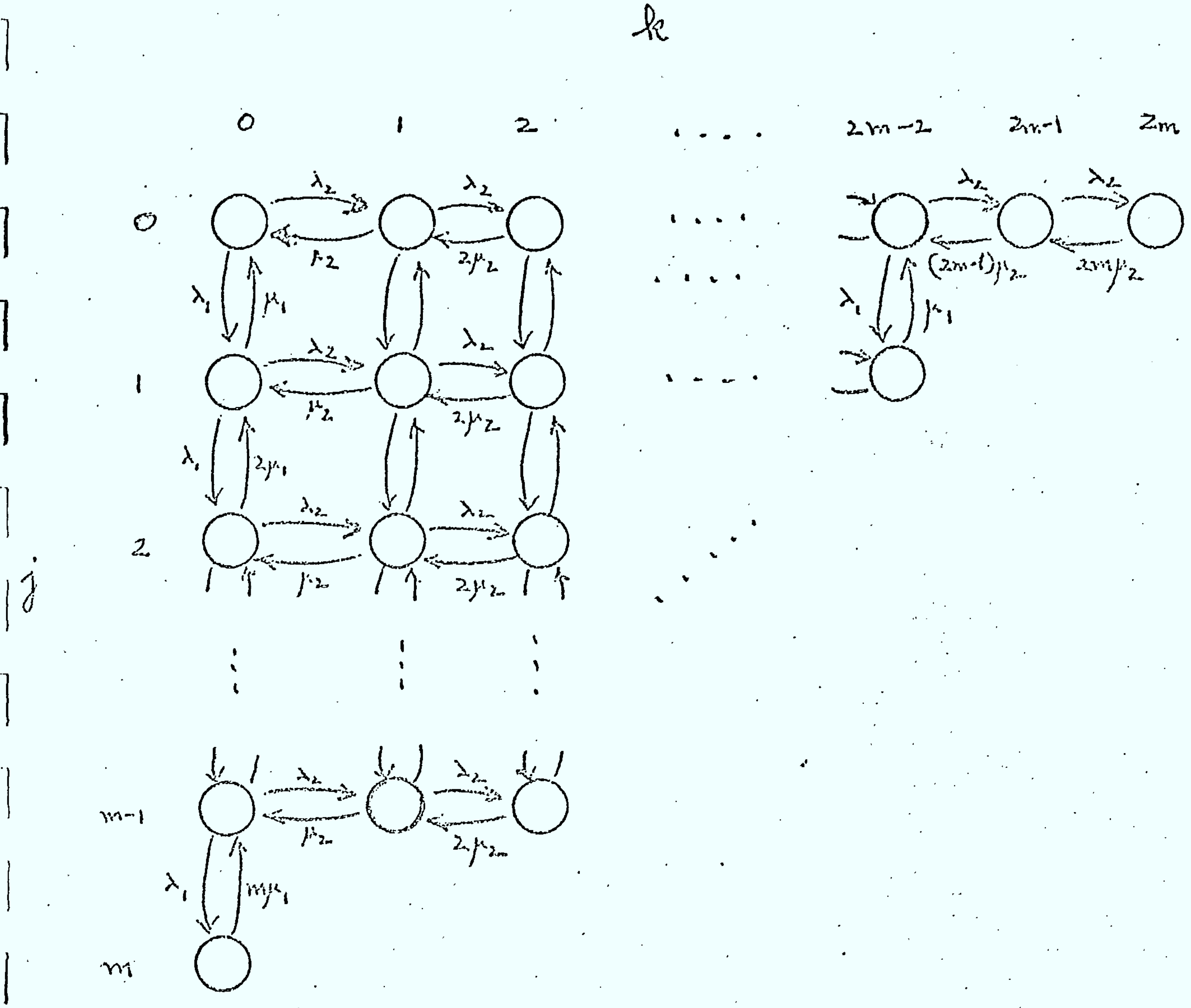


FIGURE 16. STATE TRANSITION DIAGRAM FOR MULTICAPACITY PROBLEM

for $0<j\leq m-1$, $2j+k=2m-1$

$$(\lambda_2+k\mu_2+j\mu_1)P_{jk}=\lambda_1 P_{j-1,k}+\lambda_2 P_{j,k-1}+(k+1)\mu_2 P_{j,k+1} \;;$$

for $0<j\leq m-1$, $2j+k=2m$

$$(k\mu_2+j\mu_1)P_{jk}=\lambda_2 P_{j,k-1}+\lambda_1 P_{j-1,k} \;;$$

for $j=m$, $k=0$

$$m\mu_1 P_{m,0}=\lambda_1 P_{m-1,0} \;.$$

Note that the terms on either side of the above equations can be matched in pairs of the form:

$$P_{jk} = \frac{1}{j}\left(\frac{\lambda_1}{\mu_1}\right)P_{j-1,k}$$

$$P_{jk} = \frac{1}{k}\left(\frac{\lambda_2}{\mu_2}\right)P_{j,k-1}$$

Therefore if we can find a solution satisfying these two equations for the allowed values of j and k, then the solution will also be the steady state probabilities. It is easy to verify that the following equation achieves this.

$$P_{jk} = \frac{\rho_1^j}{j!}\,\frac{\rho_2^k}{k!}\,P_{00} \qquad \begin{array}{l} 0\leq j\leq m \\ 0\leq 2j+k\leq 2m \end{array}$$

where

$$\rho_1 = \frac{\lambda_1}{\mu_1} \quad , \qquad \rho_2 = \frac{\lambda_2}{\mu_2} \quad \text{, and}$$

$$P_{00} = \left[\sum_{j=0}^{m} \sum_{k=0}^{2m-2j} \frac{\rho_1^j}{j!} \frac{\rho_2^k}{k!} \right]^{-1}$$

REFERENCES

[1] B. Gold, "Digital speech networks", Proc. IEEE, vol. 65, no. 12, Dec. 1977, 1636-1658.

[2] M.J. Ross, A.C. Tabbot and J.A. Waite, "Design approaches and performance criteria for integrated voice/data switching", Proc. IEEE, vol. 65, no. 9, Sept. 1977, 1283-1295.

[3] A. Leon-Garcia, R.H. Kwong and A.N. Venetsanopoulos, First report of "A study of a joint stochastic process for an integrated network", prepared for the Department of Communications, Ottawa, under Contract OSU79-00041, Sept. 1979.

[4] M.J. Ross, "System engineering of integrated voice and data switches", International Conf. on Communications, June 1978, Toronto, Canada.

[5] G. Coviello and P. Vena, "Integration of circuit/packet switching by a SENET (Slotted Envelope Network) concept", Proc. Nat. Telecommunications Conf., New Orleans, Louisiana, Dec. 1975, 42.12-42.17.

[6] I. Gitman, H. Frank, B. Occhiogrosso and W. Hsieh, "Issues in integrated network design", Proc. International Conf. on Communications, 1977, 38.1-36-38.1-43.

[7] K. Kummerle, "Multiplexer performance for integrated line and packet-switched traffic", Proc. of the 2nd International Conf. on Computer Communications, Stockholm, Sweden, Aug. 1974, 517-523.

[8] C. Jenny and K. Kummerle, "Distributed processing within an integrated circuit/packet-switching node", IEEE Trans. Communications, vol. COM-24, Oct. 1976, 1098-1100.

[9] M.J. Fischer and T.C. Harris, "A model for evaluating the performance of an integrated circuit- and packet-switched multiplex structure", IEEE Trans. Communications, Feb. 1976, 195-202.

[10] B. Occhiogrosso, I. Gitman, W. Hsieh and H. Frank, "Performance analysis of integrated switching communications systems", Proc. Nat. Telecommunications Conf., 1977.

[11] W. Hsieh, I. Gitman and B. Occhiogrosso, "Design of hybrid-switched networks for voice and data", International Conf. on Communications, 1978, Toronto, Canada, 20.1.1-20.1.9.

[12] N. Keyes and M. Gerla, "Report on experience in developing a hybrid packet and circuit switched network", International Conf. on Communications, June 1978, Toronto, Canada.

[13] J.W. Forgie and A.G. Nemeth, "An efficient packetized voice/data network using statistical flow control", International Conf. on Communications, 1978, Toronto, Canada.

[14] H. Rudin, "Studies in the integration of circuit packet switching", International Conf. on Communications, 1978, Toronto, Canada.

[15] H. Miyahara and T. Hasegawa, "Integrated switching with variable frame and packet", International Conf. on Communications, 1978, Toronto, Canada.

[16] H. Miyahara and T. Hasegawa, "Performance evaluation of modified multiplexing technique with two types of packet for circuit and packet switched traffic", International Conf. on Communications, 1979, Boston, Mass.

[17] L.H. Chang, "Analysis of integrated voice and data communication network", Ph.D. Thesis, Department of Electrical Engineering, Carnegie-Mellon University, Nov. 1977.

[18] I. Gitman and H. Frank, "Economic analysis of integrated voice and data networks: a case study", Proc. IEEE, vol. 66, no. 11, Nov. 1979, 1549-1570.

[19] C.J. Weinstein, M.L. Malpass and M.J. Fischer, "Data traffic performance of an integrated circuit- and packet-switched multiplex structure", M.I.T. Lincoln Lab. Technical Note, 1978-41, Oct. 1978.

[20] E. Arthurs and B.W. Stuck, "A theoretical traffic performance analysis of an integrated voice-data virtual circuit packet switch", IEEE Trans. Communications, vol. COM-27, July 1979, 1104-1111.

[21] K. Kummerle and H. Rudin, "Packet and circuit switching: cost/performance boundaries", Computer Networks, vol. 2, Feb. 1978, 3-17.

[22] G.J. Coviello, "Comparative discussion of circuit- vs. packet-switched voice", IEEE Trans. Communications, vol. 27, no. 8, Aug. 1979, 1153-1159.

[23] R.R. Anderson, G.J. Foschini and B. Gopinath, "A queueing model for a hybrid data multiplexer", B.S.T.J., vol. 58, Feb. 1979, 279-300.

[24] R.B. Cooper, Introduction to Queueing Theory, MacMillan, 1972.

[25] I.L. Mitrany and B. Avi-Itzhak, "A many-server queue with service interruptions", Operations Research, vol. 16, no. 3, 1968, 628-638.

[26] L. Kleinrock, Queueing Systems, Vol. I, John Wiley, New York, 1975.

[27] U. Yechiali and P Naor, "Queueing problems with heterogeneous arrivals and service", Operations Research, vol. 19, 1971, 722-734.

[28] J. Hsu, "A general queueing model for buffer storage problems", IEEE Trans. Communications, vol. COM-21, June 1973, 744-747.

[29] L. Kleinrock, Queueing Systems, Vol. II, John Wiley, 1976.