ANALYSIS OF TRAFFIC PROBLEMS OF INTEGRATED NETWORKS

Thesis by

Yurdaer N. Doğanata

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1987

(Submitted May 1, 1987)

ACKNOWLEDGMENTS

It is my pleasure to acknowledge people by whom I have been influenced throughout the course of this research. I am most grateful to Professor E. C. Posner for his constant encouragement and patience. His constructive suggestions, comments and questions improved this work and made it a lot clearer. He also made it possible for me to meet several people with whom I had valuable discussions that contributed significantly to my thesis.

I would like to thank Dr. A. A. Lazar, Dr. R. Mazumdar and Joseph Ferrandiz from Colombia University, for the stimulating and encouraging talks we had.

Special thanks are also extended to Dr. G. Hess and Dr. Jim Mikulski from MOTOROLA Communications and Electronics, Inc., for providing the documents about MOTOROLA trunked mobile system.

Finally, I wish to express my appreciation and love to my wife, Zinnur, for always being there standing by me and supporting me by all means.

ANALYSIS OF TRAFFIC PROBLEMS OF INTEGRATED NETWORKS

ABSTRACT

A new approach to the analysis of Markovian queueing networks is developed and applied to traffic problems of voice-data integration networks and trunked mobile radio networks. This approach is shown to be computationally much less complex for large systems compared to previous work.

In the integrated networks we consider, two classes of users share the system facilities. If all the servers are busy, the first class of users are queued, but the second class of users are blocked and cleared from the system. The performance objective of such an integrated network is to trade the time delay performance of the first class of users against the blockage performance of the second class of users to keep the grade of service as high as possible for both classes of traffic.

The key-state approach introduced in this thesis is what makes the analysis of the corresponding Markovian queueing network model of these integrated networks computationally less complex than that of previous work. The performance of integrated networks is investigated under several control strategies and new exact closed-form expressions are obtained for the equilibrium probabilities of the corresponding Markovian models. The results are extended to a more general Markovian process where a bulk of arrivals and departures are allowed.

The key-state approach is expected to become a standard tool for analyzing large queueing networks such as will arise when Integrated Services Digital Networks (ISDN) become widely deployed in the next five years.

ANALYSIS OF TRAFFIC PROBLEMS OF INTEGRATED NETWORKS

LIST OF FIGURES AND TABLES

to my mother and father

CHAPTER 1
INTRODUCTION

1.1 Integrated Networks and their importance

Today, telecommunication networks are dedicated to a single telecommunication service. The first dedicated network, telegraphy, was developed 100 years ago to meet the requirements of a specific service where messages were served on a store-and-forward basis using very little switching activity [1]. Later on, telephony was invented, which brought a new dimension to telecommunications service. To be able to transmit human voice from one point to another meant more instantaneous and personal communications. Therefore, the evolution of telephony proceeded differently from that of the telegraph network. As society transformed itself from a labor-intensive to a knowledge-intensive society, new communication services with new requirements had to be developed to satisfy the evolving needs of the information society. A very natural result of this historical development was the proliferation of new dedicated networks.

As far as their technological realization is concerned, the dedicated networks are classified into circuit-switched and packet-switched networks [2], [3]. Circuit switching, which has been used for telephony over the years, is a much older technology than packet-switched technology. The basic difference between a packet- and circuit-switched network is that circuit switching requires the dedication of the actual physical circuit to the users at either end before communication begins. In a packet-switched network, however, data messages or pieces of messages called packets are transmitted from source to destination using a routing algorithm, sharing transmission facilities. Telephone networks were originally developed for handling voice traffic. These networks can therefore be very inefficient when transmitting data. Since the data messages are usually short messages, with long periods of inactivity, it is very costly to dedicate a channel to a single pair of users who do not utilize the transmission channel most of the time. Packet-switched networks have been devised primarily for transmitting data efficiently, in short, bursty pieces. Most commonly, the performance objective of a packet-switched network is to decrease the delay of data packets in the queue waiting for service. The performance objective of the circuit-switched telephone is usually stated in terms of the blocking probability, that is, if the transmission facilities are busy, an arriving call request is blocked and cleared from the system.

In order to limit the proliferation of dedicated networks, new services have been integrated into the existing dedicated networks. In this early form of integration, data messages are transmitted through analog telephone networks by using modems. This is not a full integration for two reasons. First, the required bandwidth for data transmission is larger than the available bandwith of telephone networks. Therefore, the speed of data transmission through the telephone network is limited. Second, the features requested by data customers cannot be satisfied by this technique, because of the restrictions in the switching technology. Today, the advent of digital technology in transmission and switching makes the idea of integrating services more appealing and interesting [4]. Most of the problems of early integration techniques are now being overcome.

The networks of the future will be required to handle a variety of traffic, such as digital voice, video, facsimile, and remote control of information [5], [6]. In order to fulfill such requirements, these future networks are expected to combine circuit and packet switching technologies. Digital telephone networks are being evolved in such a way that the channel transmission bandwidth is now large enough to accommodate the integration of all nonvoice services. On the other hand, packet-switched networks are evolving to handle other types of traffic besides interactive data. It is no longer much of an obstacle to establish new services for integrated networks because of electronic switching. As many of the technological components become available for integrated networks, the concept of having an Integrated Services Digital Networks (ISDN), which is defined as a network evolved from telephony Integrated Digital Network that provides end-to-end digital connectivity to support a wide range of services [7]-[8], is starting to turn into reality.

1.2 Traffic problems of Integrated Networks:

The traffic problems of dedicated networks and their performance analysis have been well studied in the literature [9]-[14]. The fundamental results of Queueing Theory can be used to obtain the performance of dedicated networks. Queueing Theory arises naturally from the type of traffic problems dealt with in a communication network. The concept of server that is used very frequently in Queueing Theory is here the transmission facility in communication networks. The state, which is defined as the number of customers in the system, becomes the number of calls or data packets in the transmission facilities. The performance measure one should use for a dedicated network varies according to the type of traffic served in the system. In a packet-switched network arriving packets are buffered and processed. The time spent in the buffer is then a very good measure of the packet-switched network performance. On the other hand, in the older circuit-switched telephony network, the blocking probability of arriving calls is a significant measure of performance.

The determination of state probabilities is the most important problem of Queueing Theory. The behavior of queueing networks can be understood completely from the state probabilities, assuming memoryless arrivals and service times.

Although existing fundamental results of queueing theory can be applied to most dedicated network problems, these results are no longer valid in the analysis of integrated networks. This is basicaly because in an integrated network, transmission facilities are shared by different types of traffic with different performance objectives. As an example, in a voice-data integration network, the voice traffic, which requires circuit-switched service, is multiplexed with the data traffic, which requires packet-switched service. Therefore, the performance objective of a voice-data integration network is not only to reduce the blocking probability but also to keep the mean waiting time of data packets below a certain level. As a very natural consequence of the sharing process, there must always be a tradeoff between the blockage and time-delay performance of integrated networks.

The analysis of integrated networks depends on the arrival process, the service time distribution, and the service discipline. The idea of service discipline or priority rules arises from different performance objectives. Objectives change from network to network depending upon the purpose and the capacity of the network. For example, the buffer size for the packet-switched traffic is finite because of practical restrictions. If the number of packets in the queue is larger than the buffer size most of the time, then this implies that most of the packet arrivals will be lost. Therefore, in this case the performance objective must be to keep the number of packets in the queue less than a certain threshold at all traffic levels. This problem arises mostly at high traffic intensities. Hence, the service discipline selected to achieve the particular objective described above may not be good at low packet traffic intensities. This shows that the selection of service discipline is strongly dependent on the traffic level. The priorities and sharing protocols are then set according to the prespecified performance objectives.

We note that the arrival processes are assumed to be Poisson (memoryless), and the service time distributions are assumed to be exponential (memoryless) in most traffic theory problems. In fact, numerous but not all real life situation can be adequately modeled by these assumptions [9]-[16].

The necessity of developing a new approach which can be applied to a large class of queueing networks, including integrated networks, is the motivation of this thesis. Our primary interest is to understand the steady-state behavior of integrated networks, which requires the determination of the steady-state probabilities. In Chapter 2, we introduce a new approach for finding closed-form expressions for the steady-state probabilities. This new approach, which we call key-state approach, eliminates the computational problems that arise in previous work. Assuming that the arrival process is Poisson and the service time distribution is exponential, we investigate the performance results for several service disciplines by employing the key-state technique.

1.3 Overview and organization.

Most of the models proposed for integrated networks shared by multiple classes of users lead to Markovian queueing networks [18]-[27]. The determination of the equilibrium probabilities of these Markovian queueing networks requires the solution of a set of linear equations known as global balance equations. In the first section of the Chapter 2, mathematical tools and definitions are introduced which yield the derivation of the global balance equations. The algebraic techniques that have been studied previously and their computational problems are discussed in Section 2.2. Because of these computational problems, in most cases the behavior of the system cannot be expressed transparently by using the previous techniques. We therefore introduce a new approach in Section 2.3 that makes it possible to find closed-form expressions for the equilibrium probabilities.

In Chapter 3, we apply the key-state approach to one of the most prevalent type of integrated network, voice-data integration networks. The integration of voice and data is a very important step in the evolution of the Integrated Services Digital Network concept. Two service disciplines, First-Come, First-Served and Preemptive Movable-Boundary, are investigated. Closed-form expressions are obtained for the voice and data performance measures. The performance results of these service disciplines are compared to each other and to the results for dedicated networks. At the end of Chapter 3, the solution found for the equilibrium probabilities is extended to a more general case.

Chapter 4 introduces another type of integrated network, the Trunked Mobile Radio Network whose transmission facilities are shared by two different types of traffic as in the case of a voice-data integration network. However, here both types of traffic are voice traffic. A sharing algorithm is proposed for the Trunk Mobile Radio originally introduced by Motorola [28]. The performance of the original traffic control strategy is compared with the performance of other control strategies obtained by key-state approach in Chapter 4.

The last two chapters are devoted to possible extentions of the key-state approach to other traffic problems and to conclusions. The results of the key-state approach are generalized to integrated bulk arrival and departure systems. At each Poisson arrival instant, these systems accept a bulk of arrivals, which consist of perhaps more than one customer, chosen from one or both of two classes of users. Future communication networks are expected to handle different services, hence they may be modeled as bulk arrivals and departure systems, for example, information broadcast systems. Finally, the thesis discusses some open reserach problems and concludes with a brief summary.

CHAPTER 2

KEY-STATE APPROACH

2.1 Mathemetical tools of performance analysis

The assumptions made on the arrival and the departure processes make the integrated network models into Markovian queueing networks, where the state transitions probabilities satisfy the Markovian property. That is, the probability of the next state value depends only upon the current state value; the distribution of the transition time between states is memoryless. This property is satisfied only if the inter-arrival and inter-departure time (for a busy server) distribution are exponential.

The consideration of Markovian systems with discrete state space will be central to the performance analysis of integrated queueing networks. Therefore, this section is devoted to the explanation of how we deal with Markovian systems.

We consider a Markovian system with discrete state space where $X(t)$ denotes the state of the system at time t. $X(t)$ will take on values from a discrete set, S. From the definition of Markovian system, $X(t)$ depends upon past history through its current value only. We denote the probability that the system is at state X_k at time t as:

$$P_k(t) = Pr\{X(t) = X_k\}. \tag{2.1}$$

The difference between the rate at which the system enters X_k and the rate at which the system leaves X_k must be equal to the rate of change of flow into X_k, that is, zero in the steady-state. This notion provides us with a set of differential-difference equations for state probabilities. If we concentrate on state X_k, the probability flow rate into X_k is:

$$\underline{\text{Flow rate into } X_k} = \sum_{i=0}^{N} \lambda_{ki} P_i(t).$$

Here λ_{ki} is the transition rate (or arrival rate) from state i to state k, and $N+1$ is the number of states from which there are transitions to X_k. The rate at which the system leaves state X_k is given by:

$$\underline{\text{Flow rate out of } X_k} = P_k(t) \sum_{j=1}^{M} \mu_{jk}.$$

Here μ_{jk} is the transition (or departure rate) from state k to state j, and M is the number of states to which there are transitions from X_k. (Note that for an infinite-state Markovian systems, M, N are not necessarily finite numbers.) Clearly, the difference between these two flows is the effective probability flow rate into X_k, that is,

$$\frac{dP_k(t)}{dt} = \sum_{i=0}^{N} \lambda_{ki} P_i(t) - P_k(t) \sum_{j=1}^{M} \mu_{jk}. \tag{2.2}$$

If the system is ergodic then the limiting probabilites exist and therefore

$$\lim_{t \to \infty} \frac{dP_k(t)}{dt} = 0. \tag{2.3}$$

Assume that the system is stable (steady-state probabilities exist) for a set of transition rates. Then the flow rate into any state must be equal to the flow rate out of that state. Hence, for each state the following equation can be written:

$$\sum_{i=0}^{N} \lambda_{ki} P_i = P_k \sum_{j=0}^{M} \mu_{jk}, \tag{2.4}$$

where

$$P_k = \lim_{t \to \infty} P_k(t). \tag{2.5}$$

We are only interested in the steady-state behavior of the system, in other words, we are interested in the determination of the equilibrium (steady-state) probabilities. The equilibrium probabilities, P_k, are determined from the equations obtained in 2.4. The particular equation given in 2.4 is called the balance equation of state X_k. The equilibrium probabilities, P_k, must satisfy all the balance equations. Therefore the P_k are determined by solving a set of linear equations known as global balance equations. The number of these equations is equal to the number of states in the sytem.

Note that for an infinite-state Markovian system there is an infinite number of equations. These equations can be expressed in matrix form as:

$$Q\ \vec{P} = \vec{0} \tag{2.6}$$

where Q is called the transition-rate matrix and $\vec{P}$ is called the steady-state probability vector. The elements of Q are anologously given by:

$$q_{ij} = \begin{cases} \lambda_{ij}, & \text{if } i \neq j; \\ \sum_{\ell=0}^{N} \mu_{i\ell}, & \text{if } i = j \end{cases} \tag{2.7}$$

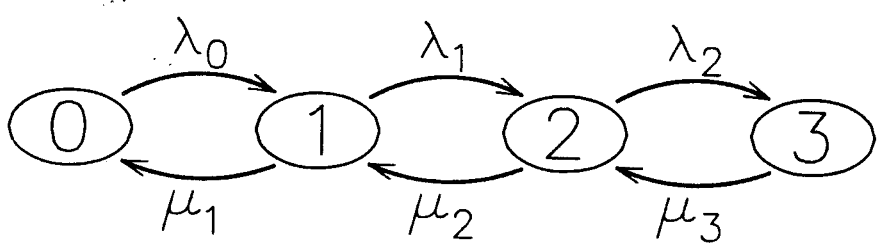
and the column vector P is given by:

$$\vec{P} = [P_0\ P_1\ \ldots]^T. \tag{2.8}$$

An alternative way of displaying the information contained in the Q matrix is by means of the state-transition rate diagram where the states are represented by ovals and the transitions by arrows from state to state. Note that the state-transition rate diagram contains exactly the same information as does the transition-rate matrix Q. An example is shown in Fig. 2.1, for the case of a birth-death process where there are only four states.

The primary interest in the problem of queueing networks is the determination of the equilibrium probabilities. Once the equilibrium probabilities are obtained, the performance results such as mean waiting time in the queue, the blocking probability of arrivals, the required buffer size, etc., can be easily found by employing the basic definitions. In the next section, techniques for finding the equilibrium probabilities are summarized, together with their computational limitations.

a.



b.

$$\lambda_0 P_0 = \mu_1 P_1$$

$$(\lambda_1 + \mu_1) P_1 = \lambda_0 P_0 + \mu_2 P_2$$

$$(\lambda_2 + \mu_2) P_2 = \lambda_1 P_1 + \mu_3 P_3$$

$$\mu_3 P_3 = \lambda_2 P_2$$

c.

$$Q = \begin{pmatrix} -\lambda_0 & \mu_1 & 0 & 0 \\ 0 & -(\lambda_1 + \mu_1) & \mu_2 & 0 \\ 0 & \lambda_1 & -(\lambda_2 + \mu_2) & \mu_3 \\ 0 & 0 & \lambda_2 & -\mu_3 \end{pmatrix}$$

Fig. 2.1 **a.** State transition rate diagram of a finite-state birth-death process.
b. Corresponding local balance equations.
c. Corresponding state-transition rate matrix Q.

2.2 Prior approach and computational problems

The equilibrium probabilities could be obtained directly by solving equation 2.6, but this classical technique is not useful for two reasons. Firstly, when the system size is infinite, it is not computationally feasible to obtain the solution directly from equation 2.6. Second, for large systems the numerical solution doesn't give insight about the system behavior.

Algebraic techniques for finding the equilibrium probabilities have been well studied in the literature [11], [14], [16]. In most of the cases it has been shown that global balance equations are satisfied by a so-called product-form solution. The networks with product-form solution defined below are called product-form networks. Multi-node queueing networks with memoryless arrivals and service times are examples of product-form networks. These networks have a multidimensional state space and global balance equations. The state of such a network is defined as:

$$S = [n_1,\ n_2, \ldots,\ n_M] \tag{2.9}$$

where n_k is the number of users in the k^{th} node, and M is the number of nodes in the network. The joint probability $P(\mathbf{n})$ is:

$$P(\mathbf{n}) = P(n_1,\ n_2, \ldots,\ n_M). \tag{2.10}$$

Surprisingly enough, although there may be interconnections and coupling between the nodes, and even constraints on the total number of customers in service, Jackson first showed that these nodes behave in a sense independently. That is, the probability $P(\mathbf{n})$ is the product-form:

$$P(\mathbf{n}) = \prod_{i=1}^{M} P_i(n_i). \tag{2.11}$$

Here $P_i(n_i)$ is the probability of having n_i users in the system at node i.

The results obtained by Jackson [29] have been generalized, and more general conditions for the existence of product form solution have been found [31]. Recently, a geometric approach was developed to study the existence of product-form solutions

[31]-[33]. This uses the geometric replication of certain building blocks in order to obtain the state-space structure of product-form networks.

In spite of great efforts to generalize the product-form solution to obtain closed-form expressions for equilibrium probabilities, there is still a large class of queueing networks where product-form solutions do not exist. This is the type of network where satisfaction of the partial balance equations does not imply the satisfaction of the global balance equations. Most of the integrated network models are unfortunately of this type.

One of the most frequently used techniques in the study of queueing systems is the moment-generating function, or equivalently the z-transform technique. The moment-generating function of a one-dimensional discrete random variable n such as system size is defined as the expectation of the indeterminate z to the n^{th} power, when the random variable n has probability distribution P_n:

$$G(z) = \sum_{n=0}^{\infty} P_n \, z^n. \tag{2.12}$$

For queueing networks where the balance equations are not readily solved by recursive techniques, especially for infinite-state queueing networks, the moment generating function is a very powerful tool in obtaining the balance equations. The performance parameters can be obtained readily from the derivatives of the moment generating function. For example, the first derivative of $G(z)$ at $z = 1$ is equal to the expected value of n. Higher order statistics of n are obtained by evaluating the higher order derivatives of $G(z)$ at $z = 1$.

In most of the integrated network models involving communication circuits shared by two classes of users, the state space structure can be divided into two disjoint subsets. The reason for attempting this division is that the balance equations in each subset have the same form. That is, there is a difference equation associated with the global balance equations of Subset 1 and there is another difference equation associated with Subset 2. This can be seen as follows. As far as the state transition rates are concerned, the behavior of the system is different when some of the communication channels are free from the behavior when all channels are busy. That is, when there is a free channel, under most service disciplines, an incoming arrival gets service immediately without any restriction. But when all the channels are occupied, the arrivals are either blocked or queued. In the latter case, the channel allocations are made according to some priority rules defined by a control scheme. Naturally, the transition rates enforced by this control scheme will be different from those obtained when the channels are free, in other words, when there is no need for a sharing algorithm. Therefore, usually Subset 2 consists of the states that form after a queue started building up, and Subset 1 covers the remaining states.

Note that the number of states in Subset 2 is infinite, since the buffer size is assumed to be infinite. The number of states in Subset 1 is finite. Previous techniques used generating functions to solve for the distribution of states in Subset 2. On the other hand, the equilibrium probabilities in Subset 1 were found by directly solving the linear equations obtained from the balance equations of each state. This method of solving the queueing network problem is not always satisfactory, because in most of the cases the number of states in Subset 1 is a quadratic function of the number of servers. Hence, the computational complexity increases quadratically. Furthermore, it is a very difficult task to obtain closed-form expressions for the equilibrium probabilities from the moment generating function. Therefore, the behavior of the system can almost never be expressed transparently using this technique.

2.3 A new approach: Key-State

The motivation for developing a new approach in the analysis of queueing networks is to avoid the computational problems mentioned in the previous section and to find closed-form expressions for the equilibrium probabilities. Problems arose from the fact that there is no single recursion on the state probabilities that covers all the state space. As a natural consequence of this observation, the state space was divided into disjoint subsets where there is a different recursion for each subset. The infinite-state subsets are solved by employing moment-generating functions, treating the finite-state subsets as the boundary of the system. As discussed in the previous section, the solution becomes tedious as the size of the subset which contains the boundary states increases.

In this study, we claim that one particular smaller subset of state space captures the essence of whole system. In other words, all other states can be rather easily found by employing the knowledge of the equilibrium probabilities in this subset only. This is a major computational simplification The number of states in this subset turns out to be much smaller than the number of states in the subsets obtained by dividing the state space according to the type of the balance equations. Hence, the number of equations to be solved decreases substantially in many cases. We call the states in this special subset key-states.

In order to make the idea clear, consider a network shared by two classes of users. We denote by $P(i,j)$ the probability of having i first-class and j second-class users in the network. Assume $\{P(0,0),\ P(0,1),\ldots,\ P(0,K)\}$ are the key states. Starting from the balance equations of the key states, it is found possible to relate the equilibrium probabilities of the key states and the neighboring states. Once the neighboring states are related to the key states, the states which do not have direct transitions to the key states are also expressed in terms of the key states by means of the balance equations of the neighboring states. (This process is equivalent to making row operations on the state transition matrix Q). Eventually, we find that the equilibrium probability $P(i,j)$ can be expressed in terms of key-state probabilities:

$$P(i,j) = \sum_{\ell=0}^{K} c_\ell(i,j) P(0,\ell). \tag{2.13}$$

Here $c_\ell(i,j)$ is called the key-state coefficient.

Instead of using the difference equations obtained from the balance equations for the equilibrium probabilities, we find equations relating the key-state coefficients themselves. In this way, we obtain a set of equations with known initial conditions. The computational problems of the previous techniques arose from the fact that the equations obtained for the equilibrium probabilities are valid only in a particular region and the equilibrium probabilities of the states (boundary states) outside this region are unknown. For large networks, these boundary states are very difficult to find. On the other hand, the initial values of key-state coefficients are obtained easily by employing (2.13) and the balance equations. As an example:

$$c_\ell(0,j) = \begin{cases} 1, & \text{if } j = \ell \\ 0, & \text{if } j \neq \ell \end{cases} \tag{2.14}$$

The key-state coefficients can be considered as the weights of the key states. As long as the limiting probabilities exist, that is, the corresponding Markovian model is ergodic, the weighted sum of key states span the whole probability space. That is, there exists at least one solution for the key-state coefficients which represents the equilibrium probabilities of the state space. This can be shown by the following example. Consider the representation given in equation (2.13). The equilibrium probability $P(i,j)$ can be expressed as:

$$P(i,j) = \vec{P}_K^T \, \vec{C}(i,j). \tag{2.15}$$

Here $\vec{P}_K$ is the key-state probability vector

$$\vec{P}_K = [P(0,0),\ P(0,1),\ldots,P(0,K)]^T \tag{2.16}$$

and $\vec{C}(i,j)$ is the key-state coefficients vector

$$\vec{C}(i,j) = [c_0(i,j),\ c_1(i,j),\ldots,c_K(i,j)]^T. \tag{2.17}$$

Provided that the key-state probabilities are greater than 0, the possible solutions for $\vec{C}(i,j)$ are readily found as:

$$\vec{C}(i,j) = [\frac{P(i,j)}{P(0,0)},\ 0,\ 0,\ldots,\ 0]^T \tag{2.18}$$

or

$$\vec{C}(i,j) = [0,\ \frac{P(i,j)}{P(0,1)},\ 0,\ldots,\ 0]^T \tag{2.19}$$

or

$$\vec{C}(i,j) = [\frac{P(i,j)}{2P(0,0)},\ \frac{P(i,j)}{2P(0,1)},\ 0,\ldots,\ 0]^T, \quad \text{etc...} \tag{2.20}$$

Therefore, it is always possible to find solutions for $\vec{C}(i,j)$ in terms of the equilibrium probabilities. Note that the solution is not unique. There can be infinitely many solutions for the key-state coefficients and the existence of equilibrium probabilities is sufficient for the existence of the solution. But since our primary interest is in finding the equilibrium probabilities, the solutions given in (2.18), (2.19) and (2.20) are not useful.

The existence of the key-state coefficients is independent of the selection of the key states. The selection of key states, however, is very important in obtaining useful recursions on the key-state coefficients. For arbitrary selected key states, it may computationally be very hard to find key-state coefficients. The key states must be selected in such a way that the recursions found for the equilibrium probabilities $P(i,j)$ can be easily reflected (transformed) to recursions on the key-state coefficients $c_\ell(i,j)$. Furthermore, this selection is done so as to make it possible to obtain the key-state coefficients of the boundary states recursively, starting from the initial values, in a way which avoids solving a large set of linear equations.

The idea given above will be clarified in Chapter 3, where the key-state approach is applied to voice-data integrated network models.

2.4 Comparison with previous work:

The key-state approach in the analysis of queueing network problems does not require the solution of a set of linear equation to obtain the steady-state probabilities of boundary states. The number of unknowns to be solved in the key-state approach is just equal to the number of key states, at least in most queueing network models. Compared with previous work, this approach cuts the algebraic work to linear in the size of problem from quadratic.

In the previous work, the recursions found for the equilibrium probabilities were attempted to be solved by using the moment-generating function technique. Assuming that the system is stable, the initial conditions were obtained by canceling the poles of the generating function outside the unit circle. The boundary states were solved for by using their balance equations. It is computationaly and physically better to have closed-form expressions for the probabilities themselves than to have merely closed-form expressions for the moment generating function of the equilibrium probabilities.

In this thesis, the recursions found for the key-state coefficients are solved for directly in the time domain. Using the fact that the equilibrium probabilities must approach zero at infinity in a stable queueing system, the key-state probabilities are solved for explicitly. As a result, the behavior of the system can be expressed transparently from the closed-form expressions obtained for the equilibrium probabilities.

The next section will clarify the usefulness of key states approach.

CHAPTER 3

VOICE-DATA INTEGRATION NETWORKS

3.1. A model for Integration Voice-Data:

In this chapter, we will focus on the problem of multiplexing two types of traffic: *voice* and *data*. Voice traffic requires *circuit* switched service, whereas data traffic consists of *packet*-switched traffic. The voice traffic is *blocked* if transmission facilities are not avaliable. On the other hand, the packet-switched traffic is allowed to *queue*. The integration technique investigated here is a TDM (Time-Division Multiplex) scheme in which the TDM frame is partitioned into equal-length time slots. A time slot is assigned for every incoming request if there is a slot available. In general, the number of time slots assigned for different users varies according to their bandwith. Here, one time slot is assumed to be enough for each type of user. Once the assignment is made, users keep their assignments from frame to frame as long as required. If there are N time slots in a frame, this means that there are N channels available for transmission. We assume that the frame length is small compared with the voice and data holding times in the system and therefore ignore the discrete nature of the frame structure. This allows us to model the system in continous time. In the literature it is shown that a continous-time model is a good approximation in the analysis of voice and data integration systems [34]-[35].

Fig. 3.1 represents this integrated multiplexer. Here, one circuit-switched type of input is multipled with packets from a queue onto a TDM link. A more general pictorial representation which multiplexes a number of circuit-switched inputs with packet-switched traffic is given in [3, Fig. 12-16].

3.2. Integration with First-Come First-Served Discipline:

Using the structure given in Fig. 3.1, we will initially investigate the First-Come, First-Served (FCFS) combining strategy, where the slot assignments are made according to first-come, first-served priority. In the next section, FCFS scheme is analyzed by using the key states approach. Closed-form expressions are found for the equilibrium probabilities and the performance measures.

3. 2. 1. Analysis:

Assuming the arrival processes are Poisson (memoryless) and the service time distributions are exponential (memoryless), let the voice arrival rate be λ_2 and the data arrival rate be λ_1. The departure rates are μ_2 and μ_1 for voice and data respectively. If all the servers are busy then the voice arrivals are blocked, but data arrivals are queued.

We define the states as:

$$S_{dv} = [d, v]; \quad 0 \leq d, \quad 0 \leq v \leq N \tag{3.1}$$

where d is the number of data packets and v is the number of voice calls in the system. The state-transition rate diagram of this system is given in Fig. 3.2.

The key states in this case are found to be $\{S_{0i}, i = 0, \ldots, N-1\}$. These are the first row states except $S_{0,N}$. Starting from S_{00}, each key state is connected to other key states and one of the second row states by means of its equilibrium equation. Hence, the second row states can be expressed in terms of only the key state probabilities. Carrying on with the same method, the third row states are connected to second row states. Iteratively, the ℓ^{th} row states are connected to $(\ell-1)^{th}$ and $(\ell-2)^{th}$ row states, and so forth. As a result, for the steady-state probability $P_v(d)$, we obtain the following representation:

$$P_v(d) = \vec{C}_v(d)^T \ \vec{K} \tag{3.2}$$

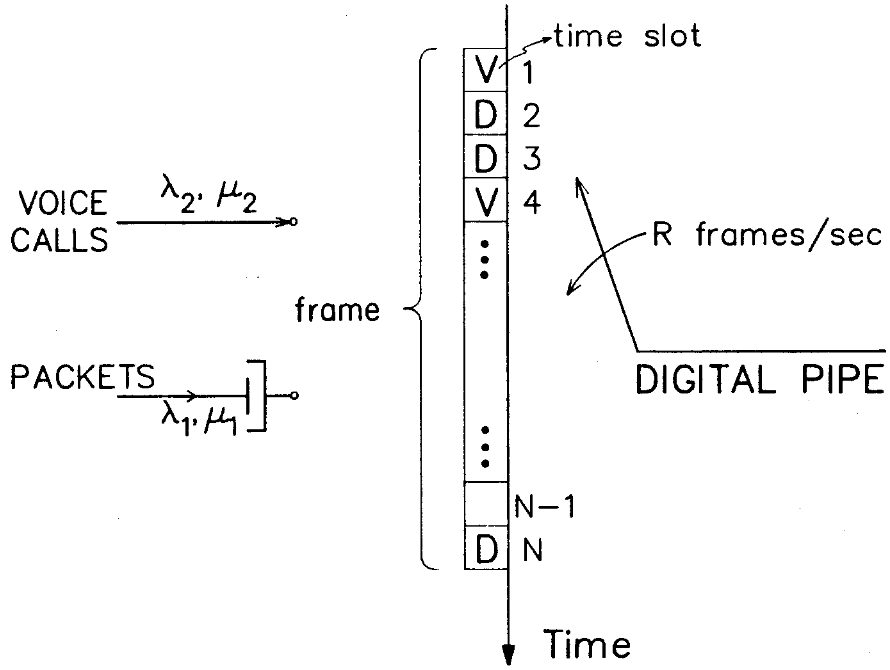


Fig. 3.1 TDM frame structure of an integrated channel.

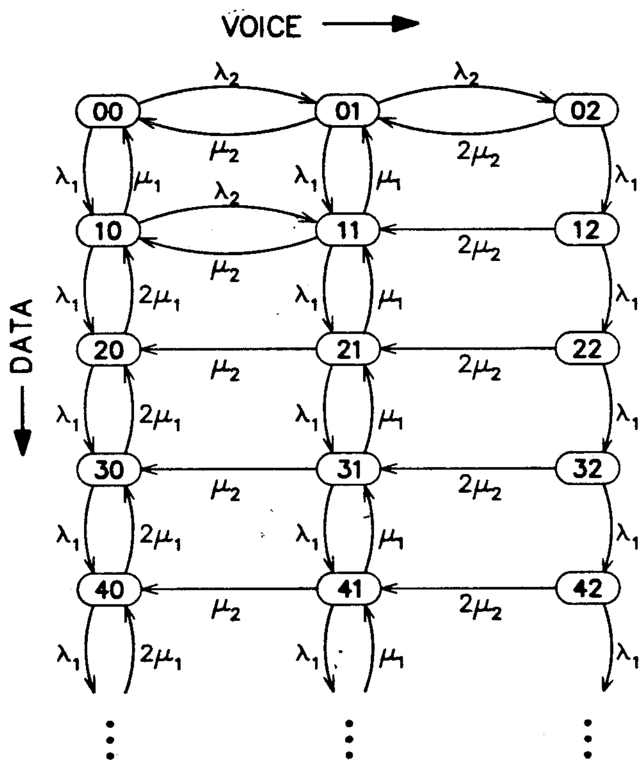


Fig. 3.2 State transition-rate diagram of FCFS scheme for $N = 2$.

where $P_v(d)$ is the probability of having d data packets and v voice calls in the system and $\vec{K}$ is the key state probability vector:

$$\vec{K} = [P_0(0) \;\; P_1(0) ... P_{N-1}(0)]^T. \tag{3.3}$$

The vector $\vec{C}_v(d)$ is called the coefficient vector and it is given as

$$\vec{C}_v(d) = [c_{v,0}(d) \;\; c_{v,1}(d) \;\; \ldots c_{v,N-1}(d)]^T \tag{3.4}$$

where $c_{v,\ell}(d)$ is the coefficient of ℓ^{th} key-state probability in the expression for $P_v(d)$. That is

$$P_v(d) = \sum_{\ell=0}^{N-1} c_{v,\ell}(d) P_\ell(0). \tag{3.5}$$

If we expand each state in a column of the state-transition rate diagram, we obtain a difference equation for the steady-state probabilities of the states located in this column. Here, expanding a state means connecting this state to its neighboring states by its equilibrium equation. As it is seen from the transition-rate diagram, the only flow to the m^{th} column states are from $(m+1)^{th}$ column states after the initial transition states, in other words, after the $(N-m)^{th}$ row.

But the last column states $P_N(d)$ do not require the knowledge of any column. By means of the balance equations obtained for the last column, we can get the following difference equation for these $P_N(d)$'s:

- $$P_N(d) = \frac{\lambda_1}{\lambda_1 + N\mu_2} P_N(d-1), \quad d \geq 0, \tag{3.6}$$

which yields:

$$P_N(d) = \frac{\lambda_2}{\lambda_1 + N\mu_2}\left(\frac{\lambda_1}{\lambda_1 + N\mu_2}\right)^d, \quad d \geq 0. \tag{3.7}$$

Also, by employing (3.2), we obtain the coefficient vector of $P_N(d)$ as:

$$\vec{C}_N(d) = \frac{\lambda_2}{\lambda_1 + N\mu_2}\left(\frac{\lambda_1}{\lambda_1 + N\mu_2}\right)^d \vec{e}_N, \quad d \geq 0. \tag{3.8}$$

Here, $\vec{e}_N$ is an N x 1 vector with all its entries zero except the N^{th} entry which is equal to 1. From (3.8) and (3.4), the entries of the coefficient vector, $c_{N,i}(d)$, can be solved for as:

$$c_{N,\ell}(d) = \begin{cases} \frac{\lambda_2}{\lambda_1+N\mu_2}\left(\frac{\lambda_1}{\lambda_1+N\mu_2}\right)^d & \text{if } i = N; \\ 0 & \text{if } i \neq N. \end{cases} \tag{3.9}$$

Equation (3.9) can also be expressed as

$$c_{N,\ell}(d) = \begin{cases} \beta_0 z_0^{-d} & \text{if } i = N; \\ 0 & \text{otherwise}, \end{cases} \tag{3.10}$$

where

$$\beta_0 = \frac{\lambda_2}{\lambda_1 + N\mu_2}; \quad z_0^{-1} = \frac{\lambda_1}{\lambda_1 + N\mu_2}. \tag{3.11}$$

The balance equations of the $(N-m)^{th}$ column states $S = [N-m, d-1]$ for $d \geq m+2$ yield the following difference equation:

- $$P_{N-m}(d) + a_m P_{N-m}(d-1) + b_m P_{N-m}(d-2) = d_m P_{N-m+1}(d-1), \quad d \geq m+2 \tag{3.12}$$

where

$$a_m = -\frac{\lambda_1 + m\mu_1 + (N-m)\mu_2}{m\mu_1}; \quad b_m = \frac{\lambda_1}{m\mu_1}; \quad d_m = -\frac{(N-m+1)\mu_2}{m\mu_1}. \tag{3.13}$$

It is not possible to obtain the steady-state probabilities directly from equation (3.12) because the difference equation is valid only for $d \geq m+2$, and the initial conditions are unknown.

Fischer and Bhat [26] have carried out the analysis of this same problem using moment generating functions. By canceling the unstable roots inside the unit circle, they obtain an expression for the z-transform of the equilibrium probabilities in terms of the unknown steady-state probabilities for $d \leq m$. They propose to find these unknown probabilities by solving the corresponding local balance equations. But the number of unknowns to be solved for is $N(N+1)/2$, which implies that for large systems the solution is unfeasible. In order to avoid the computational problems, they take $\mu_2/\mu_1 = 1$ as an approximation and obtain some closed-form expressions for mean packet delay and the voice blocking probability. Here, by using the key-state approach, we will find that it is possible to obtain a closed-form solution for the equilibrium probabilities without making this approximation and by solving only N linear equations. The exact results obtained here are compared with Bhat and Fisher's approximate result in the last section.

Equation (3.12) can be employed to get difference equations for the entries of the coefficient vector, the $c_{N-m}(d)$'s. The initial conditions of these new difference equations can be obtained from the balance equations of $P_{N-m}(d)$'s for $d \leq m+2$ by decomposing the $P_{N-m}(d)$'s into $\{P_0(0), P_1(0), \ldots, P_{N-1}(0)\}$. Substituting (3.5) into (3.12), we obtain:

$$
\begin{aligned}
&[c_{N-m,0}(d) + a_m c_{N-m,0}(d-1) + b_m c_{N-m,0}(d-2)]P_0(0)+ \\
&[c_{N-m,1}(d) + a_m c_{N-m,1}(d-1) + b_m c_{N-m,1}(d-2)]P_1(0)+ \\
&\quad\vdots \\
&[c_{N-m,N-1}(d) + a_m c_{N-m,N-1}(d-1) + b_m c_{N-m,N-1}(d-2)]P_{N-1}(0) = \\
&\qquad d_m c_{N-m+1,0}(d-1)P_0(0)+ \\
&\qquad d_m c_{N-m+1,1}(d-1)P_1(0)+ \\
&\qquad\vdots \\
&\qquad d_m c_{N-m+1,N-1}(d-1)P_{N-1}(0)
\end{aligned}
\tag{3.14}
$$

for $d \geq m+2, \quad 1 \leq m \leq N$.

Equation (3.14) is satisfied if:

$$
c_{N-m,\ell}(d) + a_m c_{N-m,\ell}(d-1) + b_m c_{N-m,\ell}(d-2) = d_m c_{N-m+1,\ell}(d-1), \tag{3.15}
$$

$d \geq m+2, \quad \ell = \{0, 1, \ldots, N-1\}$.

This set of difference equations can be solved iteratively for $\ell = 0, 1, ..., N-1$ starting from $m = 1$. The first equation to be solved is:

$$c_{N-1,\ell}(d) + a_{N-1}c_{N-1,\ell}(d-1) + b_{N-1}c_{N-1,\ell}(d-2) = d_{N-1}c_{N,\ell}(d-1), \quad d \geq 3 \quad (3.16)$$

where $c_{N,\ell}(d)$ is given in (3.10). The solution has the following form:

$$c_{N-m,\ell}(d) = A_{N-m,2m,\ell}z_{2m}^d + A_{N-m,2m-1,\ell}z_{2m-1}^d + \sum_{i=0}^{2m-2} A_{N-m,i,\ell}z_i^d, \quad d \geq m+2 \quad (3.17)$$

Here the first two terms are the complete solution of the homogenous equation, and the rest form the particular solution of the non-homogenous equation. The two roots of the characteristic equation of the difference equation, the z_i's, are given by:

$$z_{2i-1} = \frac{-a_i + \sqrt{a_i^2 - 4b_i}}{2}; \quad z_{2i} = \frac{-a_i - \sqrt{a_i^2 - 4b_i}}{2}, \quad 1 \leq i \leq m \quad (3.18)$$

which we will need later.

We can use (3.17) to rewrite (3.15) as:

$$c_{N-m,\ell}(d) + a_m c_{N-m,\ell}(d-1) + b_m c_{N-m,\ell}(d-2) = d_m \sum_{i=0}^{2m-2} A_{N-m+1,i,\ell}z_i^{d-1} \quad (3.19)$$

for $d \geq m+2$. By substituting the particular solution found (3.17) into (3.19), we obtain a recursive relation for the coefficients of the particular solution found in (3.17):

$$A_{N-m,i,\ell} = \frac{A_{N-m+1,i,\ell}\ d_m\ z_i^{-1}}{1 + a_m\ z_i^{-1}\ + b_m\ z_i^{-2}}, \qquad i = 0, 1, ..., 2m-2. \tag{3.20}$$

The remaining coefficients, $A_{N-m,2m,\ell}$ and $A_{N-m,2m-1,\ell}$, are found by making use of the initial conditions at $d = m$ and $d = m+1$, and the solution is thus completed:

$$A_{N-m,2m,\ell} = \frac{F_{N-m,\ell} - (z_{2m-1} + a_m)E_{N-m,\ell}}{(z_{2m} - z_{2m-1})\ z_{2m}^{m+2}}, \tag{3.21}$$

$$A_{N-m,2m-1,\ell} = \frac{F_{N-m,\ell} - (z_{2m} + a_m)E_{N-m,\ell}}{(z_{2m-1} - z_{2m})\ z_{2m-1}^{m+2}}, \tag{3.22}$$

where

$$E_{N-m,\ell} = \sum_{i=0}^{2m-2} A_{N-m,i,\ell}(a_m\ z_i + b_m)\ z_i^m - I_{N-m,\ell}, \tag{3.23}$$

$$F_{N-m,\ell} = b_m \sum_{i=0}^{2m-2} A_{N-m,i,\ell} z_i^{m+1} - J_{N-m,\ell}, \tag{3.24}$$

$$I_{N-m,\ell} = a_m c_{N-m,\ell}(m+1) + b_m c_{N-m,\ell}(m), \tag{3.25}$$

$$J_{N-m,\ell} = b_m c_{N-m,\ell}(m+1), \tag{3.26}$$

where $m = 1, 2, ..., N$.

From (3.5) and (3.17) the steady-state probabilities are now expressed as follows:

$$P_{N-m}(d) = \sum_{\ell=0}^{N-1} \sum_{i=0}^{2m} A_{N-m,i,\ell}\, z_i^d\, P_\ell(0), \quad d \geq m+2. \tag{3.27}$$

In Appendix A, the roots z_i which are given in (3.18) are found to be real and positive and it is proved that they satisfy the following relations:

$$z_{2i-1} > 1, \quad z_{2i} < 1, \qquad i = 0,1,2,\ldots N-1, \tag{3.28}$$

$$z_{2N} = \frac{\lambda_1}{N\mu_1}, \quad z_{2N-1} = 1. \tag{3.29}$$

At the end of this section, we show that the system is ergodic when $\lambda_1 < N\mu_1$. If the system is ergodic, (3.27) represents a probability distribution. This means that:

$$\sum_{m=0}^{N} \sum_{d=0}^{\infty} \frac{P_{N-m}(d)}{P_0(0)} < \infty. \tag{3.30}$$

In a stable ergodic system, the empty state probability can be found by normalizing the obtained solution. The normalization condition is:

$$\sum_{m=0}^{N} \sum_{d=0}^{\infty} P_{N-m}(d) = 1. \tag{3.31}$$

But from (3.27) we observe that the terms involving z_i^d diverge as d goes to infinity for odd i. This seems to contradict the stability requirement given above, because (3.27) must be convergent.

However, what it really means is that for a stable system, the coefficients of the exponents with magnitude greater than 1 must be zero. In other words:

$$\sum_{\ell=0}^{N-1} A_{N-m,2j-1,\ell}\, P_\ell(0) = 0, \qquad 1 \le m \le N, \quad 1 \le j \le m. \tag{3.32}$$

That is, in a stable system the key state probabilities must satisfy the above relations. Most of the equations in (3.32) are redundant and they are all included in the following set of N equations:

$$\sum_{\ell=0}^{N-1} A_{N-m,2m-1,\ell}\, P_\ell(0) = 0, \qquad 1 \le m \le N. \tag{3.33}$$

In other words;

$$\sum_{\ell=0}^{N-1} A_{N-m,2m-1,\ell}\, P_\ell(0) = 0 \Rightarrow \sum_{\ell=0}^{N-1} A_{N-j,2m-1,\ell}\, P_\ell(0) = 0, \tag{3.34}$$

$m \le j \le N, \quad 1 \le m \le N.$

Equation (3.34) is proved in Appendix B.

We can now conclude that, in a stable voice and data integration system, the key state probabilities must satisfy the following equations:

$$\begin{pmatrix} A_{N-1,1,0} & A_{N-1,1,1} \cdots & A_{N-1,1,N-1} \\ A_{N-2,3,0} & A_{N-2,3,1} \cdots & A_{N-2,3,N-1} \\ \vdots & \vdots & \vdots \\ A_{0,2N-1,0} & A_{0,2N-1,1} \cdots & A_{0,2N-1,N-1} \end{pmatrix} \begin{pmatrix} P_0(0) \\ P_1(0) \\ \vdots \\ P_{N-1}(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3.35}$$

In this equation, there are N equations for N key state probabilities. If all the rows were linearly independent, then the only solution would be the trivial solution, and the stability condition would fail. But we know that there exists a set of system parameters (arrival rates, departure rates) which makes this system stable, because in the limiting case when $\lambda_2 = 0$, this system is merely an M/M/N queueing system. Therefore, the rank of the above matrix must be less than N. So as to have a nontrivial solution for the key state probabilities; at least one of the above equations can be obtained from the others.

It will now be shown that the last equation is in fact redundant. By employing the first $N-1$ equations in (3.35), it is possible to express all key-state probabilities in terms of empty-state probability $P_0(0)$, i. e. we can find a set of positive numbers, ρ_ℓ, such that:

$$P_\ell(0) = \rho_\ell P_0(0), \qquad 0 \leq \ell \leq N-1. \tag{3.36}$$

Therefore, the equations given in (3.33) can be rewritten as:

$$P_0(0) \sum_{\ell=0}^{N-1} \rho_\ell \, A_{N-m,2m-1,\ell} = 0. \tag{3.37}$$

The solution for the first column steady-state probabilities is given by (3.27) as:

$$P_0(d) = \sum_{\ell=0}^{N-1} \left(A_{0,2N,\ell} \, z_{2N}^d + A_{0,2N-1,\ell} \, z_{2N-1}^d + \sum_{i=0}^{2N-2} A_{0,i,\ell} \, z_i^d \right) P_\ell(0). \tag{3.38}$$

Substituting the values of z_{2N}, z_{2N-1} into (3.38) and employing equation (3.36) and (3.37), we obtain the following expression for $P_0(0)$:

$$P_0(d) = P_0(0) \sum_{\ell=0}^{N-1} \rho_\ell \left(A_{0,2N,\ell} \left(\frac{\lambda_1}{N\mu_1}\right)^d + A_{0,2N-1,\ell} + \sum_{i=0}^{N-1} A_{0,2i,\ell}\, z_{2i}^d \right). \qquad (3.39)$$

Here the terms with z_{2i-1}^d have disappeared. Except for the first two terms, the remaining terms contain exponents of terms whose magnitudes are less than 1, so they indeed approch zero as d goes to infinity. This is also true for the second term unless $\lambda_1/N\mu_1 \geq 1$. So, assuming $\lambda_1/N\mu_1 < 1$, the following limiting expression is obtained for $P_0(d)$:

$$\lim_{d\to\infty} P_0(d) = P_0(0) \sum_{\ell=0}^{N-1} A_{0,2N-1,\ell}\, \rho_\ell. \qquad (3.40)$$

The limit of $P_0(d)$ found in (3.40) must be zero, otherwise the difference equation given in (3.12) for $P_{N-m}(d)$ is not satisfied by (3.27) for $m = N$. This implies that either $P_0(0)$ or the summation term is zero. $P_0(0)$ can not be zero, because a Markovian system is ergodic if and only if $P_0(0) > 0$ ([17], Chap. VIII, Sec. 7). Therefore

$$\sum_{\ell=0}^{N-1} P_\ell(0)\, A_{N-m,2m-1,\ell} = 0 \quad 1 \leq m \leq N-1 \Rightarrow \sum_{\ell=0}^{N-1} P_\ell(0)\, A_{0,2N-1,\ell} = 0. \qquad (3.41)$$

Hence, it is concluded that the first $N-1$ equations in (3.35) indeed imply the last one and (3.35) does have a nontrivial solution.

As a summary: the steady-state probabilities of the ergodic voice and data integration network have been found in terms of the key state probabilities in (3.27). The key state probabilities are solved for in terms of $P_0(0)$ in (3.35). The final expression for $P_{N-m}(d)$ as a function of $P_0(0)$ is given as:

$$P_{N-m}(d) = P_0(0) \sum_{\ell=0}^{N-1} \sum_{i=0}^{m} \rho_\ell \, A_{N-m,2i,\ell} \, z_{2i}^d, \quad d \geq m+2, \tag{3.42}$$

where z_{2i} is given in (3.18). The solution given in (3.42) will be completed by expressing $P_0(0)$ explicitly by employing the normalization condition that probabilities add to 1. Defining K_0 as:

$$K_0 = \sum_{d=0}^{m+1} \sum_{m=0}^{N} \sum_{\ell=0}^{N-1} c_{N-m,\ell}(d) \; \rho_\ell \tag{3.43}$$

we have

$$\sum_{d=0}^{\infty} \sum_{m=0}^{N} P_{N-m}(d) = 1 \Rightarrow$$

$$P_0(0) \left(K_0 + \sum_{d=m+2}^{\infty} \sum_{\ell=0}^{N-1} \sum_{m=0}^{N} \sum_{j=0}^{m} \rho_\ell \, A_{N-m,2j,\ell} \, z_{2j}^d \right) = 1. \tag{3.44}$$

Therefore, the empty state probability $P_0(0)$ is:

$$P_0(0) = \left(K_0 + \sum_{d=m+2}^{\infty} \sum_{\ell=0}^{N-1} \sum_{m=0}^{N} \sum_{j=0}^{m} \rho_\ell \, A_{N-m,2j,\ell} \, z_{2j}^d \right)^{-1}. \tag{3.45}$$

Or,

$$P_0(0) = \left(K_0 + \sum_{\ell=0}^{N-1} \sum_{m=0}^{N-1} \sum_{j=0}^{m} \rho_\ell \, A_{N-m,2j,\ell} \left(\frac{z_{2j}^{m+2}}{1 - z_{2j}} \right) \right)^{-1} \tag{3.46}$$

As is seen from (3.43), K_0 depends on the initial conditions $c_{N-m,\ell}(d)$ for $d = 0, ..., m+1$. These initial conditions are obtained by decomposing the $P_{N-m}(d)$'s in terms of key state probabilities starting from $d = 0$ to $d = m + 1$. For example:

$$c_{N-m,\ell}(0) = \begin{cases} 1 & \text{if } \ell = N - m \\ 0 & \text{if } \ell \neq N - m \end{cases} ; \tag{3.47}$$

$$c_{0,0}(1) = \frac{\lambda_1 + \lambda_2}{\mu_1}, \qquad c_{0,1}(1) = \frac{\mu_2}{\mu_1}, \ldots \tag{3.48}$$

Assuming that the system is ergodic, the steady-state probabilities were found in (3.42). The solution given in (3.42) represents a probability distribution if and only if $z_{2N} < 1$. Therefore, the necessary and sufficient condition for the stability of this system is:

$$\lambda_1 < N\mu_1. \tag{3.49}$$

It is no coincidence that (3.49) is also necessary and sufficient for the stability of the corresponding M/M/N queueing system (no voice traffic), because incoming voice requests are blocked unless the data queue is empty. Therefore, at high data traffic intensities, a voice-data integration network behaves like a data-only system, which in turn behaves like an M/M/N queueing system.

3.2 2. Performance Results:

In the expression found for the equilibrium probabilities in (3.42), the coefficient of z_{2i}^d is given by:

$$B_{N-m,2i} = P_0(0) \sum_{\ell=0}^{N} \rho_\ell \, A_{N-m,2i,\ell}. \tag{3.50}$$

Substituting (3.50) into (3.42), the equilibrium probabilities are obtained as follows:

$$P_{N-m}(d) = \sum_{i=0}^{m} B_{N-m,2i} \, z_{2i}^d, \quad d \geq m+2. \tag{3.51}$$

The closed-form expressions for the mean packet delay and the standard deviation of queue size can be obtained by employing (3.51), and we will now find the expressions.

3.2.2a. Mean Delay:

The mean packet delay, T, is obtained from the mean number, E[L], of packets in the system by employing Little's formula ([12], Chap. 4, pp. 140-148):

$$T = \frac{E[L]}{\lambda_1}. \tag{3.52}$$

The mean packet size in the system is given by:

$$E[L] = \sum_{m=0}^{N} \sum_{d=0}^{\infty} d \, P_{N-m}(d) = \sum_{m=0}^{N} \sum_{d=0}^{m+1} d P_{N-m}(d) + \sum_{m=0}^{N} \sum_{d=m+2}^{\infty} d \, P_{N-m}(d). \tag{3.53}$$

We define the first summation term as L_1:

$$\begin{aligned} L_1 &= \sum_{m=0}^{N} \sum_{d=0}^{m+1} d \, P_{N-m}(d) \\ &= P_0(0) \sum_{m=0}^{N} \sum_{d=0}^{m+1} \sum_{\ell=0}^{N-1} d \, c_{N-m,\ell}(d) \, \rho_\ell. \end{aligned} \tag{3.54}$$

Here L_1 depends on the initial equilibrium probabilities and is obtained by using the coefficients $c_{N-m,\ell}(d)$ for $d \leq m+1$. The second term in (3.54) can be written as:

$$\sum_{d=m+2}^{\infty} d \; P_{N-m}(d) = \sum_{i=0}^{m} B_{N-m,2i} \frac{z_{2i}^{m+2}[m+2-z_{2i}(m+1)]}{(1-z_{2i})^2}. \tag{3.55}$$

Hence the normalized delay, $\mu_1 T$, the time delay relative to the transmission time, is given by:

$$\mu_1 T = \frac{L_1}{a_1} + \frac{P_0(0)}{a_1} \sum_{m=0}^{N} \sum_{i=0}^{m} B_{N-m,2i} \frac{z_{2i}^{m+2}[m+2-z_{2i}(m+1)]}{(1-z_{2i})^2} \tag{3.56}$$

where $a_1 = \lambda_1/\mu_1$ is the data traffic. The normalized queueing time W is defined as the time spend on the queue relative to the transmission time and clearly is

$$W = \mu_1 T - 1.$$

Subtracting 1 accounts for the transmission time.

Figs. 3.3 and 3.4 show the normalized queueing time as a function of data and voice traffic respectively for different values of $\alpha = \mu_1/\mu_2$ when the voice traffic is fixed at 5 Erlangs. As expected, the mean packet queueing time increases as the data traffic, a_d, increases. Since the number of servers here is 10, the data traffic must be less than 10. Otherwise, the data queue blows up, as can be observed from Fig. 3.4. The value of W is always greater for smaller voice departure rates (larger α values). This is because if the holding time of voice calls is large, then they occupy the servers for a long time and arriving data packets cannot get service immediately. This, of course, causes an increase in the queue length.

A closed-form expression is obtained for the mean queue length in [26], assuming that $\alpha = 1$. The expression in [26] is therefore exact for $\mu_1 = \mu_2$ and is a special case of equation (3.56). We will investigate the dependency of mean delay to α and to check the accuracy of the approximate expression found in [26] for $\alpha > 1$. For this purpose, we plot normalized mean delay for three different values of α in Figs. 3.3 and 3.4. The first and second plots are obtained from (3.56) by taking $\alpha = 50$ and $\alpha = 10$. The third plot correponds to $\alpha = 1$, and it can either be obtained from the expression given in [26] or from (3.56) by taking $\alpha = 1$.

As observed in Figs. 3.3 and 3.4, at large traffic intensities the mean delay is strongly dependent on α, assuming $\alpha = 1$ does not give good results. At low traffic intensities, however, the mean queue length has weaker dependency on α. We see that here the approximation gives more accurate results, as expected.

When α is small, that is, the voice departure rate is large, the integrated system behaves like a data-only system. In this case, voice calls stay in the system for a short time and therefore it is very unlikely to find the transmission capacity occupied by voice calls; the system is open for data packets most of the time. As a result, the system accomodates intensive voice traffic without degrading data traffic performance. This can be observed in Fig. 3.4, where increasing voice traffic does not affect normalized mean packet delay significantly. This means that for small values of α, the interconnected system has the same time-delay performance as a data-only system. From Fig. 3.4, when there is no voice traffic and the data traffic is 5 Erlangs, the normalized delay is 1 or the normalized queueing time is 0, because it is not likely that an arriving packet waits before service, and so the time spent in

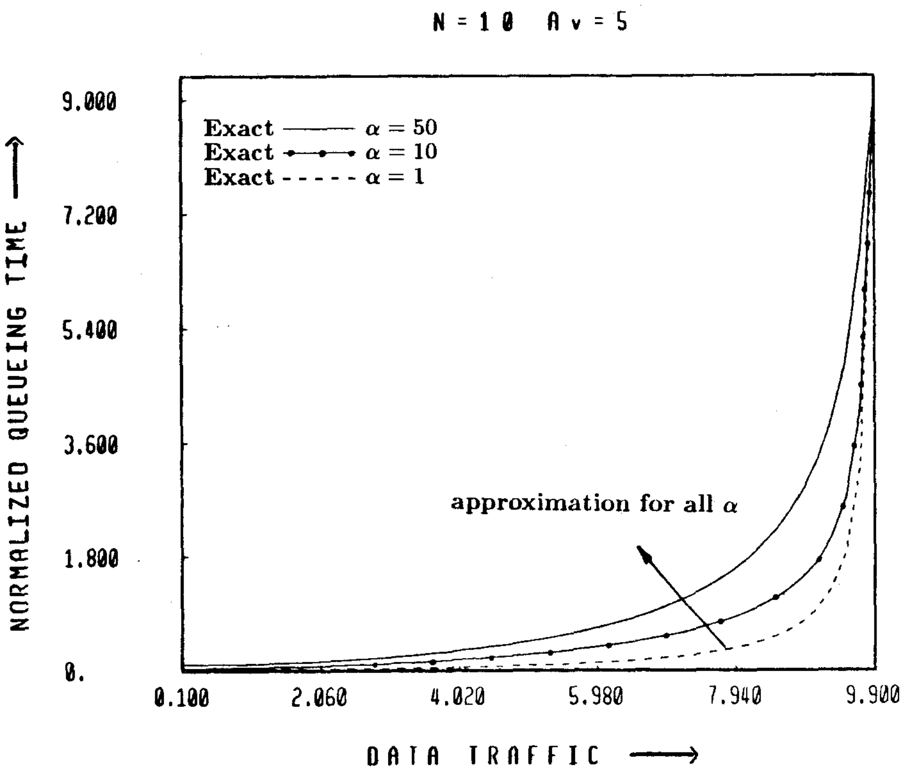


Fig. 3.3 Comparison of the normalized queueing time obtained by using exact analysis with Bhat and Fisher's approximation for different ratios of voice and data holding times ($\alpha = \mu_1/\mu_2$) and as a function of data traffic with $a_v = 5$ Erlangs.

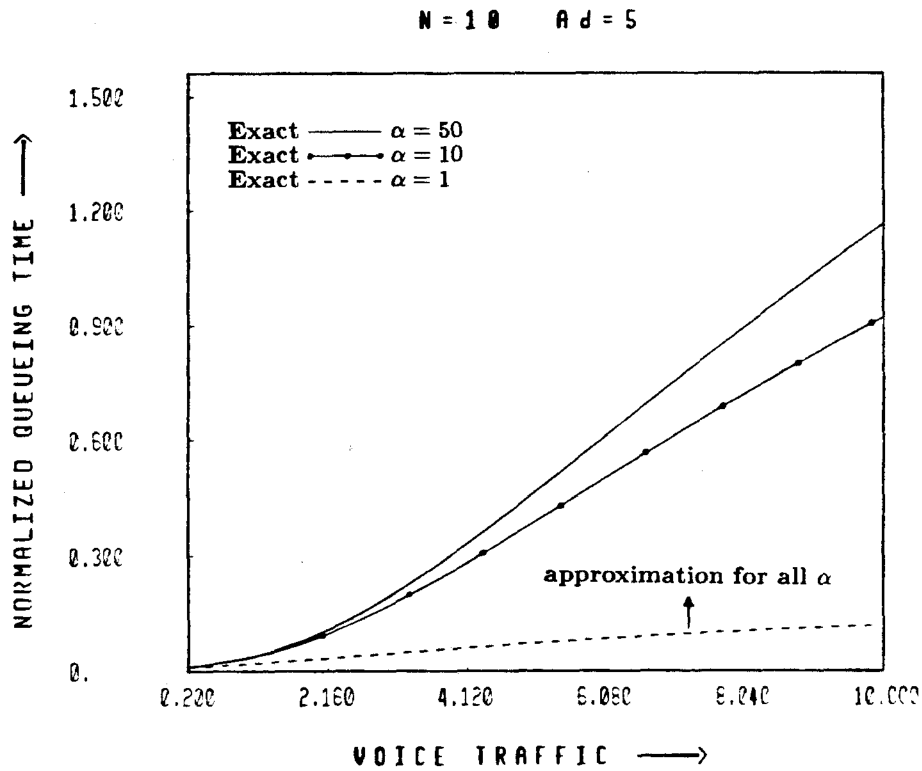


Fig. 3.4 Comparison of the normalized queueing time obtained by using exact analysis with Bhat and Fisher's approximation for different ratios of voice and data holding times ($\alpha = \mu_1/\mu_2$) and as a function of voice traffic with $a_d = 5$ Erlangs.

the system is equal to the service time. When the voice traffic is 9 Erlangs, $\alpha = 1$, the normalized packet waiting time is 0.12. That is, in order to accomodate 9 Erlangs of voice traffic together with data traffic, the priced paid is just to wait in the queue 12 percent of the mean packet service time. Therefore, compared to the analogous data-only system, the integrated system is superior since it also accommodates a second type of traffic without degrading the performance of the first type of traffic substantially.

3.2.2b. Deviation of packet size:

The mean value analysis gives us no information about the fluctuations of data and voice about their mean values. In practice, there is a very high correlation between voice messages from frame to frame, because voice messages keep their slot assignments from frame to frame as long as required. Therefore, the mean values obtained here for the queue length do not give too much insight about the system behavior. Voice calls typically last 3 minutes on the average, but data messages are much shorter (in the order of milliseconds). In case voice calls occupy all the transmission capacity, the system is closed for data traffic. Even if this doesn't happen frequently and lasts only a few seconds when it does occur, thousands of data packets may arrive to the system, and queue peaks will occur. The second order statistics of queue length, i.e., the deviations of queue length around its mean, give a better idea about system performance.

The second moment of packet size in the system is found as:

$$E[L^2] = \sum_{m=0}^{N} \sum_{d=0}^{\infty} d^2 \; P_{N-m}(d) \tag{3.57}$$

$$= \sum_{m=0}^{N} \sum_{d=0}^{m+1} d^2 \; P_{N-m}(d) + \sum_{m=0}^{N} \sum_{d=m+2}^{\infty} d^2 \; P_{N-m}(d) \tag{3.58}$$

The first term is defined as K_1, and it is found by employing the key state coefficients of initial probabilities as in (3.54):

$$K_1 = \sum_{m=0}^{N} \sum_{d=0}^{m+1} d^2 \; P_{N-m}(d)$$
$$= P_0(0) \sum_{m=0}^{N} \sum_{d=0}^{m+1} \sum_{\ell=0}^{N-1} d^2 \; c_{N-m,\ell}(d)\rho_\ell. \tag{3.59}$$

The second term can be rewritten by employing the closed-form expression obtained for the equilibrium probabilities:

$$\sum_{m=0}^{N} \sum_{d=m+2}^{\infty} d^2 \; P_{N-m}(d) = \sum_{m=0}^{N} \sum_{i=0}^{m} B_{N-m,2i} \frac{z_{2i}^{m+2}(qz_{2i}^2 + z_{2i} + r)}{(1 - z_{2i})^3}. \tag{3.60}$$

where $q = m(m+2)$ and $r = (m+2)^2 + 1$. The standard deviation of packet size in the system can be found from the mean packet size and the second moment of L as follows:

$$\sigma_L = \sqrt{E[L^2] - E^2[L]}; \tag{3.61}$$

it does not further simplify. For plotting, we define the normalized standard deviation as:

$$s_L = \frac{\sigma_L}{E[L]} \tag{3.62}$$

Fig. 3.5 and Fig. 3.6 show the normalized standard deviation of packet size in the system as a function of data and voice traffic respectively. The plots are obtained by employing equation (3.62) for different values of α. For high traffic intensities and for high values of α, s_L is much larger than 1. This means that the standard deviation of packet size is much larger than its mean. The result is not surprising, because, as explained previously, when α is large it is likely for the transmission capacity to be occupied by voice calls, and this causes the formation of queue peaks. As soon as voice calls leave the system, data packets utilize the whole transmission capacity to empty the buffers. Therefore, large variations will occur in the number of packets in the system.

3.2.2c. Blocking probability of voice calls:

The voice arrivals are blocked when all the time slots are occupied. In other words, they are blocked if the total number of data packet and voice calls in the system is greater than $N-1$. Therefore, the voice blocking probability can be expressed as:

$$P_B = 1 - \sum_{i=0}^{N} \sum_{d=0}^{N-i-1} P_i(d) \tag{3.63}$$

$$= 1 - P_0(0) \sum_{i=0}^{N} \sum_{d=0}^{N-i-1} \sum_{\ell=1}^{N} c_{i,\ell}(d) \rho_\ell . \tag{3.64}$$

Equation (3.64) gives the "exact" voice blocking probability, where the coefficients $c_{i,\ell}(d)$ are obtained as in (3.47) and (3.48) and ρ_ℓ's are obtained from (3.35) and (3.36). Bhat and Fisher indicated in [26] that the blocking probability of voice calls does not have a strong dependency on α, and so they proposed to use the closed-form expression that they found by assuming $\alpha = 1$. The exact result given in (3.64) and the approximation given in [26] are compared in Figs. 3.7 and 3.8. The expression found in [26] is exact for only $\alpha = 1$.

Equation (3.64) is plotted as a function of data trafic in Fig. 3.7 and as a function of voice traffic in Fig. 3.8 for $\alpha = \mu_1/\mu_2$ greater than 1. Our aim is to

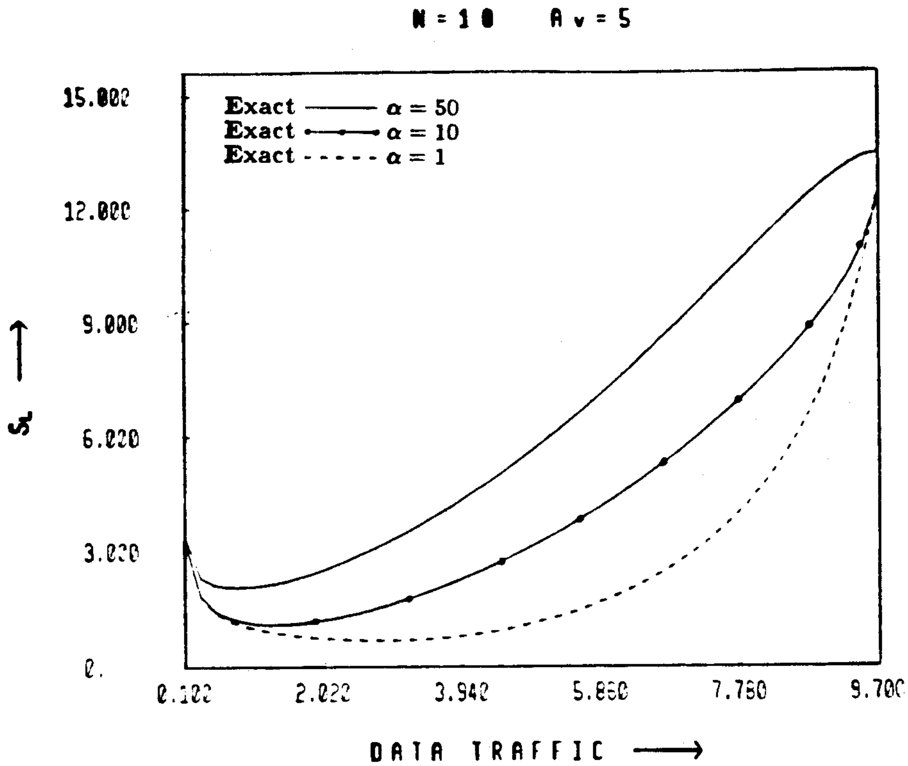


Fig. 3.5 Normalized standard deviation of packet size as a function of data traffic, when $N = 10$ and voice traffic is 5 Erlangs for $\alpha = 1$, 10, 50.

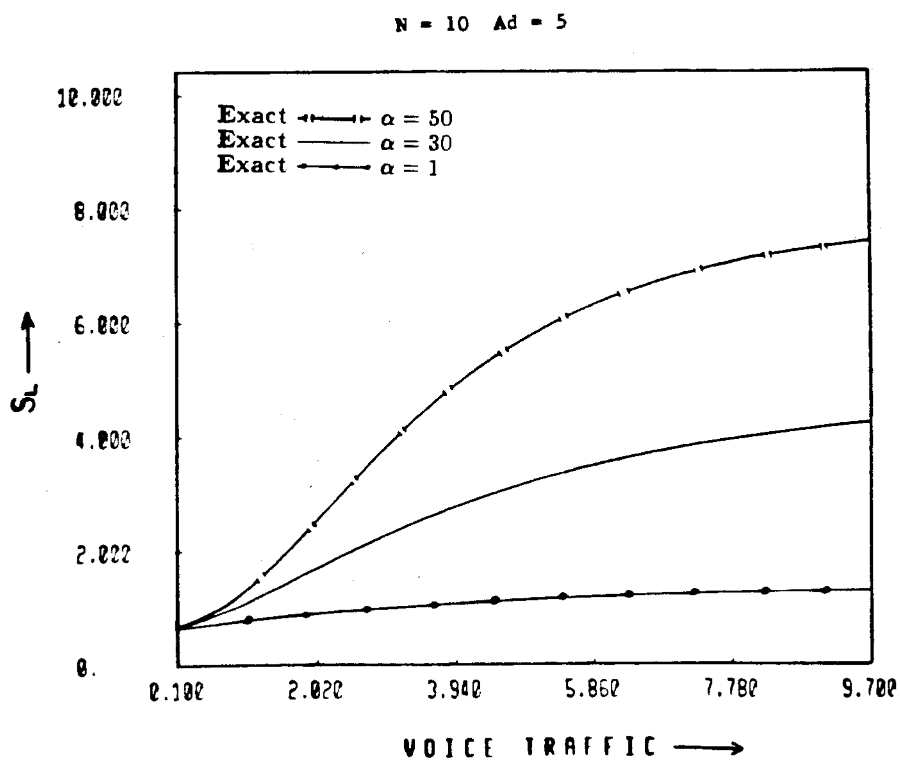


Fig. 3.6 Normalized standard deviation of packet size as a function of data traffic, when $N = 10$ and data traffic is 5 Erlangs for $\alpha = 1$, 10, 50.

check the dependency of blocking probability on α and to understand if we can use the closed-form approximation found in [26] when $\alpha > 1$. Figs. 3.7 depicts that at low data traffic intensities ($a_d < 2$) and at high data traffic intensities ($a_d > 8$), Bhat and Fisher's approximation does indeed give good results.

Figs. 3.7 and 3.8 depict the voice-blocking probability as a function of data and voice traffic respectively for $\alpha = 1, 30$ and 50. At high data traffic intensities, the voice blocking probability is close to 1, which means that the system is closed to the voice traffic most of the time.

3.2.2d. Effect of voice holding time on queue peaks:

From the designer's point of view, the fraction of time the packet size exceeds a finite threshold, say T, is important in designing an integrated system. The number of buffers in the system must be large enough to accomodate almost all data traffic. Therefore, the designer would like to know something about the maximum queue size. The probability of packet queue size being greater than a finite threshold, T, is given by:

$$Pr(d > T) = \sum_{m=0}^{N} \sum_{d=T+1}^{\infty} P_{N-m}(d) \tag{3.65}$$

$$= \sum_{m=0}^{N} \sum_{i=1}^{m} \frac{z_{2i}^{T+1}}{1 - z_{2i}} B_{N-m,i} \tag{3.66}$$

Here, we assume that $T > N+1$, because the closed-form expression found in (3.42) is valid here only for $T > N + 1$.

By employing equation (3.66), we obtain Figs. 3.9 and 3.10 for $Pr(d > T)$ as a function of T (Fig. 3.9) and as a function of data traffic (Fig. 3.10). In these plots, $\alpha = \mu_1/\mu_2$ is taken 30, the voice traffic, a_v, is 5 Erlangs and the data traffic, a_d, is 9 Erlangs. These figures depict that for larger values of α, the buffer sizes must be

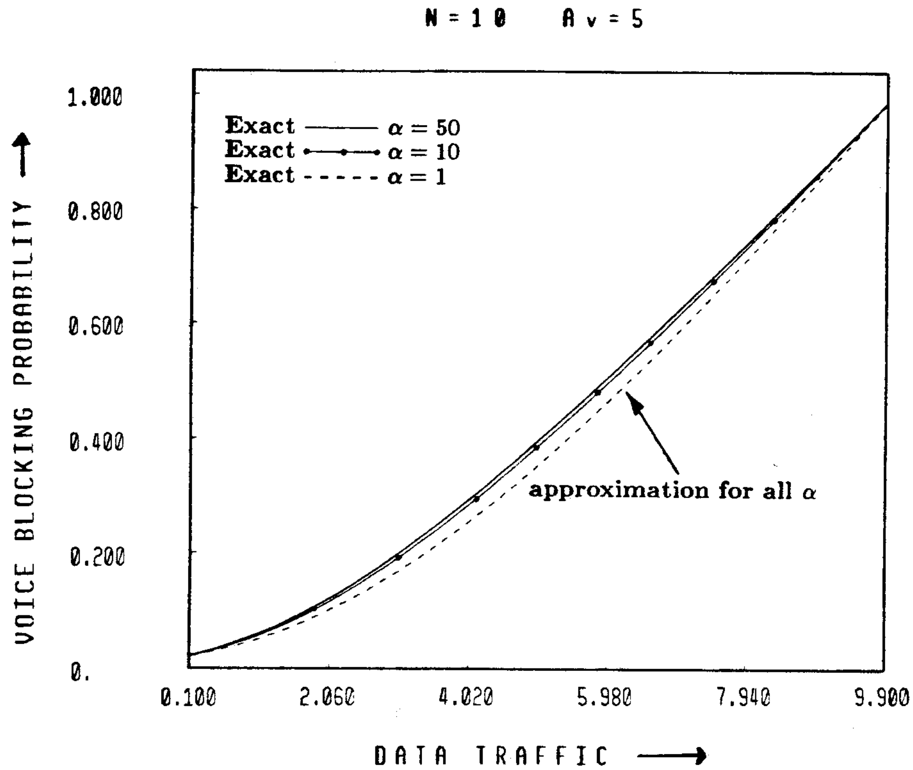


Fig. 3.7 Comparison of the voice blocking probability obtained by using exact analysis with Bhat and Fisher's approximation for different ratios of voice and data holding times ($\alpha = \mu_1/\mu_2$) and as a function of data traffic.

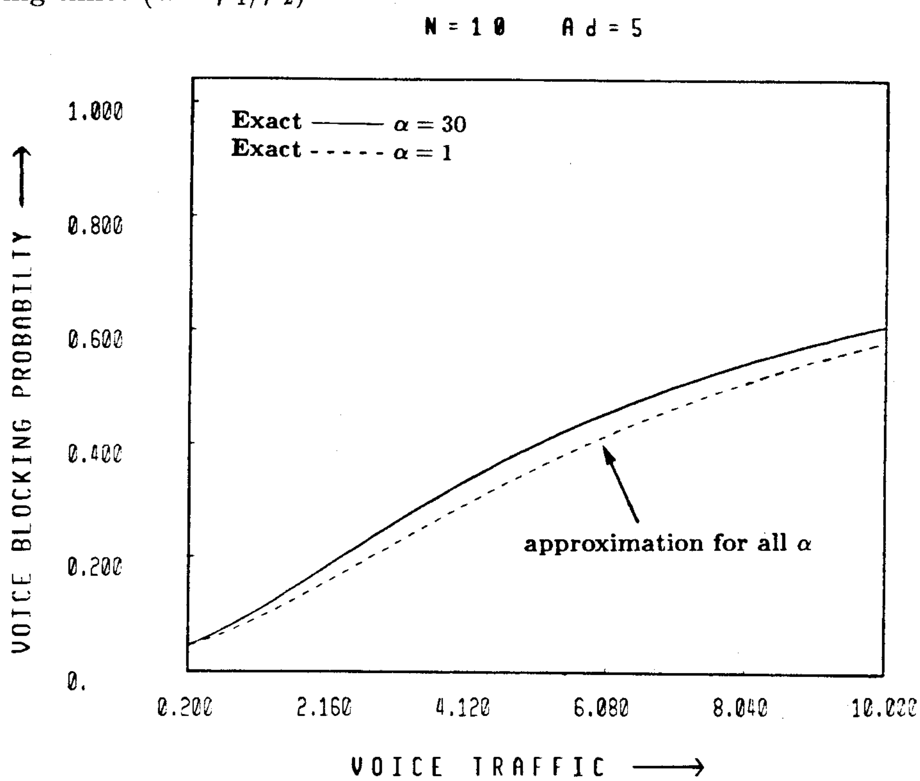


Fig. 3.8 Comparison of the voice blocking probability obtained by using exact analysis with Bhat and Fisher's approximation for different ratios of voice and data holding times ($\alpha = \mu_1/\mu_2$) and as a function of voice traffic.

kept larger. As an example, in Fig. 3.9 at $T = 51$, we obtain $Pr(d > T) = 0.01$ for $\alpha = 1$. That is, one percent of the time, the packet size in the system is larger than 51. Hence, in order not to lose packets more than one percent of the time, the buffer size must be 51. On the other hand, for $\alpha = 30$, the buffer size must be 122 to keep the same performance. At large traffic intensities and for large values of alpha, the necessity of buffering the data increases.

Mathematically speaking, the roots z_{2i} converge to 1, which makes $Pr(k > T)$ increase as can be observed in (3.66). The dependency of the roots on α can be seen as follows:

$$z_{2i} = \frac{-a_i - \sqrt{a_i^2 - 4b_i}}{2} \tag{3.67}$$

where

$$a_i = -\frac{\lambda_1 + i\mu_1 + (N-i)\mu_2}{i\mu_1}; \qquad b_i = \frac{\lambda_1}{i\mu_1}. \tag{3.68}$$

Substituting $\alpha = \mu_1/\mu_2$, we have

$$a_i = -\left(1 + b_i + \frac{N-i}{\alpha i}\right) \approx -(1 + b_i) \qquad \text{if} \alpha \gg (N-i)/i. \tag{3.69}$$

It can easily be shown that $z_{2i} = 1$ if $a_i = -(1 + b_i)$. Therefore, if $\alpha \gg (N-1) \geq (N-i)/i$, then the roots can be assumed to be equal since they are all very close to 1. This assumption leads to the following approximation valid if $\alpha \gg (N-1)$:

$$Pr(d \geq T) \approx z^T \sum_{m=0}^{N} \sum_{i=1}^{m} \frac{B_{N-m,2i}}{(1 - z_{2i})} \tag{3.70}$$

$$\approx z^T \, \phi \tag{3.71}$$

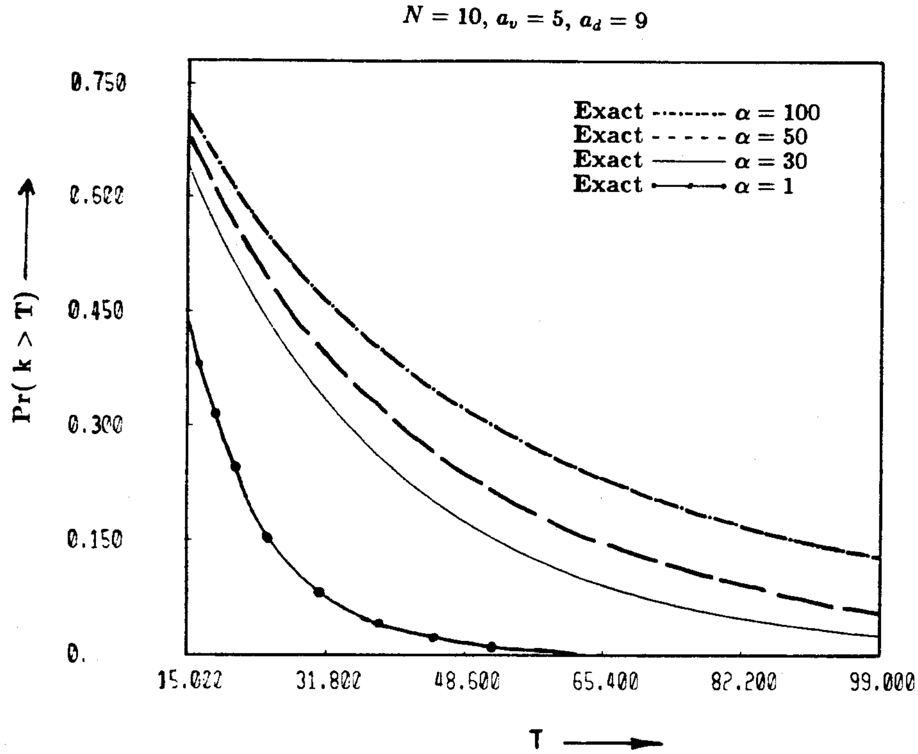


Fig. 3.9 The probability of packet size in the system is greater than the treshold T, as function of T, when $N = 10$, voice traffic is 5 Erlangs and data traffic is 9 Erlangs for $\alpha = 1, 30, 50, 100$.

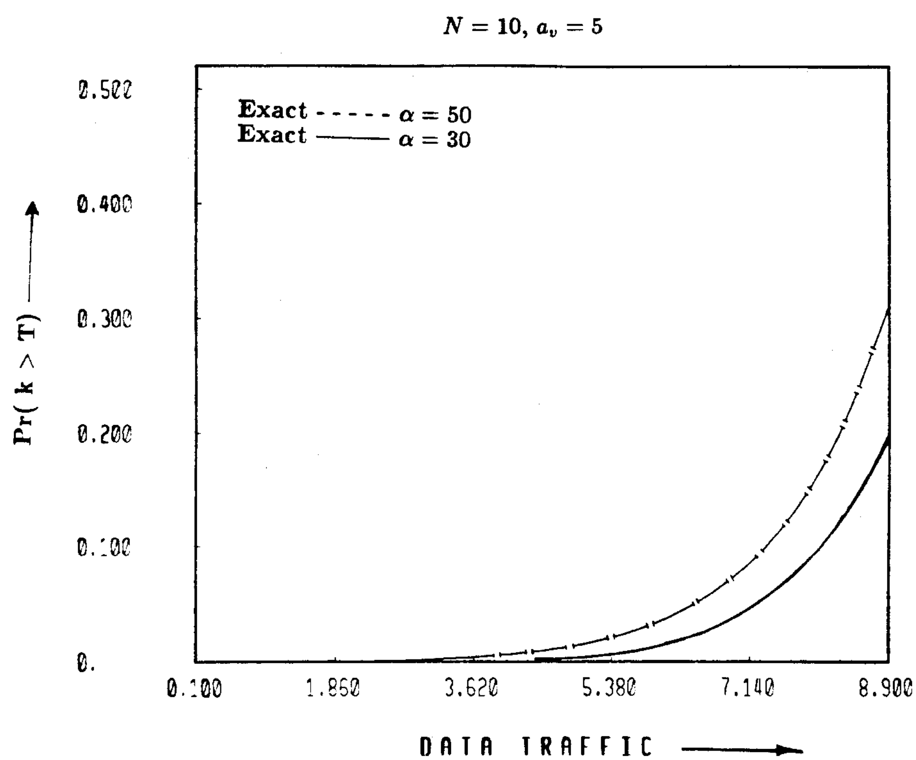


Fig. 3.10 The probability of packet size in the system is greater than the treshold $T = 50$, as a function of data traffic, when $N = 10$, voice traffic is 5 Erlangs for $\alpha = 30, 50$.

where ϕ is the sum of the coefficients of all roots:

$$\phi = \sum_{m=0}^{N} \sum_{i=1}^{m} \frac{B_{N-m,2i}}{(1 - z_{2i})} \quad \text{and} \quad z \approx z_{2i} \quad \text{for} \quad i = 0, 1, \ldots, N. \tag{3.72}$$

In particular, for $T = 0$,

$$Pr(d \geq 0) \approx \phi. \tag{3.73}$$

At high traffic intensities, it is not likely to find the system empty, so ϕ can be assumed to be equal to 1. Therefore, we obtain the following simple expression as an approximation to $Pr(d \geq T)$, which is valid when α is much larger than $N - 1$:

$$Pr(d \geq T) \approx z^T \tag{3.74}$$

Here, z can be chosen as the arithmetic mean of all roots, since they are all very close to each other:

$$z = \frac{1}{(N+1)} \sum_{i=0}^{N} z_{2i}. \tag{3.75}$$

For example, by employing (3.74), in order not to lose packets more than one percent of the time, the buffer size is determined as follows:

$$\text{Buffer size} = T = \frac{\log_{10} 0.01}{\log_{10} z} = -\frac{2}{\log_{10} z}. \tag{3.76}$$

$Pr(d > T)$ found by (3.74) is compared with the exact formula given in (3.66) in Fig. 3.11 for $\alpha = 50$ and in Fig. 3.12 for $\alpha = 100$. These figures show that the approximation is in a good agreement with the actual results, especially for large values of α.

3.2.3 Conclusions on FCFS Strategy:

Due to the high correlation of the number of voice calls in the system from frame to frame, the data traffic performance is degraded substantially. In order

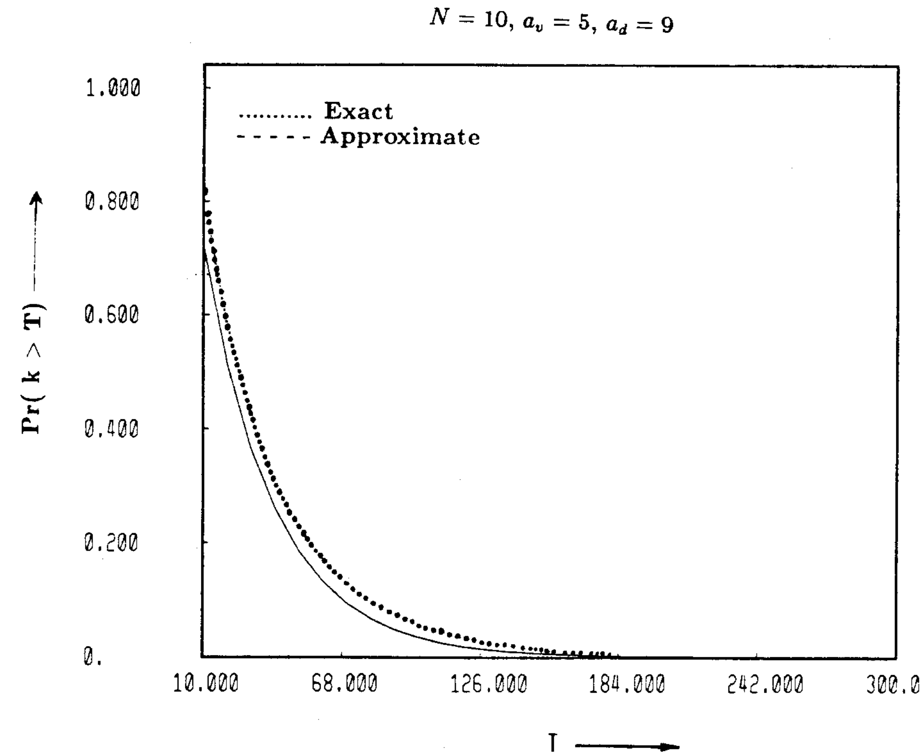


Fig. 3.11 Probability that packet size in the system is larger than T as a function of T for $\alpha = \mu_1/\mu_2 = 50$.

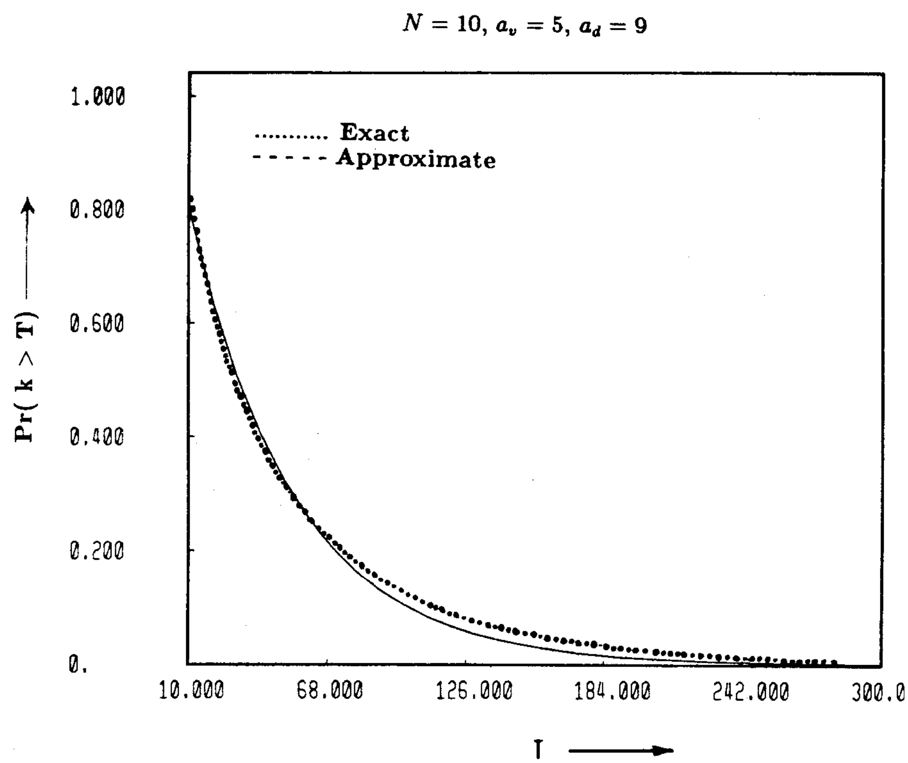


Fig. 3.12 Probability that packet size in the system is larger than T as a function of T for $\alpha = \mu_1/\mu_2 = 100$.

to improve the time delay performance of the data traffic, intervals when all the transmission capacity are occupied by voice calls only must be avoided. Another control strategy, which overcomes the problems of the FCFS control strategy, was proposed in [36], [37]. In this, the frame structure is divided into 2 parts. The first part is allocated to data traffic only. The analysis and the improvement obtained by this strategy, which is called the movable-boundary strategy, will be discussed in the next section.

3.3. Integration with Preemptive Movable-Boundary (PMB) Priority:

As indicated in the previous section, another way of controlling the service distribution between voice and data traffic is to divide the frame of N time slots into two sections. The first section containing N_1 slots is allocated to the voice traffic, while the remaining $N - N_1$ time slots, which form the second section, are reserved for data traffic only. Time slots in the first section may be used by data packets, but they are preempted by arriving voice calls if all the servers in the first section are busy. On the other hand, voice arrivals can not access the second section, so they are blocked if there is no available time slot in the first section. We model the system in continuous time, to allow state transition rate diagrams to be used. As defined previously, λ_2 is the voice and λ_1 the data arrival rate, while μ_2 is the voice and μ_1 the data departure rate. The state transition rate diagram of this model is given in Fig. 3.13.

3.3.1. Analysis:

The balance equations of first row states connect the first row states to the second row states. Therefore, the second row equilibrium probabilities can be obtained by employing only the first row states. The same argument is pursued to show that the steady-state probabilities in any row can be expressed in terms of the first row steady-state probabilities.

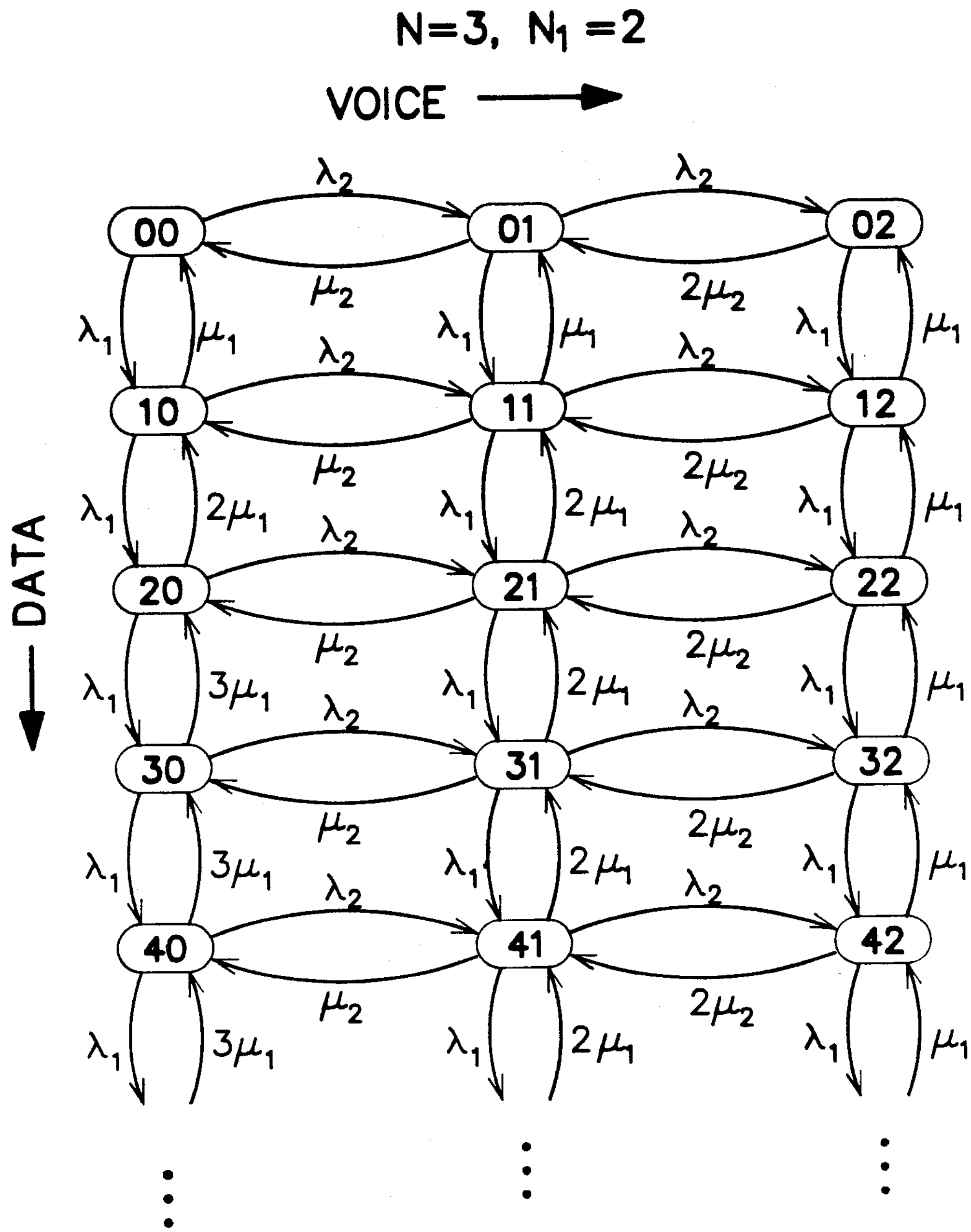


Fig. 3.13 The state transition rate diagram of PMB scheme for $N = 3$, $N_1 = 2$.

Hence, we can conclude that the key states of this Markovian state space are the first row states $S = [v, 0]$, $0 \leq v \leq N_1$.

As defined in the analysis of FCFS strategy, $P_{N_1-m}(d)$ denotes the probability of having d data packets and $N_1 - m$ voice calls in the system. It can be written in terms of key states as:

$$P_{N_1-m}(d) = \sum_{\ell=0}^{N_1} c_{N_1-m,\ell}(d) P_\ell(0) \quad d \geq 0, \qquad 0 \leq m \leq N_1, \tag{3.77}$$

where $c_{N_1-m,\ell}(d)$ is the coefficient of ℓ^{th} key state probability. The balance equation of state $S = [N_1 - m, d-1]$ yields the following difference equation for $d \geq N - N_1 + m + 1$ and $0 \leq m \leq N_1$:

$$P_{N_1-m}(d) + a_m P_{N_1-m}(d-1) + b_m P_{N_1-m}(d-2) =$$

$$d_m P_{N_1-m-1}(d-1) + e_m P_{N_1-m+1}(d-1),$$

$$d \geq M+1, \qquad 0 \leq m \leq N_1, \tag{3.78}$$

where $M = N - N_1 + m$, and

$$a_m = -\frac{M\mu_1 + (N_1 - m)\mu_2 + (1-\delta_m)\lambda_2 + \lambda_1}{M\mu_1}, \qquad b_m = \frac{\lambda_1}{M\mu_1} \tag{3.79}$$

$$d_m = -\frac{\lambda_2}{M\mu_1}, \qquad e_m = -\frac{(N_1 - m + 1)\mu_2}{M\mu_1}. \tag{3.80}$$

We substitute (3.77) into (3.78) to obtain the corresponding difference equations for the coefficients of key states:

$$c_{N_1-m,\ell}(d) + a_m c_{N_1-m,\ell}(d-1) + b_m c_{N_1-m,\ell}(d-2) =$$

$$d_m c_{N_1-m-1,\ell}(d-1) + e_m c_{N_1-m+1,\ell}(d-1)$$

$$d \geq M+1, \qquad 0 \leq m \leq N_1, \qquad 0 \leq \ell \leq N_1. \tag{3.81}$$

Note that for every difference equation in (3.78), we obtain $N_1 + 1$ difference equations in (3.81) for the key state coefficients. These equations can be put into a matrix form by making use of the following definitions. First, E is defined as a delay operator:

$$E c_{N_1-m,\ell}(d) = c_{N_1-m,\ell}(d-1). \tag{3.82}$$

Then

$$f_k(E) = 1 + a_k E + b_k E^2, \quad g_k(E) = -d_k E, \quad h_k(E) = -e_k E, \tag{3.83}$$

$$R(E) = \begin{pmatrix} f_0(E) & g_0(E) & 0 & \dots & 0 \\ h_1(E) & f_1(E) & g_1(E) & \dots & 0 \\ 0 & h_2(E) & f_2(E) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & f_{N_1}(E) \end{pmatrix}, \tag{3.84}$$

$$\vec{C}_\ell(d) = [c_{N_1,\ell}(d) \;\; c_{N_1-1,\ell}(d) \dots c_0(d)]^T. \tag{3.85}$$

Finally, (3.78) is written in matrix form as:

$$R(E)\; \vec{C}_\ell(d) = \vec{0} \qquad \ell = 0, \dots, N_1. \tag{3.86}$$

If the solution of (3.86) is in the following form:

$$\vec{C}_\ell(d) = \vec{B}_i z_i^d \tag{3.87}$$

where

$$\vec{B}_i = [b_{N_1,i}, b_{N_1-1,i}, \ldots, b_{0,i}]^T, \tag{3.88}$$

then (3.87) satisfies (3.86):

$$R(E)\ \vec{B}_i\ z_i^d = z_i^d\ R(z_i^{-1})\ \vec{B}_i = \vec{0}. \tag{3.89}$$

This implies that

$$R(z_i^{-1})\ \vec{B}_i = \vec{0}. \tag{3.90}$$

Therefore, in order to have a nontrivial solution, $R(z_i^{-1})$ must be a singular matrix. That is, the nullity of $R(z_i^{-1})$ is not zero, in fact $\vec{B}_i$ is an element of the nullspace of $R(z_i^{-1})$. Note that the degree of $\det R(z_i^{-1})$ is $2N_1+2$. This is because the degree of its diagonal entries is 2, and thus it has $2N_1+2$ roots. Assuming these roots are all distinct and real, the general form of the solution is the following:

$$\vec{C}_\ell(d) = \sum_{i=1}^{2N_1+2} k_{\ell,i}\vec{B}_i z_i^d, \quad d \geq M+1. \tag{3.91}$$

From (3.85), (3.88) and (3.91), we obtain for the entries of $\vec{C}_\ell(d)$:

$$c_{N_1-m,\ell}(d) = \sum_{i=1}^{2N_1+2} k_{\ell,i} b_{N_1-m,i} z_i^d, \quad d \geq M+1. \tag{3.92}$$

Our goal is to find the $k_{\ell,i}$, $b_{N_1-m,i}$, and z_i explicitly. Recall that each z_i is a root of the equation $detR(z^{-1}) = 0$. That is,

$$| R(z_i^{-1}) | = 0 \qquad i = 1, \ldots, 2N_1 + 2. \tag{3.93}$$

An iterative algorithm to find the determinant of $R(z^{-1})$ is obtained by expanding the last row of $R(z^{-1})$ starting from $f_{N_1}(z^{-1})$. We define the function $\phi_k(z)$ as the determinant of $R(z^{-1})$ for $N_1 = k$. Hence, the determinant of $R(z^{-1})$ is expressed iteratively as follows:

$$\begin{aligned}
\phi_{-1}(z) &= 1 \\
\phi_0(z) &= f_0(z^{-1}) \\
\phi_1(z) &= f_1(z^{-1})\phi_0(z) - h_1(z^{-1})g_0(z^{-1})\phi_{-1}(z) \\
&\vdots \\
\phi_{N_1}(z) &= f_{N_1}(z^{-1})\phi_{N_1-1}(z) - h_{N_1}(z^{-1})g_{N_1-1}(z^{-1})\phi_{N_1-2}(z)
\end{aligned}$$

where

$$| R(z^{-1}) | = \phi_{N_1}(z) \tag{3.94}$$

The roots of the determinant are readily found by using a computer program. In this study, we used the subroutines in the IMSL software library to obtain the zeros of the determinant.

The solution for $\vec{B}_i$ is found by employing (3.90). The entries $b_{N_1-m,i}$ of $\vec{B}_i$ are found iteratively as follows:

$$\begin{aligned}
b_{N_1,i} &= 1 \\
b_{N_1-1,i} &= -\frac{f_0(z_i^{-1})}{g_0(z_i^{-1})}, \\
&\vdots \\
b_{N_1-m,i} &= -\frac{h_{m-1}(z_i^{-1})}{g_{m-1}(z_i^{-1})} b_{N_1-m+2,i} - \frac{f_{m-1}(z_i^{-1})}{g_{m-1}(z_i^{-1})} b_{N_1-m+1,i}, \\
&\vdots \\
b_{0,i} &= -\frac{h_{N_1-1}(z_i^{-1})}{g_{N_1-1}(z_i^{-1})} b_{2,i} - \frac{f_{N_1-1}(z_i^{-1})}{g_{N_1-1}(z_i^{-1})} b_{1,i}.
\end{aligned}$$

The solution for the key state coefficients $c_{N_1-m,\ell}$ given in (3.92) will be completed after finding the k_i's, which are obtained by making use of the initial conditions at $d = M$ and $M+1$. We now proceed to do this. Note that from (3.81) we have

$$c_{N_1-m,\ell}(M+1) + I_{N_1-m,\ell} = e_m c_{N_1-m+1,\ell}(M). \tag{3.95}$$

$$c_{N_1-m,\ell}(M+2) + a_m c_{N_1-m,\ell}(M+1) + J_{N_1-m,\ell} = e_m c_{N_1-m-1,\ell}(M+1), \tag{3.96}$$

where

$$I_{N_1-m,\ell} = a_m c_{N_1-m,\ell}(M) + b_m c_{N_1-m,\ell}(M-1) - d_m c_{N_1-m-1,\ell}(M), \tag{3.97}$$

$$J_{N_1-m,\ell} = b_m c_{N_1-m,\ell}(M) - d_m c_{N_1-m-1,\ell}(M+1). \tag{3.98}$$

Substituting (3.92) into (3.95) and (3.96), we obtain:

$$\sum_{i=1}^{2N_1+2} k_{\ell,i} b_{N_1-m,i} z_i^{M+1} - e_m \sum_{i=1}^{2N_1+2} k_{\ell,i} b_{N_1-m+1,i} z_i^{M} = -I_{N_1-m,\ell} \tag{3.99}$$

$$\sum_{i=1}^{2N_1+2} k_{\ell,i} b_{N_1-m,i} z_i^{M+1}(z_i + a_m) - e_m \sum_{i=1}^{2N_1+2} k_{\ell,i} b_{N_1-m+1,i} z_i^{M+1} = -J_{N_1-m,\ell} \qquad (3.100)$$

for $m = 0, 1, ..., N_1$.

In (3.99) and (3.100), we have $2N_1 + 2$ equations for $2N_1 + 2$ coefficients $k_{\ell,i}$'s with ℓ fixed. These equations can be put in matrix form as:

$$\begin{pmatrix} q_{11} & q_{12} & q_{13} & \cdots & q_{1L} \\ q_{21} & q_{22} & q_{23} & \cdots & q_{2L} \\ q_{31} & q_{32} & q_{33} & \cdots & q_{3L} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{L1} & q_{L2} & q_{L3} & \cdots & q_{LL} \end{pmatrix} \begin{pmatrix} k_{\ell,1} \\ k_{\ell,2} \\ k_{\ell,3} \\ \vdots \\ k_{\ell,L} \end{pmatrix} = -\vec{S}_\ell \qquad (3.101)$$

where $L = 2N_1 + 2$ and $\vec{S}_\ell = [I_{N_1,\ell}\ I_{N_1-1,\ell} \ldots I_{0,\ell},\ J_{N_1,\ell}\ J_{N_1-1,\ell} \ldots J_{0,\ell}]^T$. The entries of the matrix $[q_{ij}]$ are given as follows:

$$q_{m+1,i} = z_i^M (b_{N_1-m,\ell,i}\ z_i - e_m b_{N_1-m+1,\ell,i}), \qquad (3.102)$$

$$q_{N_1+m+2,i} = z_i^{M+1}(b_{N_1-m,\ell,i}(z_i + a_m) - e_m b_{N_1-m+1,\ell,i}) \qquad (3.103)$$

for $m = 0, 1, .., N_1$ and $i = 1, 2, ..., 2N_1 + 2$.

Substituting the closed-form solution obtained for the key state coefficients $c_{N_1-m,\ell}(d)$ into (3.77), the equilibrium probabilities can now be expressed as:

$$P_{N_1-m}(d) = \sum_{\ell=0}^{N_1} \sum_{i=1}^{2N_1+2} k_{\ell,i} b_{N_1-m,i}\ z_i^d\ P_\ell(0), \qquad d \geq M, \qquad (3.104)$$

or

$$P_{N_1-m}(d) = \sum_{i=1}^{2N_1+2} b_{N_1-m,i} z_i^d \left(\sum_{\ell=0}^{N_1} k_{\ell,i} P_\ell(0) \right), \qquad d \geq M. \qquad (3.105)$$

If the system is stable, then the equilibrium probabilities must approach zero at infinity. That is,

$$\lim_{d \to \infty} P_{N_1-m}(d) = 0 \qquad (3.106)$$

This is possible if only the terms with $z_i \geq 1$ are exactly canceled. As in the analysis of FCFS strategy, the key state probabilities are found by making use of this stability requirement and the fact that probabilities must sum to 1.

So assuming that $z_i \geq 1$ for $i \in I$, where I is a subset of integers between 1 and $2N_1 + 2$, in order to have a stable system the key state probabilities must satisfy the following relations, one for each element of I:

$$\sum_{\ell=0}^{N_1} k_{\ell,i} P_\ell(0) = 0, \qquad i \in I. \tag{3.107}$$

These can be put into matrix form as:

$$\begin{pmatrix} k_{0,i_1} & k_{1,i_1} & \dots & k_{N_1,i_1} \\ k_{0,i_2} & k_{1,i_2} & \dots & k_{N_1,i_2} \\ \vdots & \vdots & \ddots & \vdots \\ k_{0,i_J} & k_{1,i_J} & \dots & k_{N_1,i_J} \end{pmatrix} \begin{pmatrix} P_0(0) \\ P_1(0) \\ \vdots \\ P_{N_1}(0) \end{pmatrix} = \vec{0}. \tag{3.108}$$

Here $I = \{i_1,\ i_2, \dots,\ i_J\}$.

If the queueing system is stable, then the system of linear equations (3.108) has a nontrivial solution. This means that the nullity of above matrix is nonzero. Without loss of generality, we can assume that the normalized solution of (3.108) is expressed as:

$$[1\ \sigma_1\ \sigma_2\ \dots \sigma_{N_1}]^T. \tag{3.109}$$

The key state probabilities can be related to empty state probability by employing the solution given in (3.109). In fact, σ_j is the ratio of $P_j(0)$ to the empty state probability $P_0(0)$:

$$P_j(0) = \sigma_j\ P_0(0). \tag{3.110}$$

Finally, the empty state probability $P_0(0)$ is found explicitly by using the normalization condition. The final form of the equilibrium probabilities are:

$$P_{N_1-m}(d) = P_0(0) \sum_{i=1}^{2N_1+2} b_{N_1-m,i} z_i^d \left(\sum_{\ell=0}^{N_1} k_{\ell,i}\ \sigma_\ell \right), \quad d \geq M+1. \tag{3.111}$$

3.3.2. Special case: $N = 3,\ N_1 = 1$.

In this section, closed-form expressions are obtained for the equilibrium probabilities for a case where the total number of servers is 3 and only one of them is allocated for voice traffic. The state transition rate diagram of this example has only two columns. The balance equations of the states in these columns are given as:

$$P_1(d) + a_0 P_1(d-1) + b_0 P_1(d-2) = d_0 P_0(d-1), \quad d \geq 3, \tag{3.112}$$

$$P_0(d) + a_1 P_0(d-1) + b_1 P_0(d-2) = e_1 P_1(d-1), \quad d \geq 4, \tag{3.113}$$

where $e_0 = d_1 = 0$ and

$$a_0 = -\frac{2\mu_1 + \mu_2 + \lambda_1}{2\mu_1}, \qquad a_1 = -\frac{3\mu_1 + \lambda_2 + \lambda_1}{3\mu_1}, \tag{3.114}$$

$$b_0 = \frac{\lambda_1}{2\mu_1}, \qquad b_1 = \frac{\lambda_1}{3\mu_1}, \tag{3.115}$$

$$d_0 = -\frac{\lambda_2}{2\mu_1}, \qquad e_1 = -\frac{\mu_2}{3\mu_1}. \tag{3.116}$$

The $(N_1+1)\mathrm{x}(N_1+1) = 2\mathrm{x}2$ matrix $R(z^{-1})$ defined in equation (3.84) is found for this case as follows:

$$R(z^{-1}) = \begin{pmatrix} f_0(z^{-1}) & g_0(z^{-1}) \\ h_1(z^{-1}) & f_1(z^{-1}) \end{pmatrix}. \tag{3.117}$$

By employing the definitions of $f_0,\ f_1,\ g_0,\ h_0$, given in (3.84), the determinant of $R(z^{-1})$ is obtained as:

$$\det R(z^{-1}) = f_0(z^{-1})f_1(z^{-1}) - h_1(z^{-1})g_0(z^{-1})$$
$$= r_4 z^{-4} + r_3 z^{-3} + r_2 z^{-2} + r_1 z^{-1} + 1, \qquad (3.118)$$

where

$$r_4 = b_0 b_1, \quad r_3 = a_0 b_1 + a_1 b_0,$$

$$r_2 = a_1 a_0 + b_1 - e_1 d_0, \quad r_1 = a_0 + a_1.$$

Let the roots of this quartic determinant polynomial be $z_1,\ z_2,\ z_3,\ z_4$. Not very useful closed-form expressions can be obtained for these roots by making use of the Descartes-Euler or Ferrari solutions described in [40], and we shall review this. The quartic equation given in (3.118) is transformed to the following reduced formed through the substitution $z^{-1} = y^{-1} - \frac{r_3}{4r_4}$:

$$y^{-4} + py^{-2} + qy^{-1} + r = 0.$$

The roots of this transformed equation are the four sums:

$$(\pm\sqrt{s_1} \pm \sqrt{s_2} \pm \sqrt{s_3})^{-1}$$

with the sign of the square roots chosen so that

$$\sqrt{s_1}\sqrt{s_2}\sqrt{s_3} = -q/8.$$

Here $s_1,\ s_2,\ ,s_3$ are the roots of the cubic equation

$$x^3 + \frac{p}{2}x^2 + \frac{p^2 - 4r}{16}x - \frac{q^2}{64} = 0,$$

which then, of course, has to be solved.

Assuming now that we have the four roots $z_1, \ldots, z_4$, the solution of the key-state coefficients is, as in (3.92),

$$c_{1-m,\ell}(d) = \sum_{i=1}^{4} k_{\ell,i} b_{1-m,i} z_i^d, \qquad m = 0, 1 \quad d \geq 3 + m. \tag{3.119}$$

From the iterative algorithm given for $b_{N1-m,i}$, we obtain $b_{1,i} = 1$ and

$$\begin{aligned} b_{0,i} &= -\frac{f_0(z_i^{-1})}{g_0(z_i^{-1})} \\ &= \frac{1 + a_0 z_i^{-1} + b_0 z_i^{-2}}{d_0 z_i^{-1}}, \qquad i = 1, 2, 3, 4. \end{aligned} \tag{3.120}$$

The vector $\vec{S}_\ell$ defined in (3.101) is found as:

$$\vec{S}_\ell = [I_{1,\ell} \; I_{0,\ell} \; J_{1,\ell} \; J_{0,\ell}]^T \tag{3.121}$$

where

$$I_{1,\ell} = a_0 c_{1,\ell}(2) + b_0 c_{1,\ell}(1) - d_0 c_{0,\ell}(2), \tag{3.122}$$

$$I_{0,\ell} = a_1 c_{0,\ell}(3) + b_1 c_{0,\ell}(2), \tag{3.123}$$

$$J_{1,\ell} = b_0 c_{1,\ell}(2) - d_0 c_{0,\ell}(3), \tag{3.124}$$

$$J_{0,\ell} = b_1 c_{0,\ell}(3). \tag{3.125}$$

By employing equations (3.102) and (3.103), we obtain the entries of the matrix Q as:

$$q_{1j} = b_{1j}\ z_j^3, \qquad q_{2j} = z_j^3\ (z_j - e_1\ b_{1j}),$$

$$q_{3j} = b_{1j}\ z_j^3\ (z_j + a_0), \qquad q_{4j} = z_j^4\ (z_j + a_1 - e_1\ b_{11}) \quad j = 1,2,3,4. \tag{3.126}$$

From (3.101), the coefficient vector $\vec{K}_\ell$ is solved for as:

$$\vec{K}_\ell = -Q^{-1}\ \vec{S}_\ell, \tag{3.127}$$

where

$$\vec{K}_\ell = [k_{\ell,1}\ k_{\ell,2}\ k_{\ell,3}\ k_{\ell,4}]^T. \tag{3.128}$$

At least one of the zeros of the determinant polynomial must be greater than 1, otherwise the solution would not be unique. For if we assume that all zeros are less than 1, then the stability requirement given in (3.106) is satisfied automatically. In this case, the only remaining equation to find the relationship between $P_0(0)$ and $P_1(0)$ would be the normalization equation. But this is only one equation for two unknowns. Therefore, an infinite number of distributions would satisfy the state balance equations. This would contradict the uniqueness of the solution.

By using the Routh Hurwitz criterion [39], it is possible to show that half of the roots of the determinant polynomial given in (3.118) are greater than or equal to one and the other half of the roots are less than one. Without loss of generality, we can number the four roots found from (3.118) so that z_1, z_2 are less than 1 and z_3, z_4 are greater than or equal to 1. Then, in order to have a stable system, we see from (3.107) that the key state probabilities $P_0(0)$, $P_1(0)$ must satisfy

$$k_{03}P_0(0) + k_{13}P_1(0) = 0 \tag{3.129}$$

$$k_{04}P_0(0) + k_{14}P_1(0) = 0. \tag{3.130}$$

Therefore,

$$P_1(0) = -\frac{k_{03}}{k_{13}} P_0(0). \tag{3.131}$$

$$= -\frac{k_{04}}{k_{14}} P_0(0). \tag{3.132}$$

Note that equations (3.129) and (3.130) are linearly dependent. This must be, otherwise there would not be a nontrivial solution. But we know there exists a nontrivial solution as long as the stability requirement is satisfied, which is:

$$a_d < N - E_v = N - a_v(1 - P_B). \tag{3.133}$$

Here P_B is the Erlang-B blocking probability of the voice traffic, a_v the offered voice traffic and a_d the data traffic, while E_v is the expected number of voice calls in the system, and the carried voice traffic is $a_v(1 - P_B)$.

The stability condition given in (3.133) is intuitively appealing, since it means that the data queue is at equilibrium only for

$$\frac{a_d}{(N - E_v)} < 1. \tag{3.134}$$

Here N is the total number of servers and E_v is the mean number of servers occupied by voice traffic. Therefore, on the average $N - E_v$ servers are available for data traffic. Hence the stability requirement given in (3.133) is equivalent to the stability requirement of an M/M/$(N - E_v)$ queueing system. However, the known results of M/M/n queueing networks cannot be used to derive (3.134) because they require the assumption the that data queue reaches equilibrium at each voice state. This assumption is valid if only the voice holding time is much larger than data holding time and offered data traffic is below a certain threshold.

By employing (3.132), the $\vec{\sigma}$ defined in (3.109) is found as:

$$\vec{\sigma} = [1 \quad -\frac{k_{0,4}}{k_{1,4}}]^T. \tag{3.135}$$

Using this, we are ready to give the closed-form expressions for the equilibrium probabilities. By employing (3.111), the first-row equilibrium probabilities are:

$$P_0(d) = \sum_{i=1}^{4} k_{0,i}\ b_{0,i}\ z_i^d\ P_0(0) - \sum_{i=1}^{4} k_{1,i}\ b_{0,i}\ z_i^d \frac{k_{0,4}}{k_{1,4}} P_0(0) \tag{3.136}$$

$$= \sum_{i=1}^{4}\left(k_{0,i} - k_{1,i}\frac{k_{0,4}}{k_{1,4}}\right) b_{0,i} z_i^d P_0(0), \quad d \geq 4, \tag{3.137}$$

and the second-row probabilities are:

$$P_1(d) = \sum_{i=1}^{4} k_{0,i} b_{1,i}\ z_i^d\ P_0(0) - \sum_{i=1}^{4} k_{1,i}\ b_{1,i}\ z_i^d\ \frac{k_{0,4}}{k_{1,4}} P_0(0) \tag{3.138}$$

$$= \sum_{i=1}^{4}\left(k_{0,i} - k_{1,i}\ \frac{k_{0,4}}{k_{1,4}}\right) z_i^d P_0(0), \quad d \geq 3. \tag{3.139}$$

By making use of (3.131), the equilibrium probabilities can now be rewritten as:

$$P_0(d) = P_0(0)(\psi_{01} z_1^d + \psi_{02} z_2^d), \qquad d \geq 4, \tag{3.140}$$

$$P_1(d) = P_0(0)(\psi_{11} z_1^d + \psi_{12} z_2^d), \qquad d \geq 3, \tag{3.141}$$

where

$$\psi_{01} = b_{01}(k_{01} - \frac{k_{04}}{k_{14}} k_{11}), \qquad \psi_{02} = b_{02}(k_{02} - \frac{k_{04}}{k_{14}} k_{12}), \tag{3.142}$$

$$\psi_{11} = (k_{01} - \frac{k_{04}}{k_{14}} k_{11}), \qquad \psi_{12} = (k_{02} - \frac{k_{04}}{k_{14}} k_{12}). \tag{3.143}$$

The empty state probability $P_0(0)$ is found by applying the normalization condition:

$$P_0(0) = \left(\frac{\psi_{01} z_1^4 + \psi_{11} z_1^3}{1 - z_1} + \frac{\psi_{02} z_2^4 + \psi_{12} z_2^3}{1 - z_2} + K_0 \right)^{-1}, \tag{3.144}$$

where

$$K_0 = \sum_{m=0}^{1} \sum_{d=0}^{2+m} (c_{1-m,0}(d) - \frac{k_{04}}{k_{14}} c_{1-m,1}(d)). \tag{3.145}$$

The expected number of data packets, $E(L)$, in the system is:

$$E(L) = \sum_{d=0}^{\infty} d(P_0(0) + P_1(0)). \tag{3.146}$$

From (3.146), the normalized mean queueing time W is obtained from Little's result as:

$$W = \frac{E(L) - a_d}{a_d} = \frac{E(L)}{a_d} - 1. \tag{3.147}$$

The following expression is found for $E(L)$ by making use of the closed-form expressions obtained in (3.140) and (3.141):

$$E(L) = P_0(0) \left(\sum_{i=1}^{2} \psi_{0i} \frac{z_i^3 (3 - 2z_i)}{1 - z_i} + \sum_{i=1}^{2} \psi_{1i} \frac{z_i^4 (4 - 3z_i)}{1 - z_i} + E(L_0) \right) \tag{3.148}$$

where

$$E(L_0) = \sum_{m=0}^{1} \sum_{d=0}^{2+m} d \ (c_{1-m,0}(d) - \frac{k_{04}}{k_{14}} c_{1-m,1}(d)). \tag{3.149}$$

The following section is devoted to numerically comparing the performance results of movable boundary, first-come, first-served and M/M/2 queueing systems.

3.4 Comparison of PMB, FCFS and M/M/2

Fig. 3.14 shows the normalized mean queueing time as a function of data traffic for the movable boundary (PMB) strategy. Here the total number of servers $N = 3$, the offered voice traffic is 1 Erlang, and only one server is available for voice traffic, that is, $N_1 = 1$. The normalized mean queueing time is computed by employing equation (3.147). Since the corresponding fixed-boundary system in this case is an M/M/2 queueing system, the results obtained for PMB strategy are compared to the results of the M/M/2 case. The results show clearly that the PMB strategy is quite superior to the fixed boundary strategy, as expected. For the same queueing time, we see that data traffic intensity is almost doubled in the PMB case. For example, when the normalized queueing time is one unit, the data traffic carried is 1 Erlang in the M/M/2 queueing system and 1.85 Erlangs in PMB system when $\alpha = 1$. (Recall that α was equal to μ_1/μ_2 with μ_1^{-1} is the mean holding time of a data packet and μ_2^{-1} is the mean holding time of a voice call.)

As expected , the dependence of queueing time on α is negligible for low data traffic intensity. This is because the number of channels reserved for data use is greater than the data traffic. So, it is not likely for the data arrivals to occupy the secondary channels that are also shared by voice calls. Therefore, the mean holding time of voice calls does not affect the system performance at low data traffic intensities. But at high data traffic intensities, the reserved channels are insufficient to accomodate the data traffic, and data arrivals attempt to use the secondary servers. In this case, if the voice calls occupy the secondary servers for a long time, data arrivals cannot access the system, and a data queue forms. Consequently, the queueing time is substantially affected by α . In Fig. 3.14, the normalized queueing time is plotted as a function of data traffic for $\alpha = 1$, 10 and 25. For data traffic values smaller than 1 Erlang, the difference between the three curves is negligible, i. e., the dependence upon α is very small, but this is not the case for high data-traffic levels.

On the contrary, as discussed previously, the FCFS scheme is very sensitive to $\alpha = \mu_1/\mu_2$ at all traffic levels. The performance degrades suddenly as α increases. In Fig. 3.15, the normalized queueing time is plotted as a function of data traffic in the FCFS case for $\alpha = 1, 10, 25$. Here again, the number of servers is 3 and the offered voice traffic is 1 Erlang. At low data-traffic intensities, we observe that as far as the time delay performance is concerned, the FCFS scheme is worse than the M/M/2 queueing system. On the other hand, at high traffic intensities it is even better than the PMB scheme. This is because at high traffic intensities, in the FCFS scheme data packets occupy all the servers (slots) until all the queue is emptied. However, in the PMB scheme voice calls are given preemptive priority over data packets. Therefore, voice calls are allowed to preempt data packets in service occupying slots allocated to voice if necessary for the voice call to receive service. As a result, at high traffic intensities the expected number of servers available for data traffic in PMB case is less than the number of slots available in FCFS case.

The blocking probability of voice calls in the PMB scheme is simply given by the Erlang-B formula, because voice calls do not experience any interference from data traffic. That is, there are always N_1 servers (slots) available for voice traffic. As explained above, although the data packets are allowed to use a slot allocated for voice traffic, arriving voice calls preempt those data packets from service when there is no other slot available for the voice call. We therefore have:

$$P_B = \frac{a_v^{N_1}/N_1!}{\sum_{j=0}^{N_1} \frac{a_v^j}{j!}}, \tag{3.150}$$

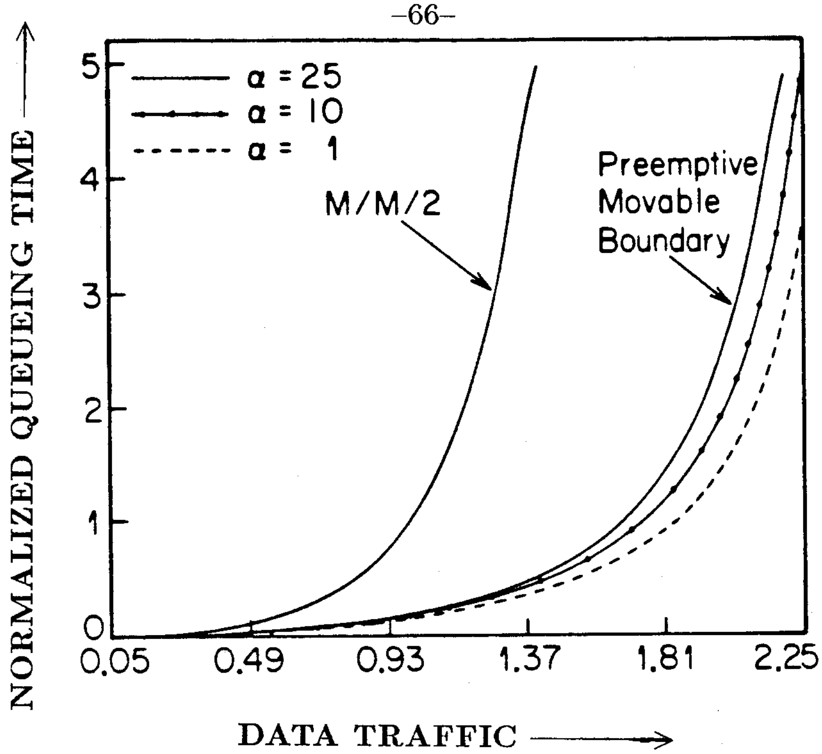


Fig. 3.14 The normalized queueing time as a function of data traffic for PMB and M/M/2 schemes for three different values of α. Here, $N = 3$, $N_1 = 1$ and $a_v = 1$.

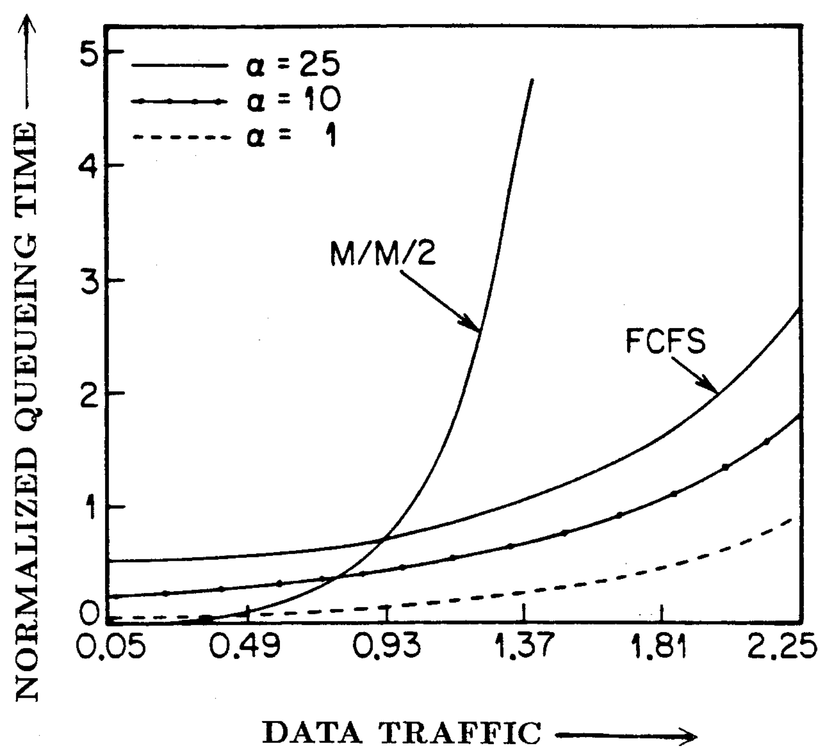


Fig. 3.15 The normalized queueing time as a function of data traffic for FCFS and M/M/2 schemes for three different values of α when $N = N_1 = 3$ and $a_v = 1$.

where a_v is the offered voice traffic and N_1 is the number of channels available for voice traffic. For example, we earlier considered in this section $a_v = 1$ and $N_1 = 1$. Therefore, the voice-blocking probability is found from (3.150) as:

$$P_B = \frac{a_v}{1 + a_v} = 0.5. \tag{3.151}$$

Note that P_B is independent of data traffic for the reason described above.

In Fig. 3.16, the blocking probabilty of voice calls is plotted as a function of data traffic for the PMB and FCFS scheme. As observed previously, in the FCFS case, P_B has a very slight dependence on α and increases almost linearly with data traffic. Here we also observe that at low data traffic levels, P_B is smaller in the FCFS than in the PMB case. But at high data traffic levels, the situation is reversed.

Time-Delay Performance		
Rank	Low Traffic*	High Traffic*
1	PMB	FCFS
2	M/M/2	PMB
3	FCFS	M/M/2

* (Carried data traffic)

Table 1. *Comparison of the Time-Delay Performance of PMB and FCFS Schemes for Voice-Data Integration in a Slotted-Channel system.*

Voice Blockage Performance		
Rank	Low Traffic*	High Traffic*
1	FCFS	PMB
2	PMB	FCFS

* (Carried data traffic)

Table 2. *Comparison of the Voice-Blockage Performance of PMB and FCFS Schemes for Voice-Data Integration in a Slotted-Channel system.*

In Tables 1 and 2, the time-delay and voice blockage performances of the FCFS and PMB strategies are qualitatively compared. The time-delay performance of an M/M/2 data-only system is also compared to the other strategies in Table 1.

3.5 Comparison with the previous work:

Two types of approximation have been proposed for the analysis of the movable-boundary scheme for a large numbers of channels (slots in a frame). The first approximation [34] is appropriate for the *underload* region of operation, $a_d < N - N_1$. The second approximation, a so called *fluid-flow* approach [35], [42], is appropriate for the *overload* region of operation, $a_d > N - N_1$. Both approximations adopt a continous-time model of the integrated system. That is, the frame length is assumed to be small compared to voice and even the data service time.

In the underload region, assuming that the voice holding time is much larger than the data holding time, i. e. $\alpha \gg 1$, the mean queue length could be obtained for each voice state by making use of the classical results of an $M/M/(N-i)$ queueing system. This is because the data queue is assumed to reach the equilibrium for each of the $N_1 + 1$ voice states. Therefore, the mean data queue length is approximated in prior work as:

$$E(L) \approx \sum_{i=0}^{N_1} E\{d/i\}\ P_v(i), \quad a_d < N - N_1, \tag{3.152}$$

where $P_v(i)$ is the probability of having i voice calls in the system. Since the voice calls use N_1 channels without any data interference, $P_v(i)$ is the Erlang probability of an $M/M/N_1$ loss system:

$$P_v(i) = \frac{a_v^i/i!}{\sum_{j=0}^{N_1} a_v^j/j!} \quad i = 0, 1, ...N_1, \tag{3.153}$$

with a_v as the offered voice traffic. Here $E\{q/i\}$, the mean queue length in an

M/M/$(N-i)$ queueing system, is given by:

$$E\{q/i\} = P_0 \frac{a_d^{N-i}}{(N-i)!} \frac{\frac{a_d}{(N-i)}}{\left(1 - \frac{a_d}{(N-i)}\right)^2}, \tag{3.154}$$

with

$$P_0 = \left(\sum_{k=0}^{N-i-1} \frac{a_d^k}{k!} + \frac{1}{1 - \frac{a_d}{(N-i)}} \right)^{-1}. \tag{3.155}$$

In the overload region the behavior of data queue has been approximated based on *fluid-flow* approach [42], [3, sect. 12-4-3, p. 702]. This approach assumes intense voice activity and high buildup of the data queue. It has been observed in [3, p. 706] that the data queue length behaves almost like a continuous variable in the overload region. Therefore, the queue length can be represented by an $x(t)$. The probability density function that there are x packets in the system at equilibrium, conditioned on having i voice calls present, is defined as $f_i(x)$. A set of forward equations can be obtained [3, p. 707] which represent the evolution of the density function $f_i(x)$ at steady state. These are first-order differential equations and the solution yields the density $f_i(x)$. The mean queue length is then approximated by the following integral:

$$E(L) \approx \int_0^\infty x \sum_{i=0}^{N_1} f_i(x)\,dx. \tag{3.156}$$

By using this fluid-flow approximation in the data overflow region, the following expression was obtained in [3, equation (12-166)] for the normalized mean queueing time when $N_1 = 1$:

$$W \approx \frac{\alpha(a_d - (N-1))a_v}{a_d N - (a_d + a_v(1-P_B))(1+a_v)^2}, \quad a_d > N-1. \tag{3.157}$$

Fig. 3.17 compares the underload and overload approximations of mean queueing time obtained by employing (3.152) and (3.157) with the fully exact (except for time quantization) result obtained in equation (3.147). The system parameters are $\alpha = 25$, $N = 3$, $N_1 = 1$, and the offered voice traffic is $a_v = 1$. Therefore, the underload region is defined by $a_d < 2$, and the overload region is defined by $a_d > 2$. It is clear from Fig. 3.17 that in the neighborhood of $a_d = 2$, both approximations deviate from the actual curve drastically. However, the underload region approximation does give extremely good results at very low data traffic intensities ($a_d < 1$). On the other hand, the overload approximation is only good at very high data traffic intensities ($a_d > 2.25$) and large α values. This is simply because for the justification of the fluid-flow approach the data queue must be large. So the new results here provide the first accurate method for obtaining the performance parameters of the movable boundary scheme at all traffic levels. Since the approximations are very poor at the medium data traffic levels, to be able to obtain the actual results in this region is very important. Also, this new method yields closed-form expressions for the equilibrium probabilities. This makes it not only possible but even easy to obtain all the important parameters besides mean delay and blocking probability, such as second order statistics, the required buffer size for data packets, etc.

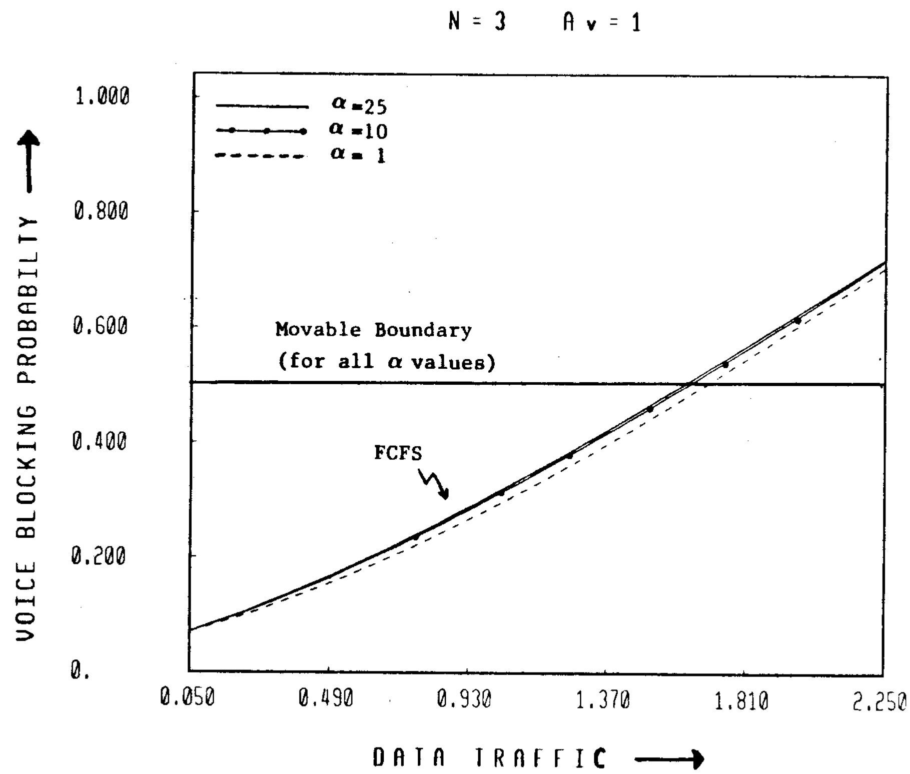


Fig. 3.16 Comparison of the blocking probability of voice calls as a function of data traffic in PMB and FCFS cases. Here $N = 3$, $N_1 = 1$ and $a_v = 1$.

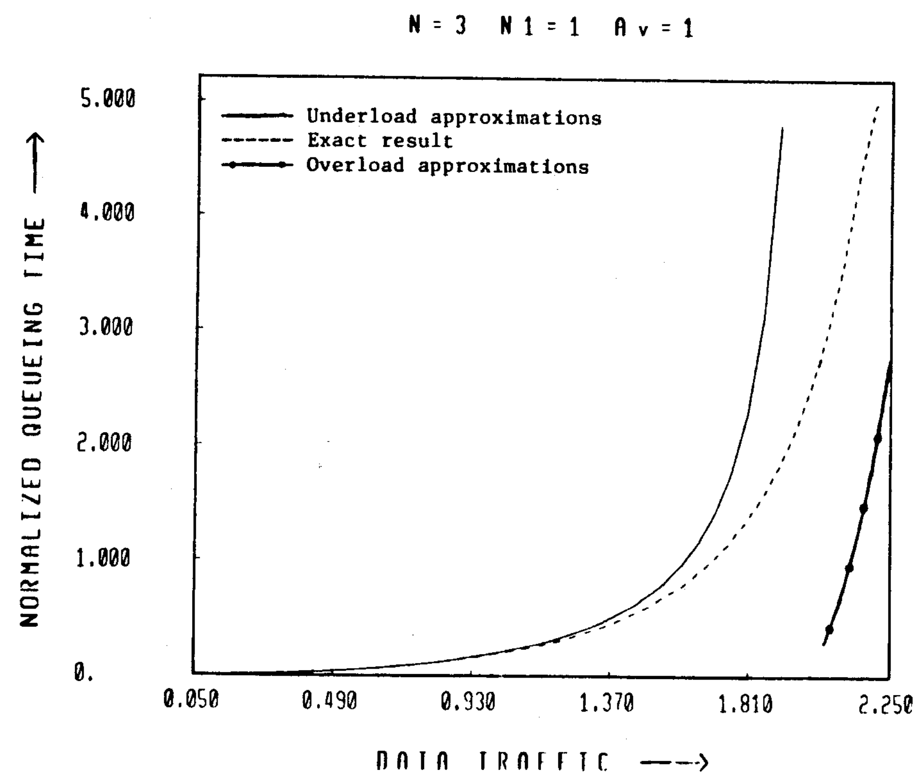


Fig. 3.17 Comparison of the underload and overload mean queueing time approximations to the actual result for PMB case. Here, $\alpha = 25$, $N = 3$, $N_1 = 1$ and $a_v = 1$.

CHAPTER 4

TRUNKED MOBILE RADIO NETWORKS

Although Trunked Mobile Radio networks are physically much different from voice-data integration networks, their queueing models and traffic problems are exactly the same. Instead of the time slots of a TDM frame, here the frequency slots of a portion of the communication spectrum reserved for mobile communication are shared. There is a repeater corresponding to each frequency slot, the place where the messages are transmitted and received. The detailed description of this system is given below.

4.1 Introduction to Trunked Mobile Radio:

In the last several years, centralized interconnect trunked systems have been introduced to service both dispatch-type mobile users and telephone line users [28]. The term trunked means that the telephone line is interfaced to the radio system at the repeater, which allows both dispatch and telephone line users to access the system simultaneously. Thus, two different types of communications utilize the same system facilities. As far as communications traffic is concerned, the basic difference between these two types of users is their average holding (service) times. A typical dispatch communication is conducted between the members of a mobile fleet or between the base station and mobiles. These are very short messages compared to an average telephone conversation, such as conducted by interconnect users. We call the telephone line users interconnect users and the dispatch-type mobile users dispatch users.

Fig. 4.1 shows the basic components of the trunked mobile system, which was originally designed by Motorola [28]. We note that only a fraction of the repeaters can be physically connected to interconnect calls. Therefore, the number of repeaters available for interconnect users is restricted. On the other hand, dispatch

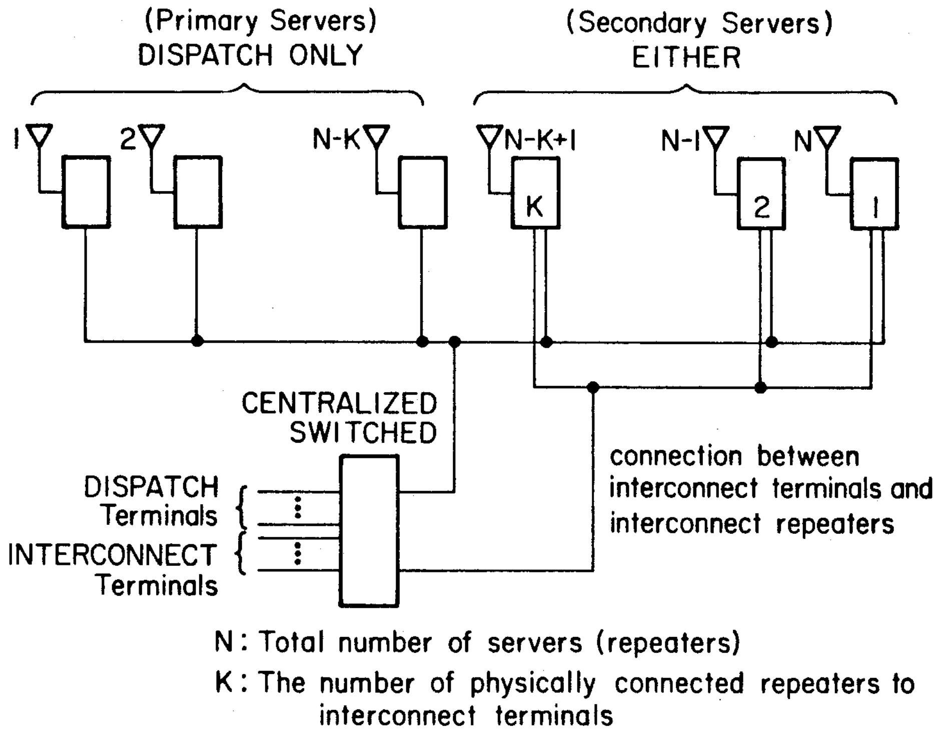


Fig. 4.1 Basic components of a trunked mobile system.

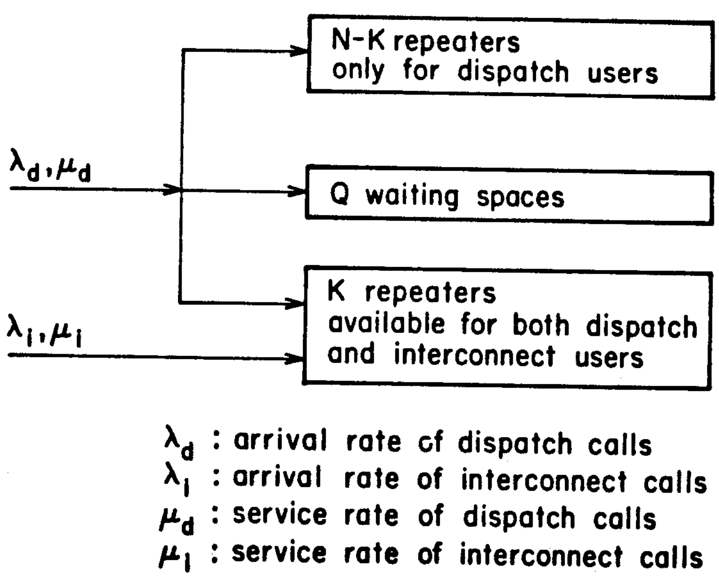


Fig. 4.2 Flow diagram of trunked mobile traffic.

users have the ability to access the system from any one of the repeaters. This is because the "real" purpose of the system is to serve the dispatch users.

In any trunking system which serves both interconnect and dispatch users, it is thus important for the system to be able to intelligently allocate the valuable "air time" resources between the two. "Intelligently" means the decision about the channel resources to be allocated is based on the current traffic demands. The parameters related to the current traffic are measured periodically, and channel resources are re-allocated in order to optimize the system performance until the next measurement.

As an example, in the evening when dispatch traffic is usually light and several channels are likely to be available at any given point in time, it makes sense for interconnect users to be able to utilize these available channels. On the other hand, in the morning, when dispatch traffic is at a peak and the dispatch demand is greater relative to the interconnect demand than usual, it makes sense that channels previously available for interconnect users be switched temporarily to dispatch use only until the peak subsides.

4.2 A Sharing Algorithm for a Trunked Mobile Radio

In this section, we propose a computer-based service sharing algorithm for the Trunked Mobile Radio based on a trade-off between the blocking probability for interconnect users and the average dispatch delay. We should note that in the original Motorola design, both dispatch and interconnect users can be queued. Different from the original strategy, here we assume that only dispatch users can be queued, and the interconnect users are blocked if there is no repeater available.

The shared service algorithm of the original Motorola design is basically to increase or decrease the number of interconnected repeaters to meet the demands of traffic based upon the system owner's preference for grade of service of dispatch versus interconnect call traffic, called a targeted operating point. Therefore, the system parameters such as the the number of interconnected repeaters are adjusted to provide the targeted grade of service. As an example, if the dispatch delay is less than its target value, the number of interconnect calls allowed is increased. If it is greater than its target, then the number of interconnect calls allowed is decreased. Since the aim of this algorithm is to keep the operation in a region where all targeted values are met, the operating point may not be the optimum one as far as minimizing the dispatch delay or the interconnect blockage is concerned. This is because when the targeted operating point is reached, the algorithm doesn't make any modification to the number of interconnected terminals to increase the grade of service. In order to find the best operating point, the average dispatch and interconnect blocking probability must be calculated at all traffic intensities and for all possible number of connections to the interconnect terminals.

In a Trunk Mobile Radio shared by blockable interconnect calls and delayed dispatch calls, a decision on a good operating point involves a trade between the blocking of interconnect calls and the queueing of dispatch calls. Therefore, we aim to find closed-form expressions for interconnect blocking probability and dispatch delay and understand this trade-off both quantitatively and qualitatively. Once closed-form expressions are obtained, the number of interconnect connections required for the targeted dispatch delay and interconnect blocking probability can be computed and necessary adjustment made automatically. Furthermore, the best operating point can be reached by trading off the dispatch delay and the interconnect blocking probability for a given traffic intensity.

Since the real purpose of the original Motorola design is to serve dispatch users, the interconnect blocking probability is usually high. Other control strategies can be used to improve the interconnect blockage performance without substantially degrading the dispatch time-delay performance. The traffic-control strategies discussed in Chapter 3 can be adapted for this purpose. These strategies can be implemented in Motorola's Trunk Mobile Radio by modifying the software (possibly some hardware) of the system common central controller. The performance of these modified strategies is compared with that of the simpler strategy we proposed in this section.

4.2.1 System Traffic Modelling and Analysis

In developing a sharing algorithm for a trunked mobile system, the following model is used:

(i) The arrivals of dispatch and interconnect call requests are independent and are assumed to be coming from memoryless (Poisson) sources of infinite population.

(ii) In both traffic streams, the call durations are exponentially distributed. That is, service times are memoryless and independent of everything else. (The average call duration for dispatch calls will later be assumed much less than that for interconnect calls.)

(iii) We call the interconnect repeaters secondary servers and the remaining (dispatch-only) repeaters primary servers. When a call request from an interconnect user arrives, it is lost (blocked calls cleared) if the secondary servers are busy, i.e., the call is cleared from the system. If the primary servers are busy but there are empty secondary servers, the dispatch arrival is served. Otherwise it is placed in an infinite queue. Fig. 4.2 shows a flow diagram of this trunked mobile traffic.

(iv) The mean service time for interconnect traffic, $1/\mu_i$,is much longer than the mean service time for dispatch traffic, $1/\mu_d$, i.e

$$1/\mu_i >> 1/\mu_d.$$

(v) Incoming dispatch call requests search for an empty repeater starting from the leftmost repeater, whereas incoming interconnect calls start from rightmost repeater, i.e, if there is more than one empty repeater, dispatch calls use the leftmost, and interconnect calls use the rightmost. (The reason for this allocation is to avoid dispatch calls occupying secondary servers when a primary would do. Otherwise, the blocking probability of interconnect calls would increase beyond that intended.)

Using the fact that the average holding time of an interconnect call is much larger than the dispatch holding time, the conditional steady-state probabilities of dispatch users can be simply found by employing the results of basic queueing theory. But in order to find the mean dispatch delay, we need to find the steady-state probabilities of interconnect users as well. It will be seen that the steady-state probabilities of interconnect users depend on the conditional blocking probabilities of interconnect users given in (4.9). So, if we know these conditional blocking probabilities, we can find the unconditional steady-state interconnect probabilities, which then yield both the mean dispatch delay and the blocking probability of an interconnect request. Possible blocking conditions of interconnect calls and how to calculate them are given in Section 4.2.2. The details of the analysis are given below.

We have an N-server queueing system, where N is the total number of repeaters. We will view the number of servers (repeaters) occupied by dispatch users as a random process that varies according to how many servers are assigned to interconnect traffic. We define the interconnect state as the number of interconnect-occupied servers. The interconnect state varies much more slowly than the dispatch state, which is the total number of dispatch-occupied servers. This is simply because we have assumed that

$$\frac{1}{\mu_i} \gg \frac{1}{\mu_d}.$$

It is therefore a good approximation to assume that the time spent in each interconnect state is much longer than the time it takes the number of dispatch requests present in the system to reach steady-state behavior, given the number of interconnect calls in progress. This steady-state approximation of Honig [22] is the basis for our analysis. Given that i servers are currently assigned to interconnect traffic, the probability that there are k dispatch calls in the system is simply found by analyzing the resulting M/M/$(N-i)$ queueing system [11].

We define the conditional state probabilities of dispatch traffic as

$P_{d/i}(k) = \mathbf{Pr}($ there are k dispatch users in the system, given that i servers are occupied by interconnect calls$)$.

From the assumed equilibrium at steady-state, we have

$$\lambda_d P_{d/i}(0) = \mu_d P_{d/i}(1)$$
$$\lambda_d P_{d/i}(1) = 2\mu_d P_{d/i}(2)$$
$$\vdots$$
$$\lambda_d P_{d/i}(N-i-1) = (N-i)\mu_d P_{d/i}(N-i).$$

These equations yield

$$P_{d/i}(k) = P_{d/i}(0) a_d^k \frac{1}{k!}, \qquad 0 \le k \le N-i. \tag{4.1}$$

For $k \ge N-i$, however, the arriving dispatch requests are queued, so the steady-state probabilities satisfy

$$\lambda_d P_{d/i}(N-i) = (N-i)\mu_d P_{d/i}(N-i+1)$$
$$\lambda_d P_{d/i}(N-i+1) = (N-i)\mu_d P_{d/i}(N-i+2)$$
$$\vdots$$

These equations imply that

$$P_{d/i}(k) = P_{d/i}(0)a_d^k \frac{(N-i)^{N-i-k}}{(N-i)!}, \qquad k \geq N-i, \tag{4.2}$$

where

$$P_{d/i}(0) = \left[\sum_{k=0}^{N-i-1} \frac{{a_d}^k}{k!} + \frac{1}{(N-i)!}{a_d}^{N-i} \frac{1}{1-\frac{a_d}{(N-i)}} \right]^{-1} \tag{4.3}$$

and

$$a_d = \frac{\lambda_d}{\mu_d}$$

is the dispatch traffic.

Given that there are i assigned interconnect calls, the mean number of queued dispatch requests is readily found by using (4.2) and (4.3):

$$\begin{aligned} E(q/i) &= \sum_{q=1}^{\infty} q P_{d/i}(N-i+q) \\ &= \frac{P_{d/i}(0)}{(N-i)!} a_d^{N-i} \sum_{q=1}^{\infty} q \frac{a_d^q}{(N-i)^q}, \end{aligned}$$

$$E(q/i) = \frac{P_{d/i}(0)}{(N-i)!} a_d^{N-i} \frac{\frac{a_d}{(N-i)}}{\left[1-\frac{a_d}{(N-i)}\right]^2}. \tag{4.4}$$

From this, the mean number of queued dispatch requests is found as follows:

$$E(q) = \sum_{\ell=0}^{\mathbf{K}} P_I(\ell) E(q/\ell) \tag{4.5}$$

where $\mathbf{K}$ is the number of interconnect repeaters and $P_I(\ell)$ is the probability of having ℓ interconnect users in the system. This probability can be found by using the method of entrance probability described in [22], which we now describe.

$P_I(\ell)$ depends not only on the previous interconnect state but also on the number of dispatch calls occupying the secondary servers. Assuming there are ℓ interconnect users in the system, where $\ell < K$, an arriving interconnect request is blocked if the remaining $K-\ell$ secondary servers are all occupied by dispatch calls. Therefore, the transition rate to the $(\ell+1)^{th}$ state is the arrival rate times the entrance probability, which we define as

$\phi_\ell = \mathbf{Pr}($ a new interconnect request is able to be served, given that there are ℓ interconnect calls in the system), for $0 \leq \ell \leq K$. (ϕ_K is, of course, 0.)

The rest of the analysis is similar to the analysis of an M/M/K blocking system except that the arrival rates in each state ℓ are decreased by the factor ϕ_ℓ.

Assuming exponential service times and using the equilibrium equations, $P_I(\ell)$ can be found as follows:

$$\lambda_i P_I(0)\phi_0 = \mu_i P_I(1)$$
$$\lambda_i P_I(1)\phi_1 = 2\mu_i P_I(2)$$
$$\vdots$$
$$\lambda_i P_I(K-1)\phi_{(K-1)} = K\mu_i P_I(K).$$

Therefore,

$$\begin{aligned} P_I(\ell) &= \frac{P_I(0)}{\ell!} a_i^\ell \phi_0\phi_1 \ldots \phi_{\ell-1} \\ &= \frac{P_I(0)}{\ell!} a_i^\ell \prod_{k=0}^{\ell-1} \phi_k, \end{aligned} \tag{4.6}$$

where

$$a_i = \frac{\lambda_i}{\mu_i},$$

the offered interconnect traffic, and

$$P_I(0) = \left[\sum_{m=0}^{K} \frac{a_i^m}{m!} \prod_{j=0}^{m-1} \phi_j \right]^{-1}. \tag{4.7}$$

Finally, substituting (4.7) into (4.6),

$$P_I(\ell) = \frac{\frac{a_i^\ell}{\ell!} \prod_{j=0}^{\ell-1} \phi_j}{\sum_{m=0}^{K-1} \frac{a_i^m}{m!} \prod_{j=0}^{m-1} \phi_j}. \tag{4.8}$$

The blocking probability of interconnect users can be found from equation (4.8) by employing the following equation:

$$P_B = \sum_{\ell=0}^{K-1} P_{B/\ell} P_I(\ell) + P_I(K) \tag{4.9}$$

where

$P_{B/\ell} =$ **Pr** (an arriving interconnect request is blocked, given that there are exactly ℓ interconnect calls in the system).

Here,

$$P_{B/\ell} = 1 - \phi_\ell \tag{4.10}$$

Therefore, (4.9) can be written by employing (4.10) as

$$P_B = \sum_{\ell=0}^{K-1} (1 - \phi_\ell) P_I(\ell) + P_I(\mathbf{K}). \tag{4.11}$$

As is seen from (4.5), (4.8) and (4.11), the blocking probability for interconnect users and the expected number of dispatch users waiting for service depend on the entrance probabilities ϕ_ℓ or on the conditional probabilities $P_{B/\ell}$. The next section describes how to calculate these conditional blocking probabilities, thus solving our problem.

4.2.2 Blocking Conditions for Interconnect Traffic

It is clear that an incoming interconnect call is blocked precisely when all the secondary servers are busy. The secondary servers are busy if

(i) there are already K interconnect calls in the system, where K is the total number of repeaters connectable to interconnect calls, or

(ii) the number of dispatch users, d, in the system is greater than or equal to $N - i$, when there are i interconnect users in the system and $0 \leq i \leq K - 1$, or

(iii) the number of dispatch users, d, in the primary (dispatch-only) repeaters is less than $N - K$, but there is at least one dispatch user in the secondary repeaters, and the rest of the secondary repeaters are occupied by interconnect calls.

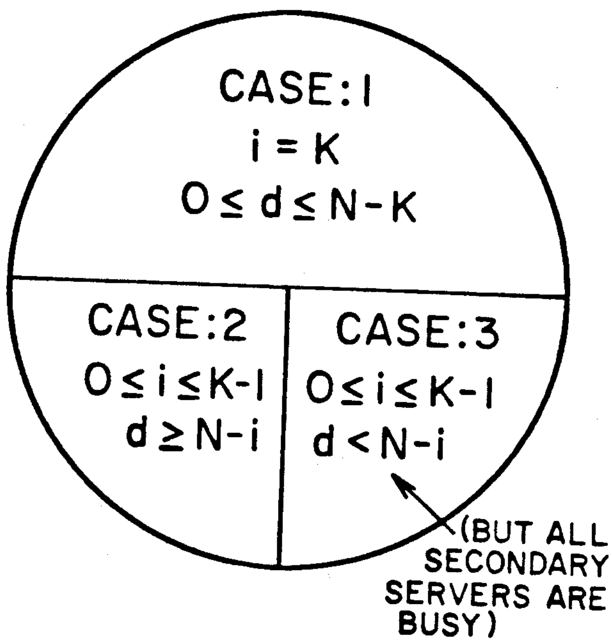


Fig. 4.3 Possible blocking cases for interconnect users when all the secondary repeaters are busy.

The three possible blocking cases for interconnect users are illustrated in Fig. 4.3. Obviously, these are disjoint events. One possible evolution of these blocking conditions is illustrated below. In the following, **D** corresponds to a dispatch user, I corresponds to an interconnect user, and '.' corresponds to empty server. Here time increases as the index j on t_j increases; $N = 15$, $K = 5$.

(i) $i = K$

i : the number of interconnect users in the system

$t_1 \rightarrow$ **D . D D D D D D D D I I I I I**

$t_2 \rightarrow$ **D D D D D D D D D D I I I I I**

$t_3 \rightarrow$ **D D D D D D D D D . I I I I I**

(ii) $d \geq N - i$

d: number of dispatch users in the system; $i = 3$

$t_1 \rightarrow$ **D D D D D D D D D D I I D D I**

$t_2 \rightarrow$ **D D D D D D D D D D I I D D I**

(iii) $d + i = K$ in the secondary repeaters, while $d < N - K$ in the primary repeaters; $i = 3$

$t_1 \rightarrow$ **D D D D D D D D D D I D I D I**

$t_2 \rightarrow$ **D D D D D D D D . D I D I D I**

$t_3 \rightarrow$ **D D D D D . D D . D I D I D I**

$t_4 \rightarrow$ **D D D D D D D D . D I D I D I**

If all the repeaters were physically connected to the interconnect terminals, i.e., $N = K$, then the only blocking condition would be $d \geq N - i$ as described in (ii). This is the case where all arrivals can access any repeater. Such a case has been analyzed by Honig [22]. But for a trunked mobile system the first and the third conditions must also be considered.

Define the probabilities corresponding to the events described in (i), (ii), and (iii) as follows:

$P_1 =$ Pr(there are K interconnect calls in the system)

$P_{2/i} =$ Pr(the number of dispatch users, d, is greater than or equal to $N - i$ given that there are i interconnect calls in the system)

(Thus, in our prior notation, $P_{2/i} = P_{d/i}(d \geq N - i)$.)

$P_{3/i} =$ Pr(The number of dispatch users in the primary is less than $N - K$, given that there are i interconnect calls and $K-i$ dispatch calls in the secondary repeaters, where $i > 0$)

$P_{2/i}$ is obtained easily by using (4.2):

$$\begin{aligned} P_{2/i} &= \sum_{j=N-i}^{\infty} P_{d/i}(j) \\ &= P_{d/i}(0)\frac{(N-i)^{N-i}}{(N-i)!}\frac{1}{1-\frac{a_d}{(N-i)}}. \end{aligned} \tag{4.12}$$

(Note that $\frac{a_d}{(N-i)} < 1$, otherwise an infinite queue forms in steady-state). Finally, P_1 is found from

$$P_1 = P_I(K) \tag{4.13}$$

For definition, this is the probability that there are K interconnect calls in the system.

The conditional blocking probability is written by using the previous definitions as

$$P_{B/i} = P_{2/i} + P_{3/i}, \qquad 0 \leq i < K, \tag{4.14}$$

so the entrance probability ϕ_i is

$$\phi_i = 1 - P_{2/i} - P_{3/i}, \qquad 0 \le i < K. \tag{4.15}$$

Now we will calculate $P_{3/i}$ explicitly. At the start, it was assumed that if there is more than one empty repeater, an arriving dispatch call request uses the leftmost repeater. Therefore, if the event described in (iii) occurs, that is $d + i = K$ in the secondary repeaters, while at least one primary repeater is empty, then in one of the past repeater states the number of dispatch users must have been equal to $N - i$. The following diagram shows the state of the repeaters at times $t_1, t_2, \ldots, t_6$ and illustrates the blocking condition described by (iii) after time t_1:

$t_1 \rightarrow$ **D D D D D D D D D D I D I D I** *(ii)*

$t_2 \rightarrow$ **D D D D D D D D • D I D I D I** *(iii)*

$t_3 \rightarrow$ **D D D D D • D D • D I D I D I** *(iii)*

$t_4 \rightarrow$ **D D D D D • D • • D I D I D I** *(iii)*

$t_5 \rightarrow$ **D D D D D D D • • D I D I D I** *(iii)*

$t_6 \rightarrow$ **D D D D D D D D • D I D I D I** *(iii)*

Fig. 4.4 State of repeaters in blocking condition (iii)

At times t_2 through t_6, the interconnect requests are blocked, even though there are empty primary servers. The system is in a blocking state between t_2 and t_6 as is described in (iii). But before entering this state, the system comes from another blocking state as described in (ii). At time t_1, the system is in such a state and $d = N - i$. When the system is in blocking condition (iii), there are i interconnect and $K - i$ dispatch users in the secondary servers and they therefore occupy all secondary servers. On the other hand, the number of dispatch users in the primary servers may vary from 0 to $N-K-1$. Whenever a departure occurs in the secondary servers, the system gets out of the blocking state.

Since all secondary servers are busy, and there is no dispatch user in the queue for this particular case (iii), the number of dispatch users in the primary servers describes the system completely. So, we can merely define the state of the repeaters in case (iii) as the number of dispatch users in the primary servers. We let $X(t)$ denote this state at time t. The random process $X(t)$ forms a continuous-time Markov chain, because clearly the future value of $X(t)$ depends on the past only through its current value. Fig. 4.5 is the state-transition-rate diagram for this process. Concentrating on state S_j, we observe that it can be entered only from state $S_{(j-1)}$ or state $S_{(j+1)}$, and similarly, it can be left only by entering state $S_{(j-1)}$ or state $S_{(j+1)}$, or by getting out of the blocking conditions altogether, i.e., by a departure from the secondary repeaters. At steady-state, the rate at which the system leaves S_j must be equal to the rate at which the system goes into that state, because $X(t)$ is an ergodic Markov chain and therefore its steady-state probabilities exist.

Now define the conditional state probabilities as follows, for $0 \leq i \leq K$:

$Q_{j/i}$ =**Pr**(There are j dispatch users in the primary, given that all secondary servers are busy and there are i interconnect users in the secondary)

Then for $j = N - K$,

$$\begin{aligned} Q_{j/i} &= Q_{(N-K)/i} \\ &= P_{d/i}(N-i) \\ &= P_{d/i}(0)\frac{{a_d}^{(N-i)}}{(N-i)!}. \end{aligned}$$

The blocking condition described in (iii) occurs when the system is in one of the states described above and $j < N - K$. Therefore

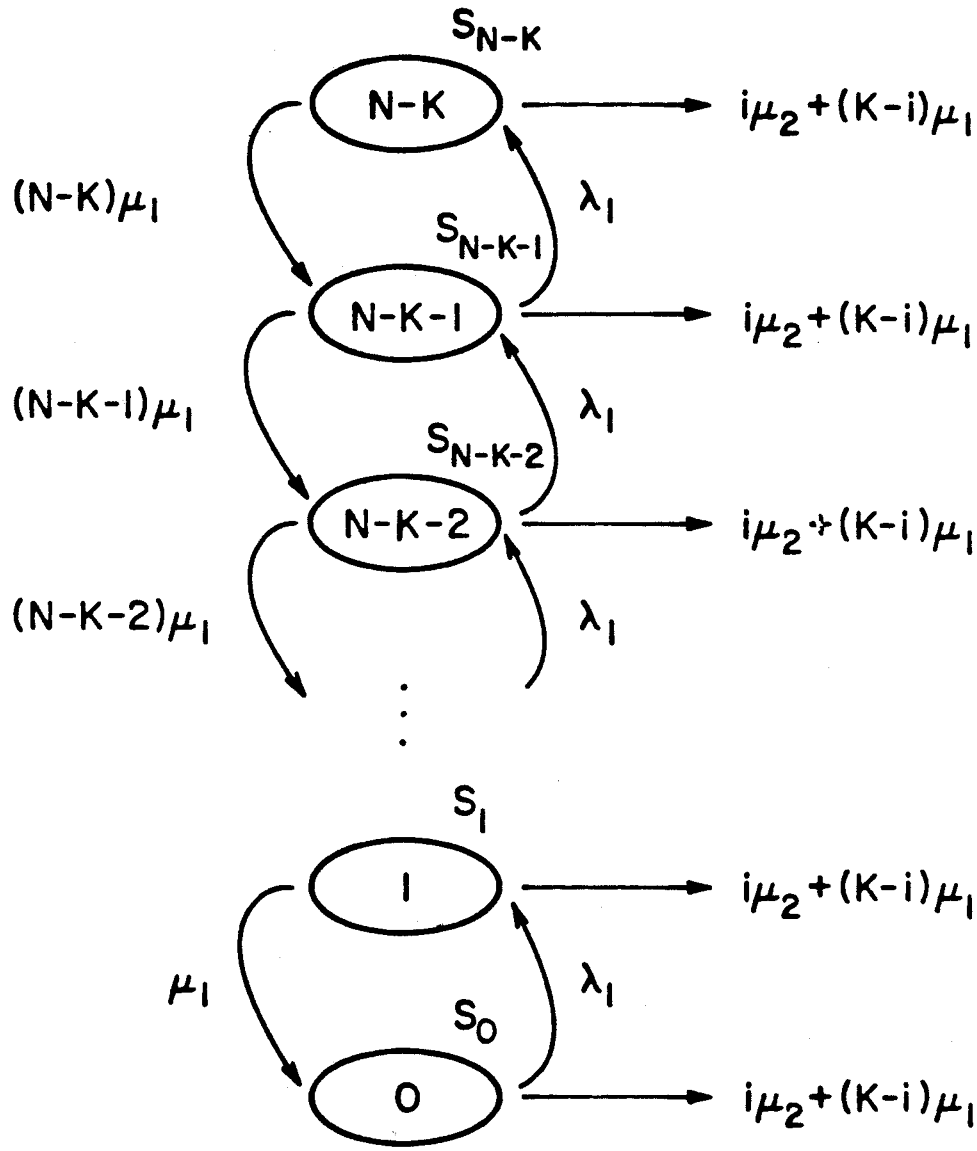


Fig. 4.5 State transition diagram when secondary serves are busy.

$$P_{3/i} = \sum_{m=0}^{N-K-1} Q_{m/i}, \qquad 0 \le i < K. \tag{4.20}$$

Specifically, if we focus on state S_j we observe that the rate at which probability "flows" <u>into</u> this state at time t is given by

<u>Flow rate into</u> $S_j = (j+1)\mu_d Q_{(j+1)/i} + \lambda_d Q_{(j-1)/i}$

whereas the flow rate <u>out</u> of that state at time t is given by

<u>Flow rate out of</u> $S_j = (j\mu_d + A_i)Q_{j/i}$

where $A_i = \lambda_d + i\mu_i + (K-i)\mu_d$.

Clearly, these two rates must be equal to each other, that is,

$$\mu_d Q_{1/i} = A_i Q_{0/1}$$

$$2\mu_d Q_{2/i} + \lambda_d Q_{0/i} = (\mu_d + A_i)Q_{1/i}$$

$$\vdots$$

$$(N-K)\mu_d Q_{(N-K)/i} + \lambda_d Q_{(N-K-2)/i} = ((N-K-1)\mu_d + A_i)Q_{(N-K-1)/i}.$$

Note that $Q_{(N-K)/i} = P_{d/i}(N-i)$, because when the number of dispatch users in the primary is $N-K$ this means that all the servers are occupied, and this corresponds to the blocking case described in (ii).

In matrix form, these relationships can be written as

$$\begin{pmatrix} \delta_0 & -\mu_d & 0 & \dots & 0 \\ -\lambda_d & \delta_1 & -2\mu_d & \dots & 0 \\ 0 & -\lambda_d & \delta_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \delta_M \end{pmatrix} \begin{pmatrix} Q_{0/i} \\ Q_{1/i} \\ Q_{2/i} \\ \vdots \\ Q_{M/i} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ (M+1)\mu_d Q_{(M+1)/i} \end{pmatrix} \tag{4.21}$$

where $M = N - K - 1$ and $\delta_j = j\mu_d + A_i$. This system of linear equations is easily solved by using the Gauss elimination method. The result is

$$Q_{r/i}.C_r = (r+1)\mu_d Q_{(r+1)/i}, \qquad 0 \le r \le N-K-1, \qquad (4.22),$$

where C_r is given as

$$C_r = (\,\delta_r - \tfrac{r\mu_d\lambda_d}{C_{r-1}}\,) \qquad , C_0 = \delta_0. \qquad (4.23)$$

Thus,

$$Q_{(N-K-1)/i} = \frac{(N-K)\mu_d}{C_{N-K-1}} P_{d/i}(N-K)$$

$$Q_{(N-K-2)/i} = \frac{(N-K)(N-K-1)\mu_d^2}{C_{N-K-1}C_{N-K-2}} P_{d/i}(N-K)$$

$$\vdots$$

$$Q_{0/i} = \frac{(N-K)!\mu_d^{(N-K)}}{C_{N-K-1}C_{N-K-2}\dots C_0} P_{d/i}(N-K)$$

where

$$P_{d/i}(N-K) = P_{d/i}(0)\frac{a_d^{N-K}}{(N-K)!}.$$

$P_{3/i}$ can now be determined by employing the equation (4.20):

$$P_{3/i} = \sum_{m=0}^{N-K-1} Q_{m/i} \qquad 0 \le i < K.$$

4.2.3 Outline of the results and conclusions

Our aim is to find the blocking probability of interconnect users and the average dispatch delay as a function of the number of interconnected repeaters. First, $P_{3/i}$,

$P_i(\ell)$, and ϕ_i are computed, and then the blocking probability of interconnect users is obtained by making proper substitutions.

An outline of this computation algorithm will now be presented. The conditional blocking probability of interconnect users are given as

$$P_{B/i} = \left\{ \begin{array}{cc} P_{2/i} + P_{3/i}, & 0 \leq i < K \\ P_I(K), & i = K \end{array} \right\}. \tag{4.24}$$

$P_{3/i}$ is computed as described in the previous section, equation (4.20), and ϕ_i is simply found by the definition

$$\phi_i = 1 - P_{2/i} - P_{3/i}, \tag{4.25}$$

one minus the blocking probability of interconnect users when the number of interconnect users i in the system less than K. $P_{2/i}$ is given as in equation (12):

$$P_{2/i} = P_{d/i}(0)\frac{(N-i)^{N-i}}{(N-i)!}\frac{1}{1-\frac{a_d}{(N-i)}}. \tag{4.26}$$

$P_I(K)$ is found using the method described in Section 4.2.1 as

$$P_I(K) = \frac{\frac{a_i^K}{K!}\prod_{j=0}^{K-1}\phi_j}{\sum_{m=0}^{K-1}\frac{a_i^m}{m!}\prod_{j=0}^{K-1}\phi_j} \tag{4.27}$$

After computing the steady-state probability of interconnect users, $P_I(i)$, and the conditional blocking probability, $P_{B/i}$, of interconnect users, the blocking probability P_B is obtained by employing equation (4.9) :

$$P_B = \sum_{\ell=0}^{K-1} P_{B/\ell}P_I(\ell) + P_I(K) \tag{4.28}$$

where $P_I(\ell)$ is

$$P_I(\ell) = \frac{\frac{a_i^\ell}{\ell!}\prod_{j=0}^{\ell-1}\phi_j}{\sum_{m=0}^{K-1}\frac{a_i^m}{m!}\prod_{j=0}^{K-1}\phi_j}. \tag{4.29}$$

The mean number of queued dispatch requests is found as:

$$E(q) = \sum_{\ell=0}^{K} P_I(\ell)E(q/\ell), \tag{4.30}$$

where $E(q/\ell)$ is

$$E(q/\ell) = P_{d/i}(0)\frac{a_d^{N-\ell}}{(N-\ell)!}\frac{\frac{a_d}{(N-\ell)}}{\left[1-\frac{a_d}{(N-\ell)}\right]^2}, \tag{4.31}$$

and $P_I(\ell)$ is as given in equation (4.29). The average dispatch delay is then obtained by employing Little's formula:

$$W = E(q)/\lambda_d, \tag{4.32}$$

where W is the mean delay and $E(q)$ is the mean number of dispatch users in the queue given in (4.30). Fig. 4.6 describes the hierarchy of our computations.

Fig. 4.7 compares the interconnect blocking probability with the average delay for various traffic levels as computed from equations (4.28) and (4.32) using the method just described. The number of repeaters connectable to interconnect calls, K, is shown on the abscissa. The ordinate corresponds to blocking probability or mean dispatch delay, respectively. Obviously, when $K = 0$, the blocking probability of interconnect users is 1, and the mean dispatch delay is at its minimum. In this case, all repeaters are reserved for dispatch traffic. As we increase the number of connections to interconnect terminals, K, interconnect users start to access the repeaters that were previously reserved for dispatch traffic and the blocking

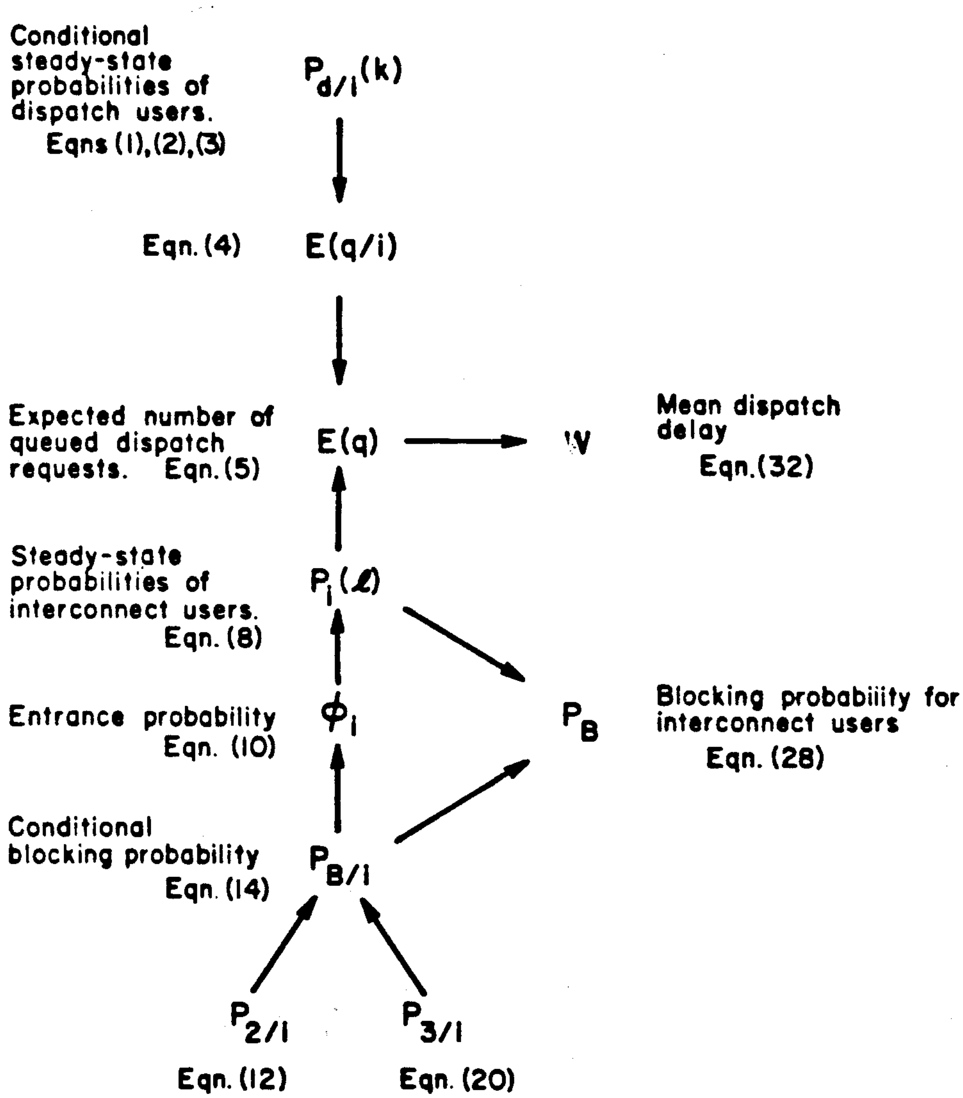


Fig. 4.6 Hierarchy in computation.

probability of interconnect users therefore decreases, as is apparent in the figure. Increasing the dispatch traffic increases the mean dispatch delay, which shifts the delay curve left and the blocking probability curve up, as we see. When the amount of dispatch traffic approaches the number of repeaters reserved for dispatch users, we see that the mean delay increases rapidly, as we would expect. The service rates used are $\mu_d = 4.01$ and $\mu_i = 0.2$, which correspond to the mean dispatch service time of 14.96 seconds and interconnect service time of 5 minutes. The arrival rates for the first comparison are $\lambda_d = 40$ customers per minute and $\lambda_i = 2$ customers per minute, for the second comparison $\lambda_d = 44$ and for the last one $\lambda_d = 48$.

At the beginning of Section 4.2 we proposed that a sharing algorithm for a trunked mobile system could be developed which would trade off the interconnect blocking probability against the average dispatch delay. The blocking probability for interconnect users and mean dispatch delay are found as described in the previous section. As a result of the comparison of these quantities, it is seen that the interconnect blocking probability decreases almost linearly as K, the number of secondary servers, increases, and it is independent of dispatch traffic for small values of K of the number of interconnected repeaters. When K is small and the dispatch traffic is light, secondary repeaters are not likely to be occupied by dispatch users, so the probability of an interconnect user being served is not affected by dispatch traffic and is thus nearly proportional to the number of secondary servers. As K increases, the dispatch users start to occupy the secondary servers, so the probability of being blocked depends not only on the number of servers but also on the number of dispatch calls in the secondary servers. Thus, the decrease in blocking probability for interconnect users is no longer linear.

Since the average holding time of an interconnect call is much larger than the average holding time of a dispatch call, the interconnect state varies very slowly.

As a result of this, when the number of interconnect users in the system is equal to K, and the dispatch traffic is greater than $N - K$, the number of dispatch users waiting for service starts to increase, and queue peaks occur as expected. As must be, this causes an increase in the average delay of dispatch calls. However, the mean delay is almost negligible when the number of primary servers is greater than the amount of dispatch traffic, because the probability of finding an empty repeater is very high for an incoming dispatch call. For example, when the dispatch traffic is 9.975 Erlangs and $K = 10$, the mean delay is 26 seconds as is seen from the second curve of Fig. 4.7. But if K is decreased by one, to $K = 8$, that is, if the number of primary servers is increased by one, the mean delay suddenly drops to 2.4 seconds; for $K = 8$ it is only 1.35 seconds.

As a result, for small values of K, as K increases, the blocking probability decreases considerably, and the mean delay remains negligible. But as soon as $N - K$ approaches the value of the dispatch traffic, the mean delay becomes very large, whereas the blocking probability for interconnect calls does not vary much.

So, in order to decrease the blocking probability of interconnect users and to keep the mean dispatch delay at a tolerable level, we conclude that the number of secondary repeaters should be the largest possible value smaller than $(N - a_d)$, where N is the total number of repeaters and a_d is the dispatch traffic.

As an example, let $a_d = 9.975$ Erlangs, $N = 20$ repeaters, $N - a_d = 10.025$. According to our criterion, the value of K must be at most 10, otherwise the dispatch queue blows up. In other words, a good operating point is to provide $N - a_d$ (rounded down) interconnect repeaters. This could have been approximately predicted in advance, as our above heuristics show.

4.3 Comparison with other Traffic Control Strategies:

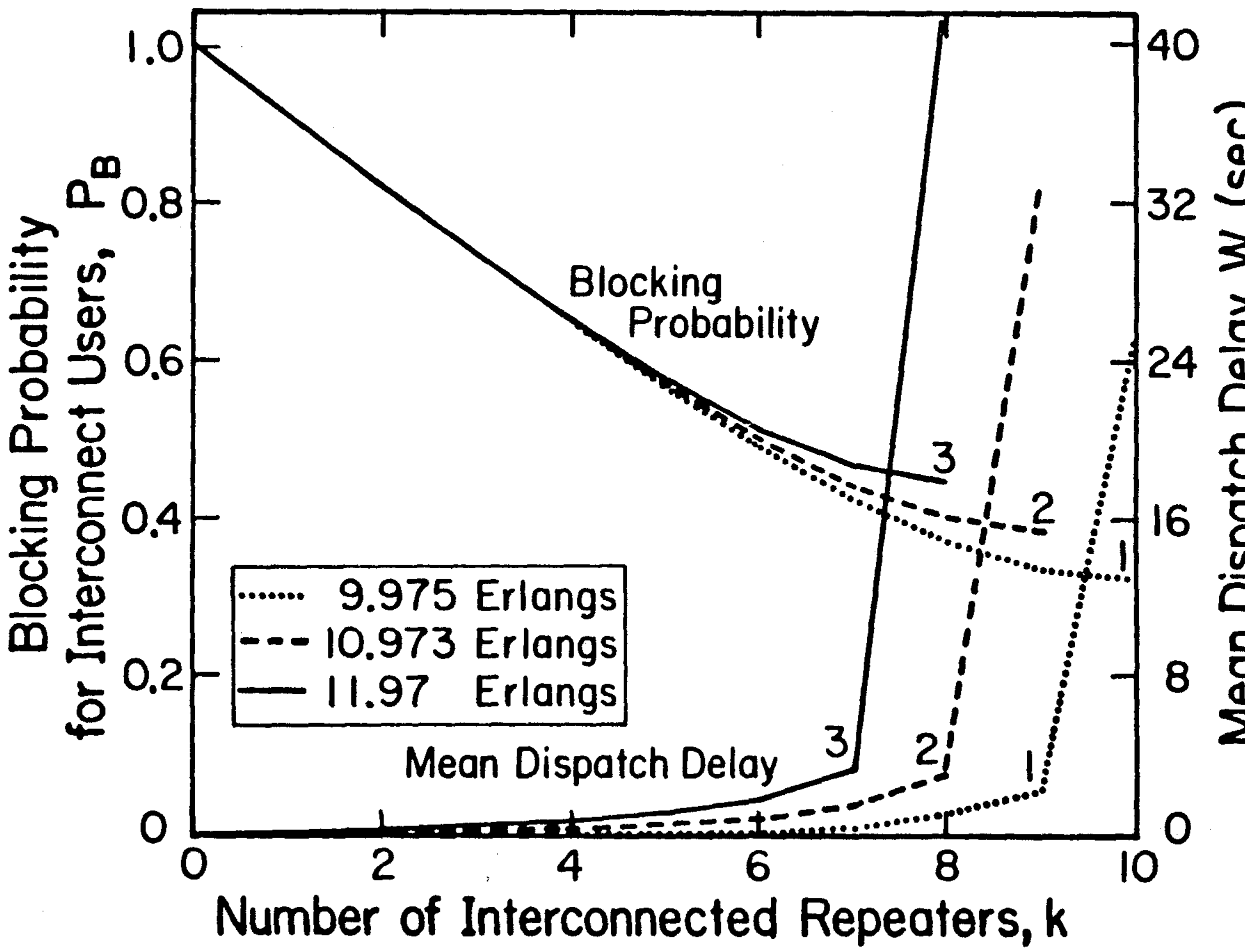


Fig. 4.7 The interconnect blocking probability and mean dispatch delay as a function of interconnected repeaters.

Several control strategies can be developed for the sharing of repeaters between dispatch and interconnect traffic. In the original strategy proposed by Motorola [28], the number of repeaters that are physically connected to interconnect terminals is controlled by an operating system according to traffic intensity. The remaining repeaters (primary) are reserved for dispatch use only. The dispatch users access the system starting from primary repeaters, and they use the secondary repeaters only if primary servers are occupied. The analysis of this scheme was given in the previous section assuming that interconnect holding time is much larger than dispatch holding time. In this section we will investigate three other control strategies by using key-state approach. The third and fourth strategies are adapted from voice-data integration networks and have already been analyzed in Chapter 3. The analysis of the second strategy is also very similar to the analysis of the others; only the coefficients of the difference equations obtained for the key-state coefficients are different.

In the second control strategy, all the repeaters have physical connections to the interconnect calls, but the number of interconnect terminals in the system at any given point in time can not exceed a threshold, say K. In this strategy, when all the repeaters are busy and there is a departure from one of the primary repeaters, interconnect traffic can utilize that repeater. This use is not possible in the first strategy. Therefore, the blockage performance of this new strategy must be better than in the original one proposed by Motorola. We call this the non-preemptive movable-boundary (NPMB) strategy, for dispatch users occupying the secondary servers are not preempted by interconnect calls, and there is no fixed boundary between the repeaters used by dispatch and interconnect calls.

The third strategy, called the preemptive movable boundary (PMB) strategy, is adapted from voice-data integration networks. The only difference between this strategy and the previous one is that if the number of interconnect calls in the system is less than the prespecified threshold and all the servers are busy, a dispatch call is preempted, that is, caused to be disconnected by an arriving interconnect call.

Finally, we consider the first-come, first-served (FCFS) strategy, where the repeater assignments are made according to first-come, first-served priority. Therefore, both dispatch and interconnect users can get service without any restriction as long as there is at least one empty repeater. The dispatch arrivals are queued while interconnect arrivals are blocked when all the repeaters are busy.

We assume, as usual, in all the strategies, that the arrival processes are Poisson, and the service distributions are exponential. This allows us to model the system as a Markovian queueing network. We can therefore define the states as:

$$S_{di} = [d, i]; \quad 0 \le d, \quad 0 \le i \le K, \tag{4.33}$$

where d is the number of dispatch calls and i is the number of interconnect calls in the system. The state-transition rate diagrams of the FCFS, NPMB and PMB control strategies are shown in Fig. 4.8. The solid-line transitions correspond to the state-transition diagram of the FCFS case when the total number of repeaters N is 2. The dashed line transitions together with the solid line transitions correspond to the NPMB case for $N = 3$ and $K = 2$. Finally, including the dotted line transitions we obtain the state-transition rate diagram of the PMB case for $N = 3$ and $K = 2$.

The similarities in the state-space structure of these control strategies make it easy to extend the results obtained for them. The analysis given below is just a brief summary of the analysis given in Chapter 3. The performance of FCFS, NPMB and PMB are compared in Figs. 4. 9 - 4. 13.

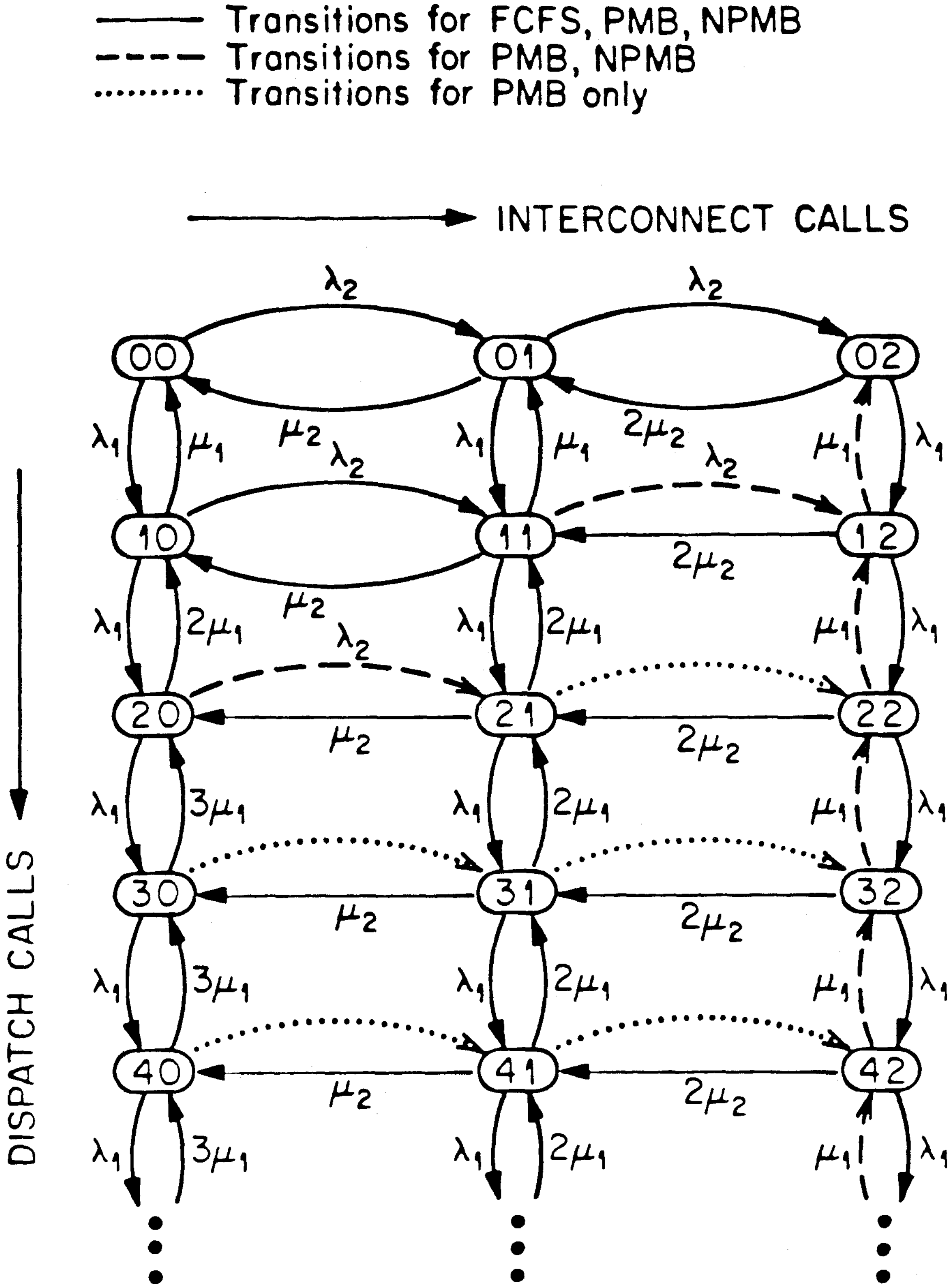


Fig. 4.8 State transition-rate diagram for FCFS, NPMB and PMB schemes.

The key states in the NPMB and MB cases are found to be $\{S_{0i}, i = 0, \ldots, K\}$. These are the first row-states. (In the FCFS case, however, S_{0K} is not necessarily a key state.) For the steady-state probability $P_i(d)$, we obtain the following representation:

$$P_i(d) = \sum_{\ell=0}^{K} c_{i,\ell}(d) P_\ell(0). \tag{4.34}$$

The balance equations of the $(K-m)^{th}$ column states $S = [K-m, d-1]$ yield the following difference equation:

$$\bullet \quad P_{K-m}(d) + a_m P_{K-m}(d-1) + b_m P_{K-m}(d-2) =$$
$$d_m P_{K-m-1}(d-1) + e_m P_{K-m+1}(d-1), \quad d \geq M. \tag{4.35}$$

Here, for the FCFS strategy, $M = m+2$, $N = K$, $1 \leq m \leq K$ and

$$a_m = -\frac{\lambda_1 + m\mu_1 + (K-m)\mu_2}{m\mu_1}; \quad b_m = \frac{\lambda_1}{m\mu_1};$$

$$e_0 = 0; \quad e_m = -\frac{(K-m+1)\mu_2}{m\mu_1}; \quad d_m = 0. \tag{4.36}$$

For the PMB strategy, $M = N - K + m + 1$, $0 \leq m \leq K$ and,

$$a_m = -\frac{\lambda_1 + (M-1)\mu_1 + (K-m)\mu_2 + (1-\delta_m)\lambda_2}{(M-1)\mu_1};$$

$$b_m = \frac{\lambda_1}{m\mu_1}; \quad d_K = e_0 = 0;$$

$$e_m = -\frac{(K-m+1)\mu_2}{(M-1)\mu_1}; \quad d_m = -\frac{\lambda_2}{(M-1)\mu_1}, \tag{4.37}$$

while for the NPMB strategy, $M = N - K + m + 1$, $0 \leq m \leq K$ and,

$$a_m = -\frac{\lambda_1 + (M-1)\mu_1 + (K-m)\mu_2}{(M-1)\mu_1}; \quad e_0 = 0; \quad d_m = 0;$$

$$b_m = \frac{\lambda_1}{(M-1)\mu_1}; \quad e_m = -\frac{(K-m+1)\mu_2}{(M-1)\mu_1}. \tag{4.38}$$

It is not possible to obtain the steady-state probabilities directly from equation (4.35), because the difference equation is valid only for $d \geq M$, and the initial conditions are unknown. But equation (4.35) can be employed to get difference equations for the entries of the coefficient vector, the $c_{K-m}(d)$'s. Equation (4.35) is satisfied if:

- $c_{K-m,\ell}(d) + a_m c_{K-m,\ell}(d-1) + b_m c_{K-m,\ell}(d-2) =$

$$d_m c_{K-m-1}(d-1) + e_m c_{K-m+1,\ell}(d-1) \tag{4.39}$$

for $d \geq M, \quad \ell = \{0, 1, \ldots, K\}$. The initial conditions are obtained by decomposing the $P_{K-m}(d)$'s in terms of key-state probabilities starting from $d = 0$ to $d = m+1$. For example:

$$c_{K-m,\ell}(0) = \begin{cases} 1 & \text{if } \ell = K-m \\ 0 & \text{if } \ell \neq K-m \end{cases}; \tag{4.40}$$

$$c_{0,0}(1) = \frac{\lambda_1 + \lambda_2}{\mu_1}, \qquad c_{0,1}(1) = \frac{\mu_2}{\mu_1}, \ldots \tag{4.41}$$

It can be shown that the solution for the key-state coefficients are as follows:

$$c_{K-m,\ell}(d) = \sum_{j=0}^{2K+2} A_{K-m,\ell,j} z_j^d \quad d \geq M \tag{4.42}$$

where the z_j are functions of the coefficients of the difference equation (in FCFS case these are the roots of the corresponding characteristic equation) and the $A_{K-m,\ell,j}$ are obtained from these balance equations applied to states $S = [K-m, M]$ and $S = [K-m, M-1]$. Here we make use of the initial conditions at $d = M-1$ and $d = M-2$.

From (4.34) and (4.42) the steady-state probabilities are now expressed as follows:

$$P_{K-m}(d) = \sum_{\ell=0}^{K} \sum_{i=0}^{2K+2} A_{K-m,i,\ell}\, z_i^d\, P_\ell(0), \quad d \geq M. \tag{4.43}$$

The key-state probabilities are related to empty-state probabilities by employing the stability requirements of the queueing network, as in (3.110):

$$P_\ell(0) = \sigma_\ell P_0(0). \tag{4.44}$$

Finally, the empty state probability $P_0(0)$ is found explicitly by using the normalization condition. The final form of the equilibrium probabilities are:

$$P_{K-m}(d) = P_0(0) \sum_{j=0}^{2K+2} z_j^d B_{K-m,j}, \quad d \geq M, \tag{4.45}$$

where

$$B_{K-m,j} = \sum_{\ell=0}^{2K+2} A_{K-m,\ell,j}\sigma_\ell. \tag{4.46}$$

By using the closed-form expression found in (4.46) for the equilibrium probabilities, the expressions for the mean dispatch delay and the blocking probability of the interconnect calls can be obtained easily.

In Figs. 4.9, 4.10 and 4.11, the normalized queueing time, i.e, the queueing time relative to the transmission time, is plotted as a function of dispatch traffic for the FCFS, NPMB and PMB cases, respectively. The number of repeaters, N, is 3, for all the cases, and the number of repeaters available for interconnect users K, is 1, for the NPMB and PMB cases. The offered interconnect traffic, a_i, is equal to 1 Erlang in all cases. Since the corresponding fixed boundary system for dispatch traffic is an M/M/2 queueing system, the results obtained for mean dispatch delay have been compared to the results of M/M/2 case. In Figs. 4.12 and 4.13, the blocking probability of interconnect calls in the FCFS and NPMB cases is compared to the blocking probability of PMB case.

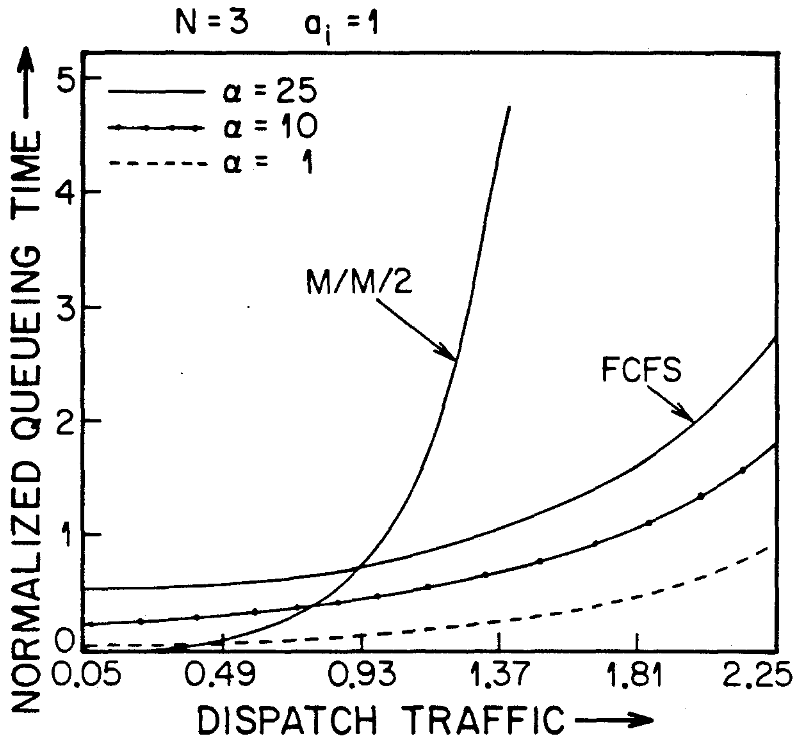


Fig. 4.9 The normlized queueing time as a function of dispatch traffic for FCFS and M/M/2 schemes for three different values of α.

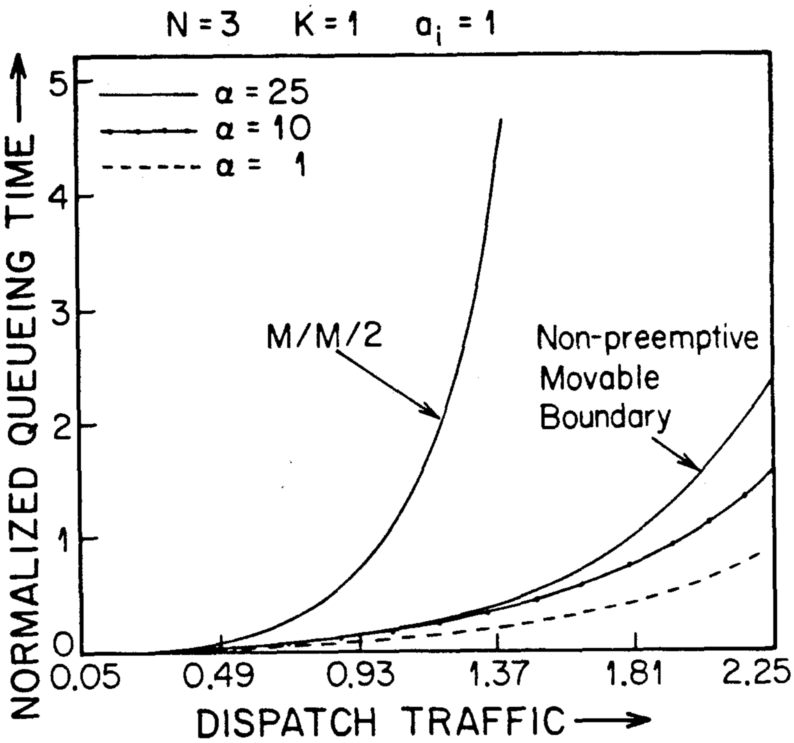


Fig. 4.10 The normlized queueing time as a function of dispatch traffic for NPMB and M/M/2 schemes for three different values of α.

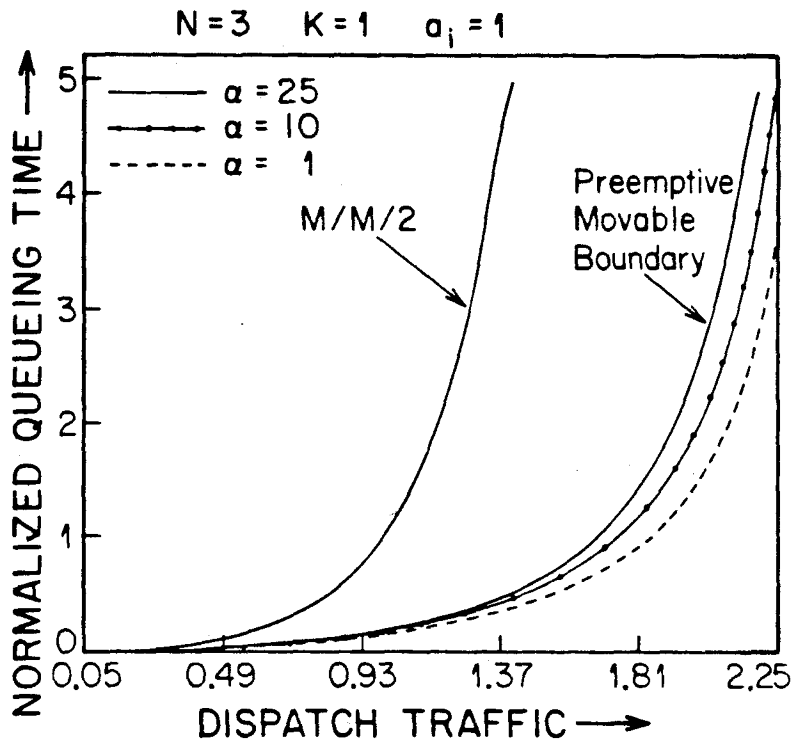


Fig. 4.11 The normlized queueing time as a function of dispatch traffic for PMB and M/M/2 schemes for three different values of α.

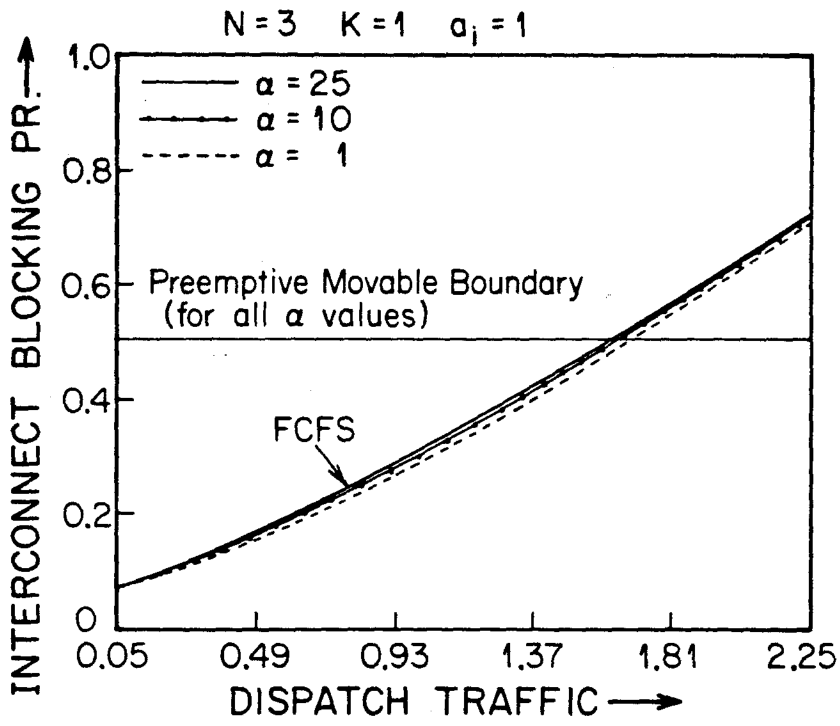


Fig. 4.12 Comparison of the blocking probability of interconnect calls as a function of dispatch traffic in PMB and FCFS cases.

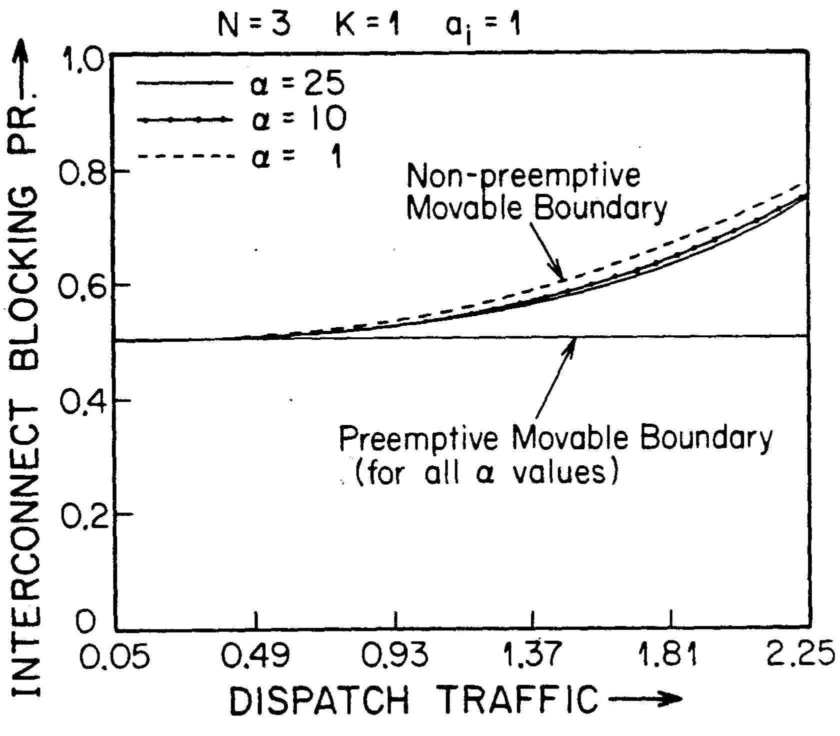


Fig. 4.13 Comparison of the blocking probability of interconnect calls as a function of dispatch traffic in PMB and NPMB cases.

As described at the beginning of this section, in the original strategy proposed by Motorola, the interconnect calls can only access to the repeaters that have physical connections to the interconnect terminals. This means that an interconnect call is blocked if all the secondary repeaters are busy, even if there are empty primary repeaters. Because of this inefficiency in utilizing the primary repeaters, the interconnect blocking probability increases. This is not so in NPMB case, where interconnect calls can use the primary repeaters provided that their number in the system is below a certain threshold.

Yet, on the other hand, as far as the dispatch grade of service is concerned, we don't gain much by the original strategy compared to the other strategies, either. This is because the number of repeaters available for dispatch use in the original strategy is the same as in the NPMB case at large dispatch traffic levels. Therefore, the NPMB strategy gives a very good idea about the delay performance of the original strategy.

As a summary, compared to the NPBM case, the original strategy has a noticeably worse blockage performance at all traffic levels, but a somewhat better time-delay performance at low traffic levels. The latter can not be considered as a real advantage because there is no congestion at low traffic levels, and so an arriving dispatch call is very likely to find empty repeaters.

Therefore, especially at high interconnect traffic intensities, the original strategy does not provide any advantages. Instead, one can choose another control scheme, one of the FCFS, PMB, or NPMB schemes, that will provide a substantial decrease in the interconnect blocking probability while keeping the dispatch grade of service high. The selection of the best control scheme depends on both the traffic intensity and the value of $\alpha = \mu_1/\mu_2$. The only way of improving the blockage performance in the original strategy would be to increase the number of physical connections to interconnect terminals, K. But this causes a substantial degradation of time-delay performance of dispatch traffic, especially at high dispatch traffic intensities and at large α values. Today's small computer technology makes it possible and easy to provide a new service discipline to the operating system of trunk mobile radio networks so as to increase the system performance by selecting the best strategy automatically.

From the comparison of those strategies, when $N = 3$, $K = 1$, $\alpha = 10$, and interconnect traffic is 1 Erlang, the interconnect blocking probability can be reduced to 30% of its original value by using the PMB strategy instead of the NPMB strategy at high traffic intensities ($a_d \approx 2$ Erlangs). The price paid here is an increase in the dispatch delay of approximately, at most, one dispatch transmission time. On the other hand, if FCFS is selected instead of NPMB, then the interconnect blocking probability would be reduced to 10% of its original value, without suffering a significant increase in the dispatch delay at all.

At low dispatch traffic intensities ($a_d \approx 0.5$ Erlangs), the reduction of the blocking probabilty can amount to 65% by selecting the FCFS strategy. The normalized dispatch queueing time, however, increases from 0.07 to 0.35. This means that the dispatch users who got service immediately when NPMB was used, now wait 35% of their holding time on the average. But the mean dispatch holding time is approximately 15 seconds. Therefore, dispatch users wait approximately 5 seconds in the queue if we use a FCFS strategy. The PMB strategy, at low and medium traffic intensities, cannot offer the same amount of reduction for the interconnect blocking probability as the FCFS scheme does. On the other hand, it does not suffer much in the dispatch time-delay performance.

In summary, we see that we should choose the FCFS strategy at high dispatch traffic intensity. At low dispatch traffic intensities, we should probably chose the PMB strategy.

CHAPTER 5

GENERALIZATIONS AND EXTENSIONS

In this chapter, the key-state approach is applied to a more general Markovian process which arises when bulk arrivals and departures are permitted. That is, when at least one server becomes free, a "bulk" of random but finite size will be accepted for service, and the system also accepts bulk arrivals at each (Poisson) arrival instant. Here we again consider two classes of users. The first-class users are queued and the second class users are blocked if there is no service facility available. The numbers of first-class and second-class users in a bulk are independent of each other. We consider particular traffic control schemes which allocate the use of service facilities between the users of different classes according to some specific priority rules. We will not directly find the solution for these problems. Instead, we focus on the solution for a more general problem. The solution for our special cases can be obtained by selecting proper transition rates. The integration models considered previously are special cases of this process as well.

5.1 Generalization of the Solution

Consider an ergodic Markovian process with state space $S = \{(i,j) : 0 \leq i < \infty, \quad 0 \leq j \leq N < \infty\}$. The set of states $\{(i,0),\ (i,1),\ (i,2),\ldots,(i,N)\}$ are called level-i states following to the notation of Neuts [41]. Note that there are infinite levels and $N+1$ states at each level. The parameter i corresponds to the number of first-class users in the system, and j corresponds the number of second-class users in the system. The transitions between states are defined by the following rates:

$\alpha_{i,j;i+\ell,j+k}$ is the transition rate from state (i,j) to state $(i+\ell, j+k)$.

We consider the class of Markovian processes where the transitions between levels i_1 and i_2 are not allowed within the same column, if $\mid i_1 - i_2 \mid > M$, nor between different columns, if $| i_1 - i_2 |> M - 1$. That is, $\alpha_{i,j;i+\ell,j+k} = 0$ if $\ell > M$ and $k = 0$ or if $\ell > M-1$ and $k \neq 0$. This means that the number of first-class users in a bulk cannot exceed M if there are no second-class users in the bulk; similarly, the number of first-class users in a bulk can not exceed $M-1$ if there is at least one second-class user in the bulk. These restrictions on the transitions between states keep the number of key states finite. Most of the queueing network models are special cases of this class. As an example, the voice and data integration models correspond to the case where $M = 1$, that is, we considered only one arrival or departure at a time. The transitions were only allowed between neighboring states on the same level and on the same column. This is an example for the class of single-arrival and single-departure systems that can be solved as a special case of bulk systems.

Packet networks and information broadcast systems can be considered as examples of systems where bulk arrivals and departures are allowed. In packet networks where an arriving packet length exceeds the protocol maximum and so has to be divided into smaller packets, a single-packet arrival can be considered as the arrival of several smaller packets. Bulk departure models, on the other hand, may arise in considering information broadcast systems where only a single unit of information can be retrieved from a data base at a time. The customers waiting for a specific piece of information receive the service simultaneously as soon as this information is broadcast.

The bulk model is also used to analyze single-arrival and departure systems. In cases where an arrival had to go through several, possibly a random number, of exponential service stations in order to complete service, instead of modeling the problem as a single arrival with an Erlangian service distribution, it can clearly be modeled as a bulk arrival system with an exponential service distribution [11, Chapter 4].

In this study, we consider the cases where the arrival and departure rates of the first-class users are independent of the number of first-class users in the system. This is generally true for most queueing networks except for those where the number of customers in the system is constant (closed queueing networks). In the latter case, the transition rates are state-dependent.

Here we will assume that transition rate of a bulk system is proportional to the number of first-class users in a bulk. This is usually the case when we model an $M/E_r/1$ system (a single server system where arrival instants are Poisson and the service distribution is Erlangian of phase r [44]) as an $M/M/1$ bulk-arrival system. The mean of the exponential distribution corresponding to the Erlangian distribution is proportional to the size r of the bulk. This assumption is essential for the solution technique we will propose. The same assumption could be made for the second-class users, but it turns out not to be a requirement for the solution. Therefore, the transition rate between two levels is assumed to be a function of only the difference of these levels.

This means

$$\alpha_{i,j;i+\ell,j+k} = \alpha_{j,j+k,\ell} \quad \text{for all } \ell.$$

Here $\alpha_{j,j+k,\ell}$ is defined as the rate of transition from (i,j) to $(i+\ell, j+k)$, the same for all $i \geq 0$. Therefore, for this class of Markovian processes:

(1) The transition rates depend only upon the difference between levels.

(2) There are no transitions within the same column if the level difference is greater than M, and there are no transitions between the different columns if the level difference is greater than $M-1$.

Fig. 5.1 shows the state diagram of an example with $M=2$ of the class of Markovian processes described above.

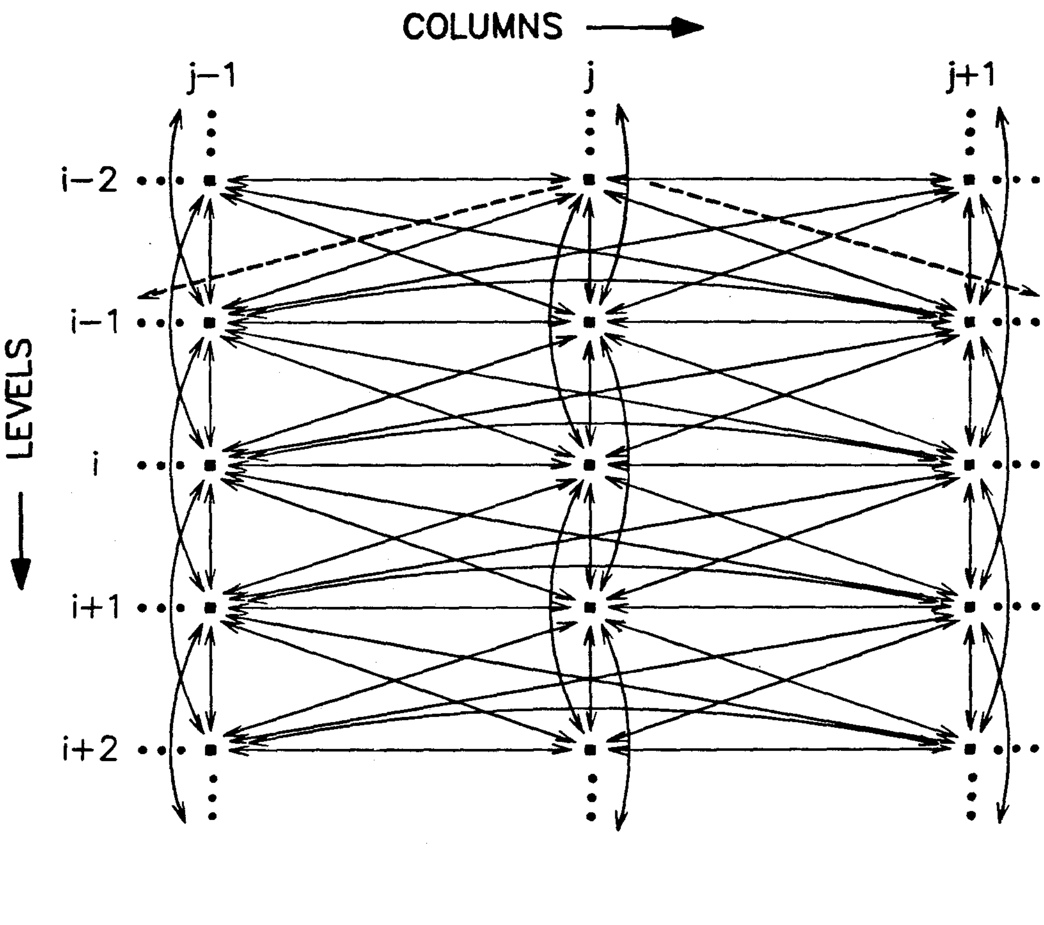


Fig. 5.1. State transition rate diagram for $M = 2$ of a Markovian process where transitions between two levels within the same columns are not allowed if the level difference is greater than 2 and the transitions between two levels among different columns are not allowed if the level difference is greater than 1.

The local balance equation for state (i,j) is obtained as follows. The probability flow rate out of (i,j) is:

$$P(i,j)\left(\sum_{\substack{\ell=-M\\ \ell\neq 0}}^{M}\alpha_{jj\ell}+\sum_{\substack{k=-j\\ k\neq 0}}^{N-j}\sum_{\ell=1-M}^{M-1}\alpha_{j,j+k,\ell}\right) \tag{5.1}$$

Likewise, the probability flow rate into (i,j) is:

$$\sum_{\substack{k=-j\\ k\neq 0}}^{N-j}\sum_{\ell=1-M}^{M-1}P(i+\ell,j+k)\alpha_{j+k,j,\ell}+\sum_{\substack{\ell=-M\\ \ell\neq 0}}^{M}P(i+\ell,j)\alpha_{jj\ell}. \tag{5.2}$$

From (5.1) and (5.2), the balance equation for state (i,j) is found as:

$$P(i,j)\left(\sum_{\substack{\ell=-M\\ \ell\neq 0}}^{M}\alpha_{jj\ell}+\sum_{\substack{k=-j\\ k\neq 0}}^{N-j}\sum_{\ell=1-M}^{M-1}\alpha_{j,j+k,\ell}\right)=$$

$$\sum_{\substack{k=-j\\ k\neq 0}}^{N-j}\sum_{\ell=1-M}^{M-1}P(i+\ell,j+k)\alpha_{j+k,j,\ell}+\sum_{\substack{\ell=-M\\ \ell\neq 0}}^{M}P(i+\ell,j)\alpha_{jj\ell}. \tag{5.3}$$

We define the delay operator E_i as:

$$E_i\ P(i,j)=P(i-1,j). \tag{5.4}$$

Note that, with E_i^ℓ the ℓ^{th} iterate of E_i, $E_i^\ell\ P(i,j)=P(i-\ell,j)$. We employ this definition to simplify the expression given in (5.3):

$$\sum_{\ell=-M}^{M}\alpha_{jj\ell}E_i^\ell P(i,j)+\sum_{\substack{k=-j\\ k\neq 0}}^{N-j}\sum_{\ell=1-M}^{M-1}\alpha_{j+k,j,\ell}E_i^\ell P(i,j+k)=0. \tag{5.5}$$

Here, $-\alpha_{jj0}$ is the transition rate out of state (i,j), and define as:

$$\alpha_{jj0} = -\left(\sum_{\substack{\ell=-M \\ \ell \neq 0}}^{M} \alpha_{jj\ell} + \sum_{\substack{k=-j \\ k \neq 0}}^{N-j} \sum_{\ell=1-M}^{M-1} \alpha_{j,j+k,\ell} \right). \tag{5.6}$$

Claim: The key states of the class of Markovian processes described above are the first $(M-1)$ level states, $\{(i,j) : 0 \leq i \leq M-1, \quad 0 \leq j \leq N\}$.

Proof: From (5.6), the equilibrium probability of state $(i+M,j)$ is obtained for $i \geq 0$ as:

$$\sum_{\ell=0}^{2M} \alpha_{j,j,\ell-M} E_i^\ell P(i+M,j) + \sum_{\substack{k=-j \\ k \neq 0}}^{N-j} \sum_{\ell=1}^{2M-1} \alpha_{j+k,j,\ell-M} E_i^\ell P(i+M,j+k) = 0. \tag{5.7}$$

By making use of the equation (5.7) at $i=0$, we see that all M^{th}-level equilibrium probabilities can be found by manipulating the lower-level probabilities. Hence, starting from M^{th}-level probabilities, all the upper-level probabilities are connected to the first $(M-1)$-level state probabilities by means of (5.7). This completes the proof of the claim that shows the key states are the first $(M-1)$-level states in the transition rate diagram.

As a result, all the equilibrium probabilities can be expressed as:

$$P(i,j) = \sum_{s=0}^{M-1} \sum_{r=0}^{N} c_{jrs}(i) P(s,r), \tag{5.8}$$

where $c_{jrs}(i)$ is the coefficient of the r^{th} key-state probability at level s.

We can now prove a theorem which gives closed-form expressions for the equilibrium probabilities of the process described at the beginning of this section. It generalizes the results found in Chapter 3 for voice-data integration networks, which are included here as a special case where $M=1$. Note that all the constraints on the transition rates are satisfied automatically when $M=1$.

Theorem: The limiting (equilibrium) probabilities of the class of Markovian processes described above are in the following form:

$$P(i,j) = P(0,0)\sum_t \phi_{jt} z_t^i. \tag{5.9}$$

Here the z_t's are the zeros of $\det\{R(z^{-1})\}$ with $z_t < 1$, while $R(z^{-1})$ is given as:

$$R(z^{-1}) = \begin{pmatrix} f_{00} & f_{10} & f_{20} & \dots & f_{N0} \\ f_{01} & f_{11} & f_{21} & \dots & f_{N1} \\ f_{02} & f_{12} & f_{22} & \dots & f_{N2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{0N} & f_{1N} & f_{2N} & \dots & f_{NN} \end{pmatrix}, \tag{5.10}$$

with

$$f_{jj} = \sum_{k=0}^{2M} z^{-k}\alpha_{j,j,k-M}, \quad f_{j+k,j} = \sum_{\ell=1}^{2M-1} z^{-\ell}\alpha_{j+k,j,\ell-M}. \tag{5.11}$$

Proof: The local balance equation obtained in (5.7) can be rewritten as:

$$f_{jj}(E_i)P(i+M,j) + \sum_{\substack{k=-j \\ k\neq 0}}^{N-j} f_{j+k,j}(E_i)P(i+M,j+k) = 0 \tag{5.12}$$

where

$$f_{jj}(E_i) = \sum_{k=0}^{2M} E_i^k \alpha_{j,j,k-M}, \quad f_{j+k,j}(E_i) = \sum_{\ell=1}^{2M-1} E_i^\ell \alpha_{j+k,j,\ell-M}. \tag{5.13}$$

The initial conditions of the equation (5.7) are found as $\{P(i,j) : 0 \leq i < M, \quad 0 \leq j \leq N\}$. These initial conditions are, however, unknown. Substituting (5.8) into (5.12), we obtain difference equations for the key-state coefficients with known initial conditions.

As a result of the above substitution, equation (5.12) yields as many difference equations as the number of key states. This means that we obtain a difference equation for every single key-state coefficient. These difference equations are given as :

$$f_{jj}(E_i)c_{jrs}(i) + \sum_{k=-j}^{N-j} f_{j+k,j}(E_i)c_{j+k,r,s}(i) = 0, \quad i \geq M, \tag{5.14}$$

$$0 \leq r \leq N, \qquad 0 \leq s \leq M-1.$$

In matrix form :

$$\begin{pmatrix} f_{00} & f_{10} & f_{20} & \cdots & f_{N0} \\ f_{01} & f_{11} & f_{21} & \cdots & f_{N1} \\ f_{02} & f_{12} & f_{22} & \cdots & f_{N2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{0N} & f_{1N} & f_{2N} & \cdots & f_{NN} \end{pmatrix} \begin{pmatrix} c_{0rs}(i) \\ c_{1rs}(i) \\ c_{2rs}(i) \\ \vdots \\ c_{Nrs}(i) \end{pmatrix} = \vec{0} \tag{5.15}$$

or

$$R(E_i)\ \vec{C}_{rs}(i) = \vec{0}, \quad 0 \leq r \leq N, \quad 0 \leq s \leq M-1. \tag{5.16}$$

As discussed in the solution of movable-boundary strategy in Sec. 3.1.1, $\vec{C}_{rs}(i)$ depends only on the zeros of $\det\{R(z^{-1})\}$. Since the degrees of the polynomials on the diagonal entries are all $2M$, the degree of the polynomial $\det\{R(z^{-1})\}$ is $2M(N+1)$. Therefore, $\det\{R(z^{-1})\}$ has $2M(N+1)$ zeros, say $\{z_1,\ z_2,\ \ldots,\ z_{2M(N+1)}\}$. Assuming that these zeros are all real and distinct, the general form of the solution is given as:

$$c_{jrs}(i) = \sum_{t=1}^{2M(N+1)} k_{rst}\ b_{jt}\ z_t^i, \qquad i \geq M. \tag{5.17}$$

Here the b_{jt}'s are the entries of the vector $\vec{B}_t$ obtained by solving the following equation:

$$R(z_t^{-1})\vec{B}_t = \vec{0} \quad t = 1, 2, \ldots 2M(N+1). \tag{5.18}$$

The k_{rst} are obtained by employing the initial conditions, as in (3.101).

Substituting the expression found for the key-state coefficients into equation (5.8), we obtain the following solutions for the equilibrium probabilities $P(i, j)$ for the non-key states. These are expressed in terms of $M(N+1)$ unknowns $P(s, r)$, $0 \leq r \leq N, \quad 0 \leq s \leq (M-1)$, which are the key-state probabilities:

$$P(i, j) = \sum_{t=1}^{2M(N+1)} b_{jt} z_t^i \left(\sum_{s=0}^{M-1} \sum_{r=0}^{N} k_{rst} P(s, r) \right), \quad 0 \leq j \leq N, \quad i \geq M. \tag{5.19}$$

Assuming the system is ergodic, the expression given in (5.19) must uniquely define a probability distribution. Let T be the subset of integers between 1 and $2M(N+1)$ and let $z_t \geq 1$ for $t \in T$. The key-state probabilities must satisfy the following relations in order to cancel the divergent terms in equation (5.19):

$$\sum_{s=0}^{M-1} \sum_{r=0}^{N} k_{rst} \ P(s, r) = 0, \quad \text{for all } t \text{ with } z_t \geq 1.$$

In matrix form,

$$\begin{pmatrix} k_{001} & k_{011} & \ldots & k_{M-1,N,1} \\ k_{002} & k_{012} & \ldots & k_{M-1,N,2} \\ \vdots & \vdots & \ddots & \vdots \\ k_{00m} & k_{01m} & \ldots & k_{M-1,N,m} \end{pmatrix} \begin{pmatrix} P(00) \\ P(01) \\ \vdots \\ P(M-1,N) \end{pmatrix} = \vec{0}. \tag{5.20}$$

Here m is the number of zeros with magnitude greater than or equal to 1.

The relationships between the key-state probabilities are obtained from (5.20). That is,

$$P(s,r) = \sigma_{rs}\ P(0,0), \quad 0 \le r \le N, \quad 0 \le s \le M-1. \tag{5.21}$$

Here σ_{rs} is the ratio of the key-state probability $P(s,r)$ and the empty-state probability $P(0,0)$. The existence and uniqueness of the equilibrium probabilities in an ergodic Markov chain guarantee a nontrivial solution of (5.20) for the key-state probabilities $P(s,r)$, $0 \le r \le N$, $0 \le s \le M-1$. The uniqueness of the relations between key-state probabilities implies that the nullity (dimension of the null space) of the coefficient matrix given in (5.20) must be 1. Hence, there is only one linearly independent solution to (5.20), and the desired solution vector is obtained by scaling this solution by $P(0,0)$. As a result of this scaling, the σ_{rs} are solved uniquely from (5.20). Finally, by substituting (5.21) into (5.19), the equilibrium probabilities can be expressed as follows (remember that $z_t < 1$ for $t \notin T$):

$$P(i,j) = P(0,0) \sum_{t \notin T} b_{jt} \left(\sum_{s=0}^{M-1} \sum_{r=0}^{N} k_{rst}\ \sigma_{rs} \right) z_t^i, \quad 0 \ge j \ge N, \quad i \ge M, \tag{5.22}$$

or

$$P(i,j) = P(0,0) \sum_{t \notin T} \phi_{jt}\ z_t^i, \quad i \ge M. \tag{5.23}$$

Here ϕ_{jt} is defined as:

$$\phi_{jt} = b_{jt} \left(\sum_{s=0}^{M-1} \sum_{r=0}^{N} k_{rst} \sigma_{rs} \right). \tag{5.24}$$

This completes the proof of the theorem.

5.2 Computational Problems:

The first computational problem arises in determining the determinant of $R(z^{-1})$ and in finding the zeros of the determinant polynomial. The polynomial

$\det\{R(z^{-1})\}$ is a $2M(N+1)^{th}$-degree polynomial in z. In cases where only the transitions between neighboring columns are allowed, $R(z^{-1})$ is a tridiagonal matrix. Generalized birth-death processes are of this type. Therefore, the algorithm (3.94) given in the previous section for the determinant of tridiagonal matrices can be used. In more general cases where transitions between non-neighboring states are allowed, conventional techniques can still be used to evaluate the determinant. We note that finding the zeros of the determinant polynomial is usually a tedious task and requires computer programs. For example, the so-called IMSL software library [43] contains some subroutines for finding the zeros of an arbitrary polynomial.

In order to calculate the coefficient vector $\vec{B}_t$ given in (5.15), the following system of linear equations must be solved for each zero of $\det R(z^{-1})$, $z_t \in T^*$:

$$R(z_t^{-1})\ \vec{B}_t = \vec{0}, \quad t \in T^*. \tag{5.25}$$

Here $T^* = \{1,\ 2, \ldots, 2M(N+1)\} - T$, that is, $T^* = \{t \mid z_t < 1\}$. This is equivalent to solving $(N+1)$ linear equations simultaneously and repeating this one for every $t \in T^*$.

In all the examples we have considered, only half of the zeros were greater than 1. We leave the proof of this observation in general as an extension to this research. Assuming that our observation can be generalized, the number of linear equations to be solved to obtain $\vec{B}_t$ is $M(N+1)$.

Finally, the coefficients k_{rst} given in (5.17) are obtained by employing the initial conditions on the key-state coefficients. From (5.8), these conditions are:

$$c_{jrs}(i) = \begin{cases} 1, & \text{if } r = j,\ s = i; \\ 0, & \text{otherwise.} \end{cases} \quad 0 \le r \le N, \quad 0 \le s \le M-1. \tag{5.26}$$

Substituting these initial conditions into (5.14), we obtain, at $i = M$:

$$\sum_t k_{rst} b_{jt} z_t^M = \frac{\alpha_{r,j,-s}}{\alpha_{j,j,-M}}, \quad 0 \le r \le N, \quad 0 \le s \le M-1. \tag{5.27}$$

Equation (5.27) depicts the fact that k_{rst} are found by solving $M(N+1)$ linear equations simultaneously.

As a result, as far as computational requirements are concerned, the price paid to get a closed-form solution increases as M and N increase. In addition to the computational work that must be done to find $\det R(z^{-1})$ and its zeros, the number of linear equations to be solved is $2M(N+1)$. The form of the solution gives insight into the system behavior, in spite of these computational difficulties.

In fact, the zeros of the $\det R(z^{-1})$ are very good indicators of the steady-state behavior. At lower traffic intensities, since the system is more likely to be empty, the lower-level states are the more probable states. Hence, we expect that the probability distribution of a level goes to zero rapidly as the level is increased. The closed-form expression given in (5.9) then suggests that the z_t must be much less than 1, since the rate of decrease of $P(i,j)$ is controlled by these zeros. On the other hand, at high traffic intensities, the higher-level states are the more probable states. Therefore z_t must be closer to 1 to make the distribution shift to higher levels. As a result, at traffic intensities where the values of some of these zeros are above a certain threshold, one can predict that the probability of being at higher levels will increase. Naturally, this will increase the size of the buffer required to store first-class users waiting for service. The problem of relating the buffer size to the values of z_t is also left as an open research problem.

5.3 Example: $M = 2$, $N = 1$

In this section, we apply the results of the previous section to a special case of a bulk arrival and departure systems with two classes of users. Here we assume that there cannot be more than two arrivals or departures at the same time. Also, in a bulk, either there are two users from the first class or one user from the first and

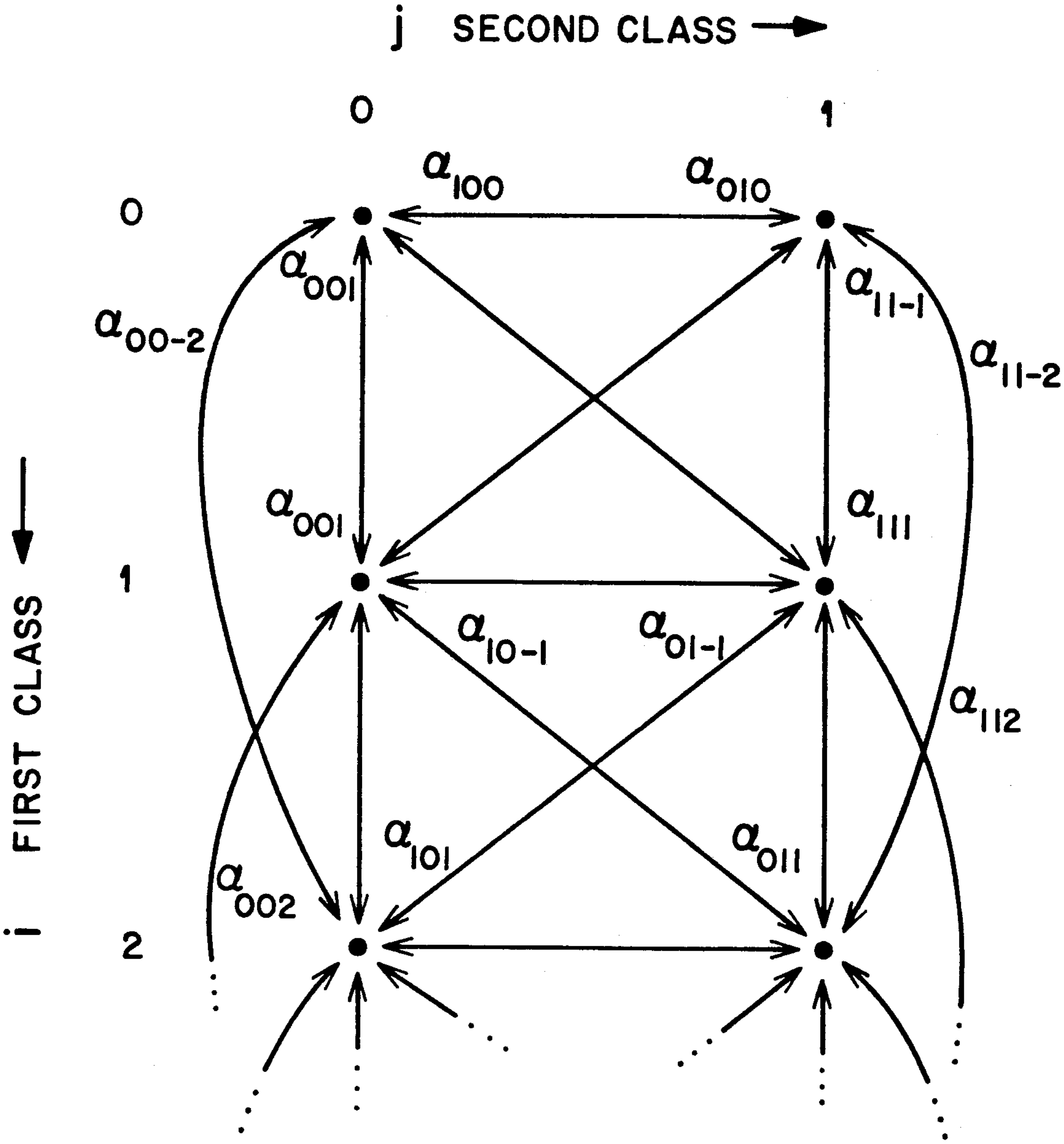


Fig. 5.2 State transistion rate diagram of a bulk arrival departure system when $M = 2$, $N = 1$.

one user from the second class. Single arrivals and departures are also allowed from both classes of traffic. In the notation of the previous section, $M = 2$ and $N = 1$. The latter assumption means that the number of second-class users in the system cannot exceed 1. The corresponding state-transition rate diagram is given in Fig. 5.2. This system can be considered as a single-server system where two first-class users or one first- and one second-class user can be served simultaneously. The second-class users can get service immediately upon arrival by preempting a first-class user. First-class arrivals are queued, and the second class arrivals are blocked if the server is busy.

The balance equations of the first column states are:

$$\sum_{\ell=-2}^{2} \alpha_{00\ell} P(i-\ell, 0) + \sum_{\ell=-1}^{1} \alpha_{10\ell} P(i-\ell, 1), \quad i \geq 2. \tag{5.28}$$

The second column states are:

$$\sum_{\ell=-2}^{2} \alpha_{11\ell} P(i-\ell, 1) + \sum_{\ell=-1}^{1} \alpha_{01\ell} P(i-\ell, 0), \quad i \geq 2, \tag{5.29}$$

where

$$-\alpha_{000} = \left(\sum_{\substack{\ell=-2 \\ \ell \neq 0}}^{2} \alpha_{jj\ell} + \sum_{\ell=-1}^{1} \alpha_{01\ell} \right), \tag{5.30}$$

$$-\alpha_{110} = \left(\sum_{\substack{\ell=-2 \\ \ell \neq 0}}^{2} \alpha_{jj\ell} + \sum_{\ell=-1}^{1} \alpha_{10\ell} \right). \tag{5.31}$$

Equations (5.30) and (5.31) can be rewritten by using the delay opearator E_i defined in (5.5):

$$f_{00}(E_i) P(i+2, 0) + f_{01}(E_i) P(i+2, 1) = 0, \tag{5.32}$$

$$f_{10}(E_i) P(i+2, 0) + f_{11}(E_i) P(i+2, 1) = 0, \tag{5.33}$$

or in matrix form:

$$\begin{pmatrix} f_{00}(E_i) & f_{01}(E_i) \\ f_{10}(E_i) & f_{11}(E_i) \end{pmatrix} \begin{pmatrix} P(i+2,0) \\ P(i+2,1) \end{pmatrix} = \vec{0}. \tag{5.34}$$

Here

$$\begin{aligned} f_{00}(E_i) &= \alpha_{002}E_i^4 + \alpha_{001}E_i^3 + \alpha_{000}E_i^2 + \alpha_{00-1}E_i + \alpha_{00-2}, \\ f_{01}(E_i) &= \alpha_{101}E_i^3 + \alpha_{100}E_i^2 + \alpha_{10-1}E_i, \\ f_{11}(E_i) &= \alpha_{112}E_i^4 + \alpha_{111}E_i^3 + \alpha_{110}E_i^2 + \alpha_{11-1}E_i + \alpha_{11-2}, \\ f_{10}(E_i) &= \alpha_{011}E_i^3 + \alpha_{010}E_i^2 + \alpha_{01-1}E_i. \end{aligned} \tag{5.35}$$

We find that $\det R(z^{-1})$ is an 8^{th} degree polynomial:

$$\det R(z^{-1}) = f_{00}(z^{-1})f_{11}(z^{-1}) - f_{01}(z^{-1})f_{10}(z^{-1}). \tag{5.36}$$

Assuming that all the roots are real and distinct, the solution for the key-state coefficients is found in (5.17):

$$c_{jrs}(i) = \sum_{t=1}^{8} k_{rst}b_{jt}z_t^i. \quad j = 0,1 \quad i \geq 2. \tag{5.37}$$

Here $r, s \in \{0,1\}$. The coefficients b_{jt} are solved from (5.18) for $1 \leq t \leq 8$:

$$\begin{aligned} b_{0t} &= 1 \\ b_{1t} &= -\frac{f_{00}(z_t^{-1})}{f_{01}(z_t^{-1})} = -\frac{f_{01}(z_t^{-1})}{f_{11}(z_t^{-1})}. \end{aligned} \tag{5.38}$$

From the balance equations obtained in (5.30) and (5.31) and from the proof of the claim about the key states of a bulk system given in the previous section, it can be shown that the key states are $\{(0,0),\ (0,1),\ (1,0),\ (1,1)\}$. Hence, the initial conditions for the key-state coefficients are as in (5.26). The coefficients k_{rst} are found by using these initial conditions to complete the solution for key-state coefficients. By employing the solution found for the key-state coefficients, the equilibrium probabilities are found by (5.19):

$$P(i,0) = \sum_{t=1}^{8}(k_{00t}P(0,0) + k_{01t}P(0,1) + k_{10t}P(1,0) + k_{11t}P(1,1))z_t^i, \quad i \geq 2, \tag{5.39}$$

$$P(i,1) = \sum_{t=1}^{8} b_{1t}(k_{00t}P(0,0) + k_{01t}P(0,1) + k_{10t}P(1,0) + k_{11t}P(1,1))z_t^i, \quad i \geq 2. \tag{5.40}$$

The key-state probabilities must satisfy the following relations to cancel the terms with $z_t \geq 1$, i.e., $t \in T$:

$$k_{00t}P(0,0) + k_{01t}P(0,1) + k_{10t}P(1,0) + k_{11t}P(1,1) = 0, \tag{5.41}$$

or in matrix form:

$$\begin{pmatrix} k_{001} & k_{011} & k_{101} & k_{111} \\ k_{002} & k_{012} & k_{102} & k_{112} \\ \vdots & \vdots & \vdots & \vdots \\ k_{00m} & k_{01m} & k_{10m} & k_{11m} \end{pmatrix} \begin{pmatrix} P(00) \\ P(01) \\ P(10) \\ P(11) \end{pmatrix} = \vec{0}. \tag{5.42}$$

Here we assumed that $z_t \geq 1$ for $t \in \{1, 2, \ldots, m\}$.

The existence and uniqueness of the solution guarantees a unique set of relations between the key-state probabilities. Therefore, from (5.42), we uniquely obtain:

$$P(1,0) = \sigma_{10}P(0,0), \quad P(0,1) = \sigma_{01}P(0,0). \quad P(1,1) = \sigma_{11}P(0,0). \tag{5.43}$$

In order to have a unique solution, the nullity of the coefficient matrix corresponding to the equations given in (5.42) must be exactly one. This implies that there can only be one linearly independent solution vector for (5.42) which gives the unique relations between key-state probabilities.

5.4 Extensions:

The generalized model given in Section 5.1 is still not the most general one. We have only considered two-dimensional Markovian processes. Application of the key-state approach to higher-dimensional problems can be effected by employing the same principles. Of course, we should expect to have to solve more difference equations as we increase the dimension of our problem.

Note that the difference equations obtained for the key-state coefficients are always constant-coefficient difference equations in the cases we have considered. The reason for that is the assumptions we made on the state-transition rates. At the beginning of this chapter, we assumed that the transition rates depend only upon the difference between levels. In cases where the state-transition rates depend directly on the level of the states, these difference equations no longer have constant coefficients. The cases in which this assumption fails will be an important area for extension of the key-state approach.

One of the most important questions about queueing networks is the existence of a product-form solution for a given state-space structure, as defined in Section 2.2. Despite efforts to categorize those queueing networks for which product-form solution do or do not exist, an approach has not been developed yet to check for this. The expression found for the equilibrium probabilities in (5.9) suggests research on developing an algorithm to check the existence of product-form solutions which depends on the factorization of the coefficient ϕ_{jt}. If there is a product-form solution, then the joint probability of having i first- and j second-class users in the system can be factored into the product of each of the marginal distributions. That is:

$$P(i,j) = P_i(i)P_j(j). \tag{5.44}$$

This is possible only if the coefficient ϕ_{jt} in (5.23) can also be factored into two coefficients, which must be functions of j or t only:

$$\phi_{jt} = \phi_j \ \phi_t. \tag{5.45}$$

Hence, (5.23) can be rewritten as:

$$P(i,j) = P(0,0)\phi_j \sum_t \phi_t z_t^i \tag{5.46}$$

or

$$P(i,j) = P_i(i) \ P_j(j), \tag{5.47}$$

as required for a product-form solution. Here $P_j(j) = P(0,0)\phi_j$ and $P_i(i) = \sum_t \phi_t z_t^i$. As a result, conditions on the factorization of ϕ_{jt} would give an idea about the existence of product-form solutions for the type of networks which equilibrium probabilities can be expressed as in (5.9).

CHAPTER 6

SUMMARY AND CONCLUSIONS

In this thesis, a new approach (key-state) which provides exact closed-form expressions for the equilibrium probabilities of a rather general Markovian queueing network is developed. Prior techniques use linear recursions on the equilibrium probabilities. In the key-state approach, recursions are developed for the key-state coefficients which eliminate the computational problems previously encountered. Chapter 2 was devoted to the explanation of this approach.

In Chapter 3, several voice-data integration techniques were investigated by using the key-state technique. We found that the optimal traffic control scheme for the distribution of output channels between two classes of users strongly depends on the ratio of the mean holding times and traffic intensities. In Tables 3.1 and 3.2, the traffic control strategies were ranked according to their performance under light and heavy traffic intensities. As desired, there was a quantitative trade off between the blockage and the time-delay performance of a traffic control strategy when one class of traffic is blocked and the other class is delayed.

In Chapter 4, we focused on the traffic problems of another integrated type of network, trunked mobile radio. Here the two classes of users were dispatch and interconnect. In the original design, the number of physical connections to interconnected terminals were controlled by an operator according to the traffic intensity. In Section 4.2, we proposed a sharing algorithm that determined the number of repeaters permitted access to interconnect use. This algorithm operated automatically, transparent to the system operator. The traffic-control strategies discussed in Chapter 3 gave better blockage performance than the original strategy. However, the proposed strategies were not readily applicable to trunked mobile radio because of the physical constraints of the equipment. But today's small-computer technology now makes it possible and even easy to modify the original dispatch-interconnect integration technique so that these new schemes can also be used. The best control scheme would then be selected automatically based on the traffic intensity and the ratio between the holding times.

Finally, in Chapter 5, the results of the key-state approach were generalized to a more general Markovian process, which arises when bulk arrivals and departures are permitted in a system shared by two classes of users. The computational complexity of finding closed-form expressions for the equilibrium probabilities increased as the size of bulk increased, but obtaining the solution was still computationally feasible. To be able to analyze bulk arrival and departure integrated systems was important for two reasons. First, the result covers integration schemes where the holding time distribution is no longer exponential but Erlangian. As an example, we considered a single-arrival, single-departure system where the arrivals had to go through several exponential service stations. This system can be modeled as a bulk arrival system with a single exponential server. Second, new protocols and integration schemes may be developed for communication services that have not yet been integrated, and these may involve the bulking of arrivals at points in the network.

APPENDIX A

The roots in FCFS strategy

Claim: The roots of the characteristic equation given in (18) are real, positive and satisfy the following relations:

$$z_{2i} < 1, \quad z_{2i-1} > 1, \quad z_{2N} = 1, \quad z_{2N-1} = \frac{\lambda_1}{N\mu_1} \qquad i = 1, 2, ..., N-1. \qquad (A-1)$$

Proof: The roots are given as:

$$z_{2i} = \frac{-a_i - \sqrt{a_i^2 - 4b_i}}{2}; \qquad z_{2i-1} = \frac{-a_i + \sqrt{a_i^2 - 4b_i}}{2} \qquad (A-3)$$

where

$$a_i = -\frac{\lambda_1 + i\mu_1 + (N-i)\mu_2}{i\mu_1}; \qquad b_i = \frac{\lambda_1}{i\mu_1}. \qquad (A-4)$$

If $a_i^2 - 4b_i > 0$ then the roots are real. From (A-4),

$$1 + a_i = -b_i - \frac{(N-i)\mu_2}{i\mu_1} < -b_i \qquad (A-5)$$

since $\frac{(N-i)\mu_2}{i\mu_1} > 0$. This implies that

$$4(1 + a_i) < -4b_i \Rightarrow a_i^2 + 4(1 + a_i) < -4b_i + a_i^2$$

$$\Rightarrow (2 + a_i)^2 < a_i^2 - 4b_i \Rightarrow \mid (2 + a_i) \mid < \sqrt{a_i^2 - 4b_i} \qquad (A-6)$$

$$\Rightarrow (a_i^2 - 4b_i) > (2 + a_i)^2 > 0 \Rightarrow \text{ the roots are real.}$$

From (A-6), for $a_i > -2$,

$$2 + a_i < \sqrt{a_i^2 - 4b_i} \Rightarrow z_{2i-1} = \frac{-a_i + \sqrt{a_i^2 - 4b_i}}{2} > 1, \qquad (A-7)$$

and for $a_i < -2$,

$$-(2+a_i) < \sqrt{a_i^2 - 4b_i} \Rightarrow z_{2i} = \frac{-a_i - \sqrt{a_i^2 - 4b_i}}{2} < 1. \qquad (A-8)$$

Furthermore, $z_{2i-1} > 1 > z_{2i} > -a_i > 0$. Hence the two roots are positive.

Finally, if $i = N$ then $1 + a_i = -b_i$, which yields:

$$z_{2N} = 1, \qquad z_{2N-1} = \frac{\lambda_1}{N\mu_1}; \qquad (A-9)$$

this completes the proof of the claim given by (A-1).

APPENDIX B

The stability equations for key-state probabilities

Claim: For $1 \le m \le N$,

$$\sum_{\ell=0}^{N-1} A_{N-j,2m-1,\ell}\ P_\ell(0) = 0 \quad \text{for } j = m \qquad (B-1)$$

implies that the summation is zero for all j with $m \le j \le N$.

Proof: From (27), the equilibrium probabilities can be written as:

$$P_{N-m}(d) = \sum_{i=0}^{2m} z_i^d \sum_{j=0}^{N-1} A_{N-m,i,j}\ P_j(0), \qquad d \ge m+2. \qquad (B-2)$$

We define $g_{N-m,i}$ as:

$$g_{N-m,i} = \sum_{j=0}^{N-1} A_{N-m,i,j} P_j(0). \qquad (B-3)$$

Substituting this into (B-2), we get :

$$P_N(d) = g_{N,0} z_0^d; \qquad (B-4)$$

$$P_{N-1}(d) = g_{N-1,0} z_0^d + g_{N-1,1} z_1^d + g_{N-1,2} z_2^d; \qquad (B-5)$$

$$P_{N-2}(d) = g_{N-2,0} z_0^d + g_{N-2,1} z_1^d + g_{N-2,2} z_2^d + g_{N-2,3} z_3^d + g_{N-2,4} z_4^d; \qquad (B-6)$$

$$\vdots$$

$$P_{N-m}(d) = g_{N-m,0} z_0^d + \ldots + g_{N-m,2m-1} z_{2m-1}^d + g_{N-m,2m} z_{2m}^d. \qquad (B-7)$$

From equations (B-4) to (B-7), it is observed that z_{2j-1}^d appears in the expressions for all P_{N-m}'s for $m \geq j$. We will show that if the coefficient of z_{2j-1}^d is zero in $P_{N-j}(d)$, then the corresponding coefficients are zero in all $P_{N-m}(d)$'s for $m \geq j$. Algebraically speaking, we will show that

$$g_{N-j,2j-1} = 0 \Rightarrow g_{N-m,2j-1} = 0, \qquad j \leq m \leq N. \qquad (B-8)$$

By employing the recursive relation obtained for the coefficient of the particular solution in (20), $g_{N-j,2j-1}$ can be written as follows:

$$g_{N-j,2j-1} = \sum_{\ell=0}^{N-1} A_{N-j,2j-1,\ell}\ P_\ell(0)$$

$$= \sum_{\ell=0}^{N-1} \left(\frac{1 + a_{j+1} z_{2j-1}^{-1} + b_{j+1} z_{2j-1}^{-2}}{d_{j+1}\ z_{2j-1}^{-1}} \right) A_{N-j-1,2j-1,\ell}\ P_\ell(0)$$

$$= \left(\frac{1 + a_{j+1}\ z_{2j-1}^{-1} + b_{j+1}\ z_{2j-1}^{-2}}{d_{j+1}\ z_{2j-1}^{-1}} \right) g_{N-j+1,2j-1}$$

$$\vdots$$

$$= \prod_{\ell=1}^{j-m} \left(\frac{1 + a_{j+\ell}\ z_{2j-1}^{-1} + b_{j+\ell}\ z_{2j-1}^{-2}}{d_{j+\ell}\ z_{2j-1}^{-1}} \right) g_{N-m,2j-1}.$$

These equations imply that if $g_{N-j,2j-1} = 0$ then $g_{N-\ell,2j-1} = 0$ for $\ell \geq j$, because

$$\frac{1 + a_{j+\ell}\ z_{2j-1}^{-1} + b_{j+\ell}\ z_{2j-1}^{-2}}{d_{j+\ell}\ z_{2j-1}^{-1}} \neq 0, \qquad \ell = 1, \ldots, N-m. \qquad (B-9)$$

The reason why (B-9) is non-zero can be explained as follows: z_2j and z_{2j-1} are the roots of the characteristic equation of the difference equation given in (19). Therefore, z_{2j} and z_{2j-1} are the roots of the polynomial given by:

$$p_j(z) = 1 + a_j\ z^{-1} + b_j\ z^{-2}. \qquad (B-10)$$

The numerator of (B-9) is found as $p_{j+\ell}(z)$. Hence, the zeros of (B-9) are found to be $z_{2(j+\ell)}$ and $z_{2(j+\ell)-1}$. From (18), it can be seen that

$$z_{2(j+\ell)-1} \neq z_{2j-1} \qquad \text{for } \ell \neq 0 \qquad (B-11)$$

because, $a_{j+\ell} \neq a_j$ and $b_{j+\ell} \neq b_j$ for $i \neq j$. Therefore, z_{2j-1} can not be a zero for (B-9). This completes the proof of the claimed implication in (B-1).

REFERENCES

[1] A. E. Joel, "The Past 100 Years in Telecommunications Switching,"IEEE Comm. Magazine, May 1984, Vol. 22, No. 5, 68-70.

[2] G. J. Coviello, "Comparative Discussion of Circuit- vs. Packet- Switched Voice," IEEE Tran. on Comm., Vol. COM-27, No. 8, August 1979, 1153-1159.

[3] M. Schwartz, *Telecommunication Networks*, Adison-Wesley, Reading, Mass., 1987.

[4] R. Vickets and T. Vilmansen, "The Evolution of Telecommunications Technology," Proceedings of the IEEE, Vol. 74, No. 9, 1231-1245.

[5] W. E. Falconer, J. A. Hooke, "Telecommunications Services in the Next Decade,"Proceedings of the IEEE, Sept. 1986, Vol. 74, No. 9, 1246-1261.

[6] T. E. Browne,"Network of the Future," Proceedings of the IEEE, Sept. 1986, Vol. 74, No. 9, 1222-1230.

[7] T. Irmer, "An Idea Turns into Reality - CCITT Activities on the way to ISDN," IEEE J. on Selected Areas in Comm., Vol. SAC-4, No. 3, May 1986. 316-319

[8] D. J. Kostas, "Transition to ISDN - An Overview," IEEE Comm. Magazine, Vol. 22, No. 1, Jan. 1984, 11-17.

[9] D. Bertsekas, R. Gallager, *Data Networks*, Prentice-Hall, New Jersey, 1987.

[10] H. Inose, *An Introduction to Digital Integrated Communications Systems*, University of Tokyo Press, Tokyo, 1979.

[11] L. Kleinrock, *Queueing Systems, Volume I:Theory.* New York, Wiley, 1975.

[12] B. W. Stuck and E. Arthurs, *A Computer and Communications Networks Performance Analysis Primer*, New Jersey, Prentice-Hall, 1985.

[13] A. S. Tanenbaum, *Computer Networks*, New Jersey, Prentice-Hall, 1981.

[14] T. L. Satty, *Elements of Queueing theory with applications*, Dover, New York, 1961.

[15] J. Medhi, *Stochastic Processes,* Wiley, New Delhi, 1982.

[16] F. P. Kelly, *Reversibility and Stochastic Networks,* Wiley, New York, 1979.

[17] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 2, Second Edition, Wiley, New York, 1966.

[18] K. Sriram, P. K. Varshney, and J. G. Shanthikumar, "Discrete-Time Analysis of Integrated Voice/Data Multiplexers with and without Speech Activity Detectors," IEEE J. on Selected Areas in Comm., Vol. SAC-1, no. 6, Dec. 1983, 1124-1132.

[19] Y. N. Doğanata, "A Shared Service Algorithm For a Trunked Mobile System," IEEE Trans. on Vehic. Tech., Vol. VT-35, No. 3, pp 93-100, Aug. 1986.

[20] J. A. Morrison, "Analysis of Some Overflow Problems with Queuing," Bell System Technical Journal, vol. 59, No. 8, pp. 1427-1447, Oct. 1980.

[21] L. Kaufman, J. B. Seery, and J. A. Morrison, "Overflow Models for Dimension PBX Feature Packages," Bell System Technical Journal, vol. 60, No. 5, May-June 1981.

[22] M. L. Honig, "Analysis of TDMA Network with Voice and Data Traffic", AT&T Bell Laboratories Technical Journal, Vol.63, No.8, October 1984.

[23] E. Arthurs and B. W. Stuck," Traffic Analysis Tools for Integrated Digital Time-Division Link Level Multiplexing of Synchronous and Asynchronous Message Streams," IEEE J. on Selected Areas in Comm., vol. SAC-1, No. 6, Dec. 1983, 1112-1123.

[24] B. Kraimeche and M. Schawartz, "Analysis of Traffic Access Control Strategies in Integrated Service Networks," IEEE Trans. on Comm., Vol. COM-33, No. 9, Oct. 1985, 1085-1093.

[25] M. J. Fischer and T. C. Harris, " A Model for Evaluating the Performance of an Integrated Circuit- and Packet-Switched Multiplex Structure," IEEE Trans. on Comm., Vol. COM-24, No.2, Feb. 1976, 195-202.

[26] U. N. Bhat and M. J. Fischer, " Multichannel Queueing systems with Heterogenous Classes of Arrivals ", Naval Res. Logist. Quarterly, Vol. 23, No. 2, 1976, 271-283.

[27] C. J. Weinstein, M. L. Malpass, and M. J. Fischer, "Data Traffic Performance of an Integrated Circuit and Packet - Switched Multiplex Structure," IEEE Trans. on Comm., Vol. COM-28, no. 6, June 1980, 873-877.

[28] MOTOROLA Product Digest: " Centralized Interconnect for 800MHz Trunked Systems," MOTOROLA Communication and Electronics, Inc. Schaumburg, Illinois, 1978.

[29] J. R. Jackson, " Job Shop-like Queueing Systems," Management Science, Vol. 10, No. 1, 1963, 131 - 142.

[30] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios, " Open, Closed, and Mixed Networks of Queues with Different Classes of Customers," J. ACM, Vol. 22, no. 2, 1975, 248-260.

[31] A. A. Lazar and T. G. Robertazzi, " The Geometry of Lattices for Markovian Queueing Networks," Proc. of the Eighteenth Conference on Information Sciences and Systems, Princeton University, Princeton, N. J., 1984, pp. 164-168

[32] A. A. Lazar and T. G. Robertazzi, " An Algebraic Topological Approach to Markovian Queueing Networks " Proc. of the Eighteenth Conference on

Information Sciences and Systems, Princeton University, Princeton, N. J., 1984, pp. 437-442

[33] A. A. Lazar and T. G. Robertazzi, " Markovian Petri Net Protocols with Product Form Solution," Proc. of the 24^{th} Annual Allerton Conference on Communications, Control and Computing, University of Illinois, Urbana, ILL, Oct. 1-3, 1986.

[34] B. Occhiogrosso et. al., "Performance Analsis of Integrated Switching Communications Systems," Proc. National Telecommunications Conference, Los Angeles, Dec. 1977, 12:4-1–12:4-11.

[35] A. Leon-Garcia, R. H. Kwong, and G. F. Williams, " Performance Evaluation Methods for an Integrated Voice-Data Link," IEEE Trans. on Comm., Vol. COM-30, No. 8, Aug. 1982, 1848-1858.

[36] K. Kummerle, " Multiplexer Performance for Integrated Line and Packet - Switched Traffic, " Proc. 2^{nd} Int. Conf. Comput. Comm., Stockholm, Sweden, Aug. 1974, 508-515.

[37] P. Zafiropoulo, " Flexible Multiplexing for Networks Supporting Line - Switched and Packet - Switched Data Traffic, " Proc. 2^{nd} Int. Conf. Comput. Commun., Stockholm, Sweden, Aug. 1974, 517 - 523.

[38] G. Coveiello and P. Vend, " Integration of Circuit / Packet Switching by a SENET (Slotted Envelope Network) Concept, " Proc. National Telecommunications Conference, New Orleans, Dec. 1975, 42-12–42-17.

[39] Chi-Tsong Chen, *Linear System Theory and Design,* Holt, Rinehart and Winston, 1984.

[40] G. Korn, T. Korn, *Mathematical Handbook for Scientist and Engineers*, Second Edition, McGraw-Hill, 1968.

[41] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*, John Hopkins, Baltimore, 1981.

[42] D. P. Gaver and P. P. Lehoczky, " Channels that Cooperatively Service a Data Stream and Voice Messages, " IEEE Trans. on Comm., Vol. COM-30, No. 5, May 1982.

[43] The IMSL LIBRARY *a set of fortran subroutines for mathematics and statistics,* IMSL, Houston, Texas.

[44] H. J. Larson, *Introduction to Probability Theory and Statistical Inference*, third edition, New york, Wiley, 1982