

MEAN DELAY ANALYSIS FOR UNIDIRECTIONAL BROADCAST STRUCTURES

by

JOSEPH WAI MING PANG

B.A.Sc., The University of British Columbia, 1983

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

Department of Electrical Engineering

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

© Joseph Wai Ming Pang, 1985

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of ELECTRICAL ENGINEERING

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date SEP. 5, 1985

Abstract

Unidirectional broadcast structures constitute a class of high performance local network architectures. They are very flexible and well suited for fiber optic implementation. The access methods used in these networks are often based on certain implicit token-passing mechanisms to provide superior delay-vs-throughput characteristics. The performance of these unidirectional broadcast protocols is evaluated in this thesis via a study on the classical token-ring system. Emphasis is placed on the analysis of mean delay-vs-throughput performance for the gated exhaustive service discipline under asymmetric traffic. The analysis involves examination of the statistical behaviour of interacting queues. A number of exact results are derived and based on these results, a very good approximation for the average delays is developed. The approximation agrees closely with exact numerical solutions over a wide range of system parameters. The implications of the approximation are also discussed.

TABLE OF CONTENTS

	Page
1. INTRODUCTION.....	1
1.1 High-performance Local Area Networks.....	1
1.2 Objectives.....	2
1.3 Previous Work.....	3
1.4 Outline.....	4
2. DEVELOPMENT OF THE LAN SYSTEM MODEL.....	6
2.1 Description of D-net and its Communication Protocol.....	6
2.2 Comparison of D-net with Token-Ring Systems.....	10
2.3 System Model and Parameters.....	11
2.4 Definitions of System Random Variables.....	15
2.5 Comments on the Relations between System Random Variables....	17
3. FORMULATION OF THE SYSTEM EQUATIONS.....	19
3.1 Imbedded Markov Chain Formulation.....	19
3.2 Moment Calculations.....	22
3.3 Approaches to the Solutions of the Basic Equations.....	23
3.4 Existence and Uniqueness of Solutions to the Basic Equations.....	24
3.5 Comments Regarding Solutions of the Basic Equations.....	26

TABLE OF CONTENTS (cont'd)

	Page
4. SOME EXACT ANALYTIC RESULTS.....	28
4.1 Overview of Results.....	28
4.2 Expected Cycle Time.....	28
4.3 Simplification of the Forcing Matrix B.....	31
4.4 The Jacobian Matrix A.....	33
4.5 An Auxilliary Result.....	35
4.6 A Difference Equation for the Second Joint Moments of Terminal Service Times.....	37
4.7 Solution for the Symmetric Case.....	45
5. MEAN DELAY ANALYSIS.....	47
5.1 Relationship between Mean Message Delays and Service Time Moments.....	47
5.2 A Fundamental Relation for $\{\tilde{v}_{ii}\}$	50
5.3 An Approximation for $\{\tilde{v}_{ii}\}$	52
5.4 Comments on the Approximation for $\{\tilde{v}_{ii}\}$	54
5.5 Comparison of Approximation and Exact Numerical Results for $\{\tilde{v}_{ii}\}$	56
5.6 Delay-vs-Throughput Results.....	65
5.7 Finite Buffering and Non-exhaustive Services.....	79

TABLE OF CONTENTS (cont'd)

	Page
6. CONCLUDING REMARKS.....	82
6.1 Summary of the Results.....	82
6.2 Suggestions for Further Work.....	83
REFERENCES.....	86
APPENDIX: Eigenvalues of the Jacobian Matrix A.....	89

LIST OF FIGURES

	Page
2.1	Architecture of D-net..... 7
2.2	Trains of Packets on D-net..... 7
2.3	The Star-coupled Version of D-net..... 9
2.4	The Open-ring Version of D-net..... 9
2.5	Schematics for the Single-server Multi-queue Model..... 12
5.1a	E_{\max} versus ρ for $N=8$ and 1 single large user..... 59
5.1b	E_{avg} versus ρ for $N=8$ and 1 single large user..... 59
5.2a	E_{\max} versus ρ for $N=8$ and 1 single small user..... 60
5.2b	E_{avg} versus ρ for $N=8$ and 1 single small user..... 60
5.3a	E_{\max} versus ρ for $N=8$ and 2 clusters of identical users..... 61
5.3b	E_{avg} versus ρ for $N=8$ and 2 clusters of identical users..... 61
5.4a	E_{\max} versus ρ for $N=16$ and 1 single large user..... 62
5.4b	E_{avg} versus ρ for $N=16$ and 1 single large user..... 62
5.5a	E_{\max} versus ρ for $N=16$ and 1 single small user..... 63
5.5b	E_{avg} versus ρ for $N=16$ and 1 single small user..... 63
5.6a	E_{\max} versus ρ for $N=16$ and 2 clusters of identical users..... 64
5.6b	E_{avg} versus ρ for $N=16$ and 2 clusters of identical users..... 64
5.7a	\tilde{v}_{ii} versus ρ for $N=8$, 1 single large user and $t_i=\rho_i^2$ 70
5.7b	$\text{var}(\tilde{v}_{ii})$ versus ρ for $N=8$, 1 single large user and $t_i=\rho_i^2$ 70
5.8a	\tilde{v}_{ii} versus ρ for $N=8$, 1 single large user and $t_i=\rho_i$ 71
5.8b	$\text{var}(\tilde{v}_{ii})$ versus ρ for $N=8$, 1 single large user and $t_i=\rho_i$ 71
5.9a	\tilde{v}_{ii} versus ρ for $N=8$, 1 single large user and $t_i=1$ 72
5.9b	$\text{var}(\tilde{v}_{ii})$ versus ρ for $N=8$, 1 single large user and $t_i=1$ 72

LIST OF FIGURES (cont'd)

	Page
5.10a \tilde{v}_{ii} versus ρ for $N=8$, 1 single small user and $t_i=\rho_i^2$	73
5.10b $\text{var}(\tilde{v}_{ii})$ versus ρ for $N=8$, 1 single small user and $t_i=\rho_i^2$	73
5.11a \tilde{v}_{ii} versus ρ for $N=8$, 1 single small user and $t_i=\rho_i$	74
5.11b $\text{var}(\tilde{v}_{ii})$ versus ρ for $N=8$, 1 single small user and $t_i=\rho_i$	74
5.12a \tilde{v}_{ii} versus ρ for $N=8$, 1 single small user and $t_i=1$	75
5.12b $\text{var}(\tilde{v}_{ii})$ versus ρ for $N=8$, 1 single small user and $t_i=1$	75
5.13a \tilde{v}_{ii} versus ρ for $N=8$, 2 clusters of identical users and $t_i=\rho_i^2$	76
5.13b $\text{var}(\tilde{v}_{ii})$ versus ρ for $N=8$, 2 clusters of identical users and $t_i=\rho_i^2$	76
5.14a \tilde{v}_{ii} versus ρ for $N=8$, 2 clusters of identical users and $t_i=\rho_i$	77
5.14b $\text{var}(\tilde{v}_{ii})$ versus ρ for $N=8$, 2 clusters of identical users and $t_i=\rho_i$	77
5.15a \tilde{v}_{ii} versus ρ for $N=8$, 2 clusters of identical users and $t_i=1$.	78
5.15b $\text{var}(\tilde{v}_{ii})$ versus ρ for $N=8$, 2 clusters of identical users and $t_i=1$	78
A.1 Right-half plane semicircle.....	93

ACKNOWLEDGEMENTS

I would like to acknowledge my appreciation to Dr. R.W. Donaldson for his encouragement and supervision throughout the course of this research..

I would also like to thank Dr. H.W. Lee for his many valuable suggestions.

Thanks are due to Ms. Charlotte Stevenson for typing the manuscript.

Finally, financial support from an NSERC postgraduate scholarship is gratefully acknowledged.

1. INTRODUCTION.

1.1 High-performance Local Area Networks.

The demand for local communication resources continues to grow very rapidly. Today, local area networks (LANs) operating at a few Mbit/s can still satisfy most current needs but the situation is likely to change in the near future. The advent of superfast microcomputers, the evolution of new system architectures (e.g. distributed processing) and the integration of data intensive services (e.g. interactive graphics, tele-conferencing) all lead to ever higher bandwidth requirements. Even with sophisticated bandwidth compression techniques, the bandwidth of the conventional co-axial cable will not be adequate for many applications in the future.

Optical fiber emerges as the most promising candidate for implementing the high-performance LANs needed in the future. The most impressive feature of optical fiber is its enormous information-carrying capacity; the theoretical limit is well above 100 Gbps-km. Several Gbps-km can be achieved using current technology. As the technique of wavelength-division multiplexing matures, the bandwidth of an optical fiber will increase many-fold. Thus, not only can optical fiber satisfy the bandwidth requirements of LANs in the foreseeable future, it will also provide great flexibility for growth in the long term. Other desirable features such as immunity to electromagnetic interference, high channel security, low signal loss, small size, light weight, etc. all make optical fiber very attractive for LAN applications. Nevertheless, limitations on current fiber optics technology do impose constraints on LAN system designs ([5-7]). For example,

the very high insertion loss of current optical taps has led designers to consider point-to-point configurations such as rings or stars instead of linear bus structures. Furthermore, bidirectional transmission on a single optical fiber also faces considerable difficulty due to the minute fiber core size and crosstalk between source and detector.

To overcome the technological difficulties with fiber optics, a number of new network schemes have been developed ([8-13]). Among these, the unidirectional broadcast structures (UBSs) with conflict-free schedulings appear to be most suitable for very high data rate applications. The access protocols used in these UBSs resemble the conventional token-passing on a ring which provides superior delay-throughput characteristics. However, these networks avoid the major pitfalls associated with the token-ring. As a result, UBSs have received much attention lately. A typical UBS called D-net will be described in Chapter 2, together with its interesting properties.

1.2 Objectives.

The primary goal of this thesis is to investigate the performances of UBSs by studying the classical token-passing protocol. The performance index will be the average message delay at each node. Unfortunately, a rigorous delay analysis under the most general conditions is very difficult. To provide mathematical tractability, a number of reasonable assumptions are made to simplify the problem. Attention is focused on the gated exhaustive service discipline because it can be analyzed mathematically and also provides insight to the performance analysis of the more practical non-exhaustive service schemes.

1.3 Previous Work.

The problem of establishing the queueing characteristics of a token-ring system has been studied by many authors in the last two decades under various titles (e.g. polling, traffic control, cyclic queues). The long history of this problem indicates both its importance and its intractability. So far, no solutions are available to the general problem, whether exact or approximate, analytic or algorithmic. Thus, we have to use less reliable and hard-to-interpret results obtained from simulations or actual network measurements to study the queueing behaviour of many practical token-ring systems. Nevertheless, rigorous analyses have been carried out for some special although unrealistic cases. It is the hope of many researchers that these analyses can shed some light on the behaviour of more practical systems.

A good summary of the more recent work on token-ring analysis can be found in [29]. With few exceptions, most of the previous work dealt with the gated and the non-gated exhaustive service disciplines with infinite buffering and Poisson arrivals. The special case with only two nodes in the system was studied in [19-21] and exact explicit delay results were obtained. Generalizations to an arbitrary number of nodes were investigated in [22-29]. Moment generating functions of cycle time, queue length and delay distributions have been obtained in [22-26]. Unfortunately, these are not closed-form solutions that facilitate easy evaluation of delay moments. Attempts have been made to derive explicit analytic expressions for the average delays but they were successful for the symmetric case only ([25-27]). For the general asymmetric case, only numerical algorithms

([27-29]) and approximate solutions ([30-33]) are available. Most of the approximations given in the literature are based on rather heuristic arguments and hence their regions of validity are unclear. In general, these approximations break down at very heavy network loadings or under highly asymmetric terminal traffic. Numerical algorithms have the advantage of generating exact results but they cannot readily show the effects of simultaneous changes in the system parameters. An asymmetric token-ring has an enormous parameter space (the set of all possible combinations of parameters) and hence resists analysis by numerical methods. Nevertheless, exact numerical results can serve as absolute references for validating approximations and simulations.

The more general cases of non-exhaustive service disciplines and finite buffering are much more difficult to deal with, as evidenced by the lack of previous work in this area. An approximate treatment of non-exhaustive service disciplines can be found in [34]. Non-exhaustive service disciplines on a two-node system were analyzed in [35]. The system with single-buffer nodes (interactive-user model) was investigated in [38] based on the work from [36-37].

1.4 Outline.

In Chapter 2, a typical UBS called D-net is described to show the resemblance of the UBSs and token-ring systems. Then the queueing model and relevant system parameters are formulated. A number of important random variables are defined and their relationships are discussed.

0 In Chapter 3, the queueing problem is solved by the method of Imbedded Markov Chain. The imbedded chain being considered is the joint terminal service times in a service cycle. A recursive functional equation is derived for the moment generating function of the imbedded chain at steady state. Differentiations of this basic functional equation yield recursive linear equations for the first and second order moments. The existence and uniqueness of solutions to these linear equations are also established.

In Chapter 4, a number of results are derived concerning the solutions of joint moments of the terminal service times. A simple difference equation which has not been reported before is presented. This equation provides the basis for a very good approximation.

In Chapter 5, the relationships between average message delays and the normalized cycle time variances are given. A fundamental relation among the delays, dictated by Kleinrock's conservation law, is also derived. An approximate solution is then developed, followed by a comparison with exact numerical results.

In the conclusion, a summary of results is given with suggestions for future research.

2. DEVELOPMENT OF THE LAN SYSTEM MODEL.

2.1 Description of D-net and its Communication Protocol.

Unidirectional broadcast structures may be different in topology and in many other aspects but their basic operations are very similar. A typical UBS called D-net is described to illustrate the basic communication protocol commonly used for UBSs. Fig. 2.1 shows the architecture of D-net. The network consists of an inbound and an outbound channel. All traffic flows unidirectionally from the outbound channel to the inbound channel. There are N stations in the network, numbered in the order shown in the figure. Each station has three taps: the receiver (R-)tap, the sensor (S-)tap and the transmitter (T-)tap. The R-taps are connected to the inbound channel to receive any broadcast messages. The S-taps and the T-taps are connected to the outbound channel as shown in Fig. 2.1; they are used to sense upstream traffic and to transmit broadcast messages, respectively. The black ball is called the locomotive generator and it has only the R-tap and the T-tap.

To transmit, a station first waits until it senses upstream traffic at its S-tap. The station then looks for the end-of-carrier (EOC). The EOC event is defined as the cessation of signal at the S-tap. It is assumed that it takes t_d seconds to detect this event. After detection of EOC, the station starts transmitting its own packets. While transmitting, a station may sense more incoming upstream traffic through its S-tap. In this case, the station aborts its transmission and waits for the next EOC. Otherwise, the station finishes its transmission. This basic operation will be repeated. Basically, a station sees a 'train' of packets separated by t_d

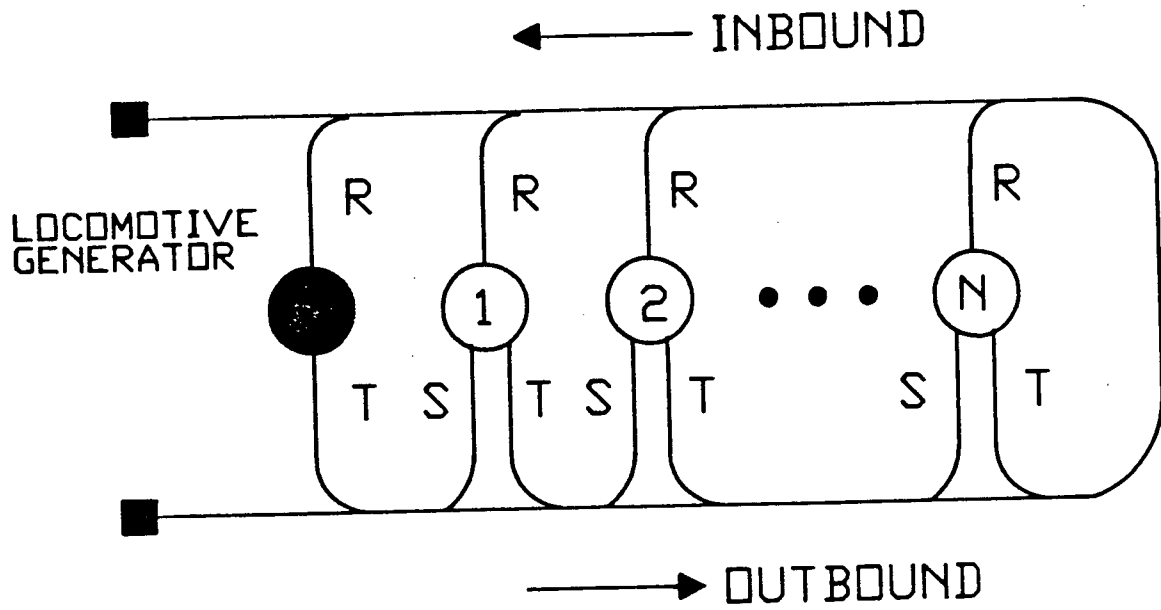


Fig. 2.1 Architecture of D-net.

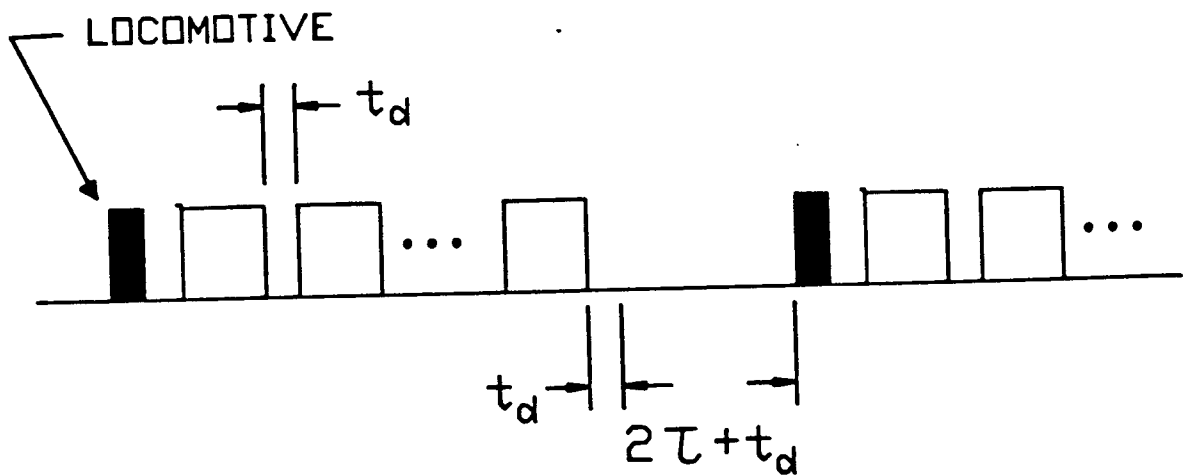


Fig. 2.2 Trains of packets on D-net.

seconds on the channel and each station attaches its packets to the end of this train. The train thus grows in size as it passes through a station with ready packets.

The locomotive generator is used to transmit a short burst called the locomotive to head the train of packets. Without the locomotive generator, the network will stall because no stations will contend for the channel unless there is some traffic. To keep the network 'alive' all the time, the locomotive generator generates a locomotive each time it sees the end-of-train (EOT) at its R-tap. The EOT event signifies the end of a cycle. It is detected when a silence period of $2t_d$ seconds after cessation of signal is noted at the R-tap of the locomotive generator. If normal operation prevails, the traffic on the unidirectional channel will consist of trains of packets separated by $2(\tau + t_d)$ seconds shown in Fig. 2.2 where τ is the propagation delay on either the inbound or the outbound channel and $2t_d$ is the time needed to detect EOT.

It is noted that the D-net unidirectional bus described here is not suitable for fiber optic implementation because of the difficulties with current optical tap technology (e.g. high signal loss). Two modifications of the basic D-net structure are given in [12] to overcome these technological difficulties. The star coupled version of D-net shown in Fig. 2.3 is suggested to enable the insertion loss of R-taps to be lumped in the coupler so that each station receives approximately the same signal power. The open-ring version shown in Fig. 2.4 can be used to avoid the distributed insertion loss at the T-taps. Since each node is active in the open-ring D-net, bypass circuits must be incorporated in case of a node failure.

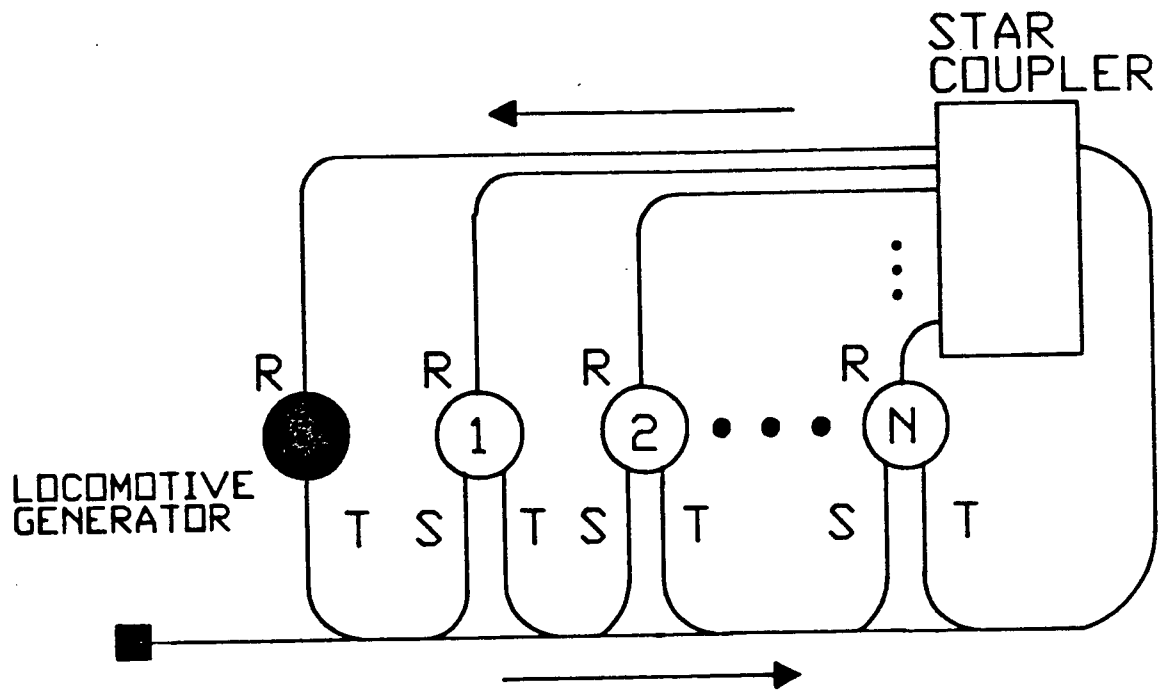


Fig. 2.3 The Star-coupled Version of D-net.

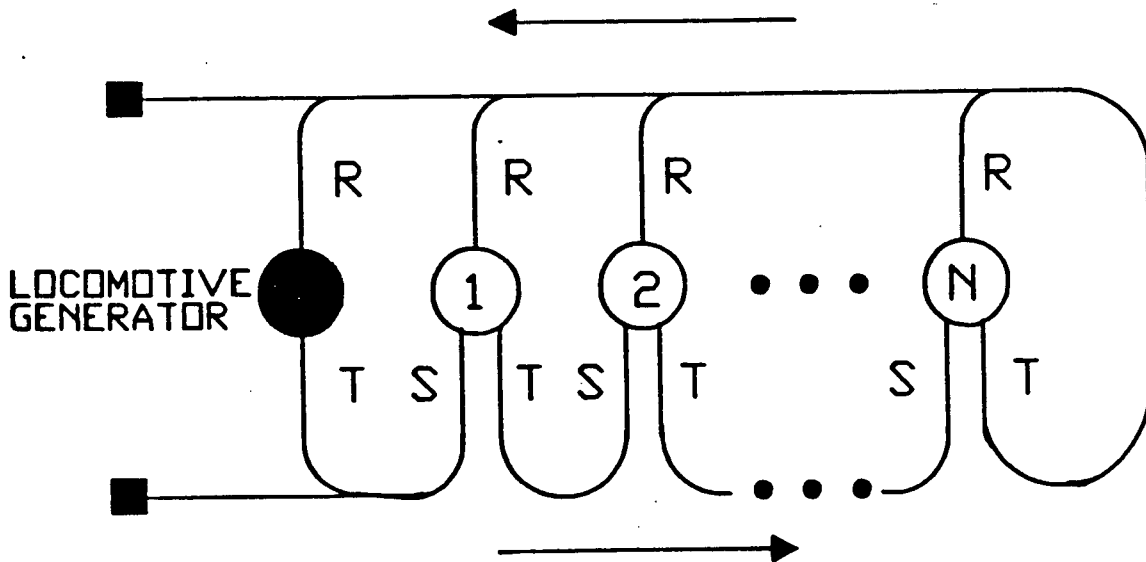


Fig. 2.4 The Open-ring Version of D-net.

2.2 Comparison of D-net with Token-Ring Systems.

By taking advantage of the inherent ordering of the network nodes, the D-net protocol provides orderly network access similar to that of token-passing on a ring. As long as the EOC detection time in D-net is comparable to the token detection time in a token-ring, the two networks will have essentially the same delay performance. From the training sequence of D-net, we may regard the switching overheads as being lumped in the inter-round silence period of $2(\tau + t_d)$ seconds. One advantage of D-net over the conventional distributed token-ring is that the impact of the inter-round overhead can be reduced by interleaving the locomotives. To do this, the locomotive generator must anticipate the length of each train and generate the locomotive not by detecting EOT but by precise timing. However, interleaving locomotives is advantageous only when the inter-round overhead is a significant portion of the cycle time (train + overhead). Calculations based on typical parameters suggest that the interleaving of locomotives is perhaps overskilled for many LAN applications. A more attractive way to reduce the ratio of overhead to cycle time is to allow statistical variations of the station access times according to the network loading, for example by not restricting service to a single packet per node per cycle. The gated exhaustive service discipline under investigation in this thesis is one example of this approach.

The reliability issues surrounding the D-net protocol and token-ring are also different. One observes that the D-net protocol is not completely distributed. The locomotive generator can be regarded as a central

controller whose failure will stall the network. Fortunately, the locomotive generator is a simple device so it can be made redundant with little additional cost. Nevertheless, the D-net protocol avoids the major pitfalls in token-passing such as lost tokens or duplicate tokens. If a locomotive is lost because of channel errors, the locomotive generator will simply time out and generate another locomotive. Considering the complexity of error detection and recovery mechanisms of conventional token-passing, the D-net protocol represents a major improvement over token-passing. Another reliability issue is related to the control passing mechanism in the D-net protocol. Recall that each station attempts to transmit immediately after detecting EOC and aborts its transmission as soon as it detects any incoming traffic at its S-tap. Since it takes a finite amount of time to detect incoming traffic, there will be collisions of a very short duration (time needed to detect incoming traffic) at the beginning of each packet. It is assumed that each packet is headed by a preamble portion for bit synchronization and the collisions mentioned above will not affect the synchronizability of the preamble portion.

2.3 System Model and Parameters.

As we have seen, the basic operation of D-net is identical to the token-ring except for the switching overheads from one station to another. For this reason, the performances of D-net and all other similar UBSs will be studied in this thesis as if they were token-ring systems. Fig. 2.5 shows the schematics of the single-server multi-queue model suitable for the analysis of the token-ring system. N queues, numbered from 1 to N , are

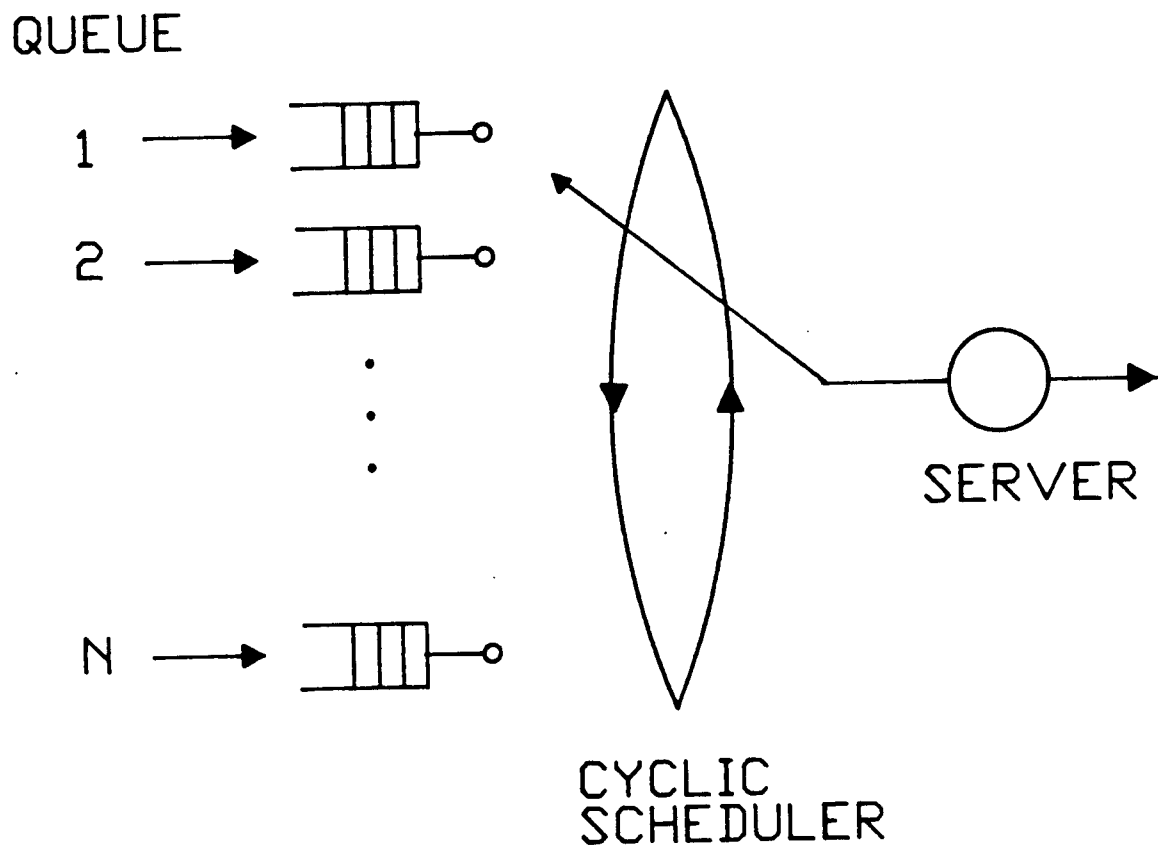


Fig. 2.5 Schematics for the Single-server Multi-queue Model.

attended by a single server in a cyclic order. Further, time is divided into contiguous slots of fixed duration. All time quantities are measured in terms of this fundamental unit. A number of parameters are needed to specify the queueing system, in order to pose a well-defined problem involving dependent queues. These parameters will be discussed in the following paragraphs.

Message arrivals. Poisson message arrival is perhaps the most popular assumption in queueing analysis. A slightly more general type of arrival will be considered here. Messages arriving at a queue within a time slot will be registered at the end of that slot. The number of message arrivals at queue i in one slot is random and its distribution is characterized by the moment generating function (MGF)¹ $\alpha_i(s)$. Arrivals within different time slots are assumed to be independent and the same assumption is made for arrivals at different queues. By letting the slot interval become vanishingly small while keeping the message arrival rate constant, one obtains Poisson message arrivals as a special case of the general independent slot arrivals considered here. Other arrival statistics can also be reasonably approximated by choosing the appropriate slot size and the MGFs $\{\alpha_i(s)\}$.

Message lengths. Message lengths are assumed random with general distributions described by MGFs $\{\beta_i(s)\}$. Although generally distributed, the

¹ The MGF of a random variable X is defined as $F(s) = E[\exp(sX)]$.

length of a message must be an integral multiple of the slot interval. Moreover, message lengths are assumed to be independent of the message arrivals and also independent among different queues.

Service discipline. The service discipline in a multi-queue system includes specifications of the service schedule, order-of-service at each queue and the number of messages to be served from each queue in a service cycle. In a token-ring system, queues are inherently served in a cyclic order (service schedule). Further, a first-come-first-serve discipline will be assumed at each queue (order-of-service). Thus, the service disciplines in this thesis will usually refer to the number of messages served from each queue in a cycle. Various service disciplines are possible; they can be gated or non-gated, exhaustive or non-exhaustive. For gated service disciplines, all messages arriving at a queue after commencement of service at that queue will not be served in that service cycle. Services not gated are called non-gated. The terms 'exhaustive' and 'non-exhaustive' are self-explanatory. In practice, almost all service disciplines are non-exhaustive in nature to avoid the possibility of very long cycles. Unfortunately, the rigorous analysis of non-exhaustive service disciplines appear to be insurmountably difficult. As a result, attention will be focused on the gated exhaustive discipline for mathematical tractability. As well, most non-exhaustive schemes can be approximated by the gated exhaustive scheme for low traffic levels. For the gated exhaustive discipline, all messages buffered at a queue when service starts at that queue will be served in that particular cycle while messages arrived during service will be left in the queue until the next cycle. This is different from the truly exhaus-

tive (non-gated exhaustive) discipline where queues are served till empty in each cycle. The non-gated exhaustive scheme will not be considered here. Interested readers can refer to [22-29] for more detail.

Switching overheads. There is a generally distributed switching overhead from queue i to the next queue characterized by the MGF $\omega_i(s)$. Similar to the message length, the switching overhead must be an integral multiple of the slot interval. Also, all switching overheads are assumed to be independent of each other. The D-net, with constant switching overheads, is an example.

Buffering. Infinite buffering will be assumed throughout this thesis. This is a reasonable assumption as long as the blocking probability is small.

In summary, the system under consideration is a single-server N-queue system. Time is divided into fixed contiguous slots. Queues are served cyclically under a gated exhaustive discipline and a first-come-first-serve order-of-service. Infinite buffering is assumed at each queue. General independent time slot message arrivals are assumed with slot arrivals characterized by the MGF $\alpha_i(s)$ at queue i . Messages arriving at queue i have generally distributed lengths described by the MGF $\beta_i(s)$. The switching overhead from queue i to the next queue is also generally distributed with MGF $\omega_i(s)$.

2.4 Definitions of System Random Variables.

To formulate the mathematical queueing problem, a number of random variables need to be defined. Let

$C_i^{(k)}$ = kth cycle time of queue i, i.e. the time between successive scan instants (instant when service starts) of queue i.

$X_i^{(k)}$ = the number of messages buffered at queue i at its scan instant in the kth cyclic service of that queue.

$Y_i^{(k)}$ = the number of messages arrived at queue i during the kth cycle of queue i.

$Z_i^{(k)}$ = the number of messages served from queue i in the kth cycle of queue i.

$W_i^{(k)}$ = the switching overhead from queue i to the next queue in the kth cycle of queue i.

$T_i^{(k)}$ = the terminal service time of queue i in the kth cycle of queue i, i.e. time for the server to serve queue i and to switch to the next queue.

All the above random variables assume non-negative integer values.

According to the definitions, a number of relationships can be written as follows:

$$Z_i^{(k)} = X_i^{(k)} \text{ for gated exhaustive service} \quad (2.1)$$

$$X_i^{(k+1)} = X_i^{(k)} - Z_i^{(k)} + Y_i^{(k)} \quad (2.2)$$

$$T_i^{(k)} = \sum_{j=1}^{Z_i^{(k)}} L_{i,j}^{(k)} + W_i^{(k)} \quad (2.3)$$

$$C_i^{(k)} = \sum_{j=i}^N T_j^{(k)} + \sum_{j=1}^{i-1} T_j^{(k+1)} \quad (2.4)$$

$$Y_i^{(k)} = \sum_{j=1}^{C_i^{(k)}} M_{i,j}^{(k)} \quad (2.5)$$

where

$L_{i,j}^{(k)}$ = length of the j th message buffered at queue i at its scan instant in the k th cyclic service of queue i .

$M_{i,j}^{(k)}$ = number of messages arrived at queue i in the j th slot of the k th cycle of queue i .

The variables $M_{i,j}^{(k)}$ and $L_{i,j}^{(k)}$ are two sets of independent and identically distributed (i.i.d.) random variables whose statistics are described by the MGFs $\alpha_i(s)$ and $\beta_i(s)$ respectively. When the lower index exceeds the upper index in any summation, the summation is defined as zero. This convention will be adopted throughout this thesis.

2.5 Comments on the Relations between System Random Variables.

Equations (2.1)-(2.5) represent the most basic relations between the system random variables for the gated exhaustive service discipline. For other service disciplines, equations (2.2)-(2.5) still hold but (2.1) has to be modified. For example, if queues are served according to the gated non-exhaustive scheme with queue i allowed to be served up to p_i messages in each cyclic service, then (2.1) will become

$$Z_i^{(k)} = \sum_{j=1}^{p_i} I(X_i^{(k)} > j) \quad (2.6)$$

where $I(\cdot)$ is the indicator function². By definition, $Z_i^{(k)} \leq X_i^{(k)}$ for all gated disciplines. Moreover, there is often a deterministic relation between $Z_i^{(k)}$ and $X_i^{(k)}$ like (2.1) and (2.6) but there are prominent exceptions such as the non-gated exhaustive case.

One notices that equations (2.1)-(2.5) are recursive in nature. Given $X_1^{(k)}, X_2^{(k)}, \dots, X_N^{(k)}$, we can obtain $Z_1^{(k)}, Z_2^{(k)}, \dots, Z_N^{(k)}$ and $T_1^{(k)}, T_2^{(k)}, \dots, T_N^{(k)}$ using (2.1) and (2.3). Then we can obtain $C_1^{(k)}, Y_1^{(k)}$ and $X_1^{(k+1)}$ from (2.4), (2.5) and (2.2). Clearly, $X_2^{(k+1)}, X_3^{(k+1)}, \dots$ can be derived successively in this manner. Thus, (2.1)-(2.5) completely characterize the evolution of the system. We shall use these equations to solve our queueing problem by the method of imbedded Markov chain in the next chapter.

² The indicator function $I(\cdot)$ is defined to be 1 if the statement within brackets is true and 0 otherwise.

3. FORMULATION OF THE SYSTEM EQUATIONS.

3.1 Imbedded Markov Chain Formulation.

The imbedded Markov chain we consider here is the joint random variable of all the terminal service times in a cycle. If we define $\vec{T}^{(k)} = (T_1^{(k)}, T_2^{(k)}, \dots, T_N^{(k)})$, then $\vec{T}^{(k)}$ will be the imbedded chain. The steady state statistics of $\vec{T}^{(k)}$ can be obtained using moment generating functions. Let $P(\vec{s})$ where $\vec{s} = (s_1, s_2, \dots, s_N)$ be the MGF of the random vector $\vec{T}^{(k)}$ as k approaches infinity. Then we can write

$$P(\vec{s}) = \sum_{\vec{t}} \exp(\vec{s} \cdot \vec{t}) \sum_{\vec{u}} p(\vec{t}|\vec{u}) p(\vec{u}) \quad (3.1)$$

where

$$\vec{t} = (t_1, t_2, \dots, t_N), \quad \vec{u} = (u_1, u_2, \dots, u_N),$$

$$\vec{s} \cdot \vec{t} = \sum_{n=1}^N s_n t_n = \text{dot (scalar) product},$$

$$p(\vec{u}) = \lim_{k \rightarrow \infty} \text{Prob}(\vec{T}^{(k)} = \vec{u}),$$

$$p(\vec{t}|\vec{u}) = \text{Prob}(\vec{T}^{(k+1)} = \vec{t} | \vec{T}^{(k)} = \vec{u}).$$

Interchanging the order of summation yields

$$P(\vec{s}) = \sum_{\vec{u}} p(\vec{u}) \sum_{\vec{t}} \exp(\vec{s} \cdot \vec{t}) p(\vec{t}|\vec{u}) \quad (3.2)$$

The transition probabilities can be expressed by

$$p(\vec{t}|\vec{u}) = \prod_{n=1}^N q(t_n | c_n) \quad (3.3)$$

where

$$q(t_n | c_n) = \text{Prob}(T_n^{(k+1)} = t_n | C_n^{(k)} = c_n),$$

$$c_n = \sum_{j=n}^N u_j + \sum_{j=1}^{n-1} t_j.$$

Thus, the summation over \vec{t} in (3.2) can be written as

$$\begin{aligned} \sum_{\vec{t}} \exp(\vec{s} \cdot \vec{t}) p(\vec{t} | \vec{u}) &= \prod_{n=1}^N \left(\sum_{t_n} \exp(s_n t_n) q(t_n | c_n) \right) \\ &= \prod_{n=1}^{N-1} \left(\sum_{t_n} \exp(s_n t_n) q(t_n | c_n) \right) \sum_{t_N} \exp(s_N t_N) q(t_N | c_N) \end{aligned} \quad (3.4)$$

Let L_N and M_N be random variables having distributions of message lengths and arrivals in a slot respectively at queue N . Then the summation over t_N in (3.4) can be evaluated as

$$\begin{aligned} &\sum_{t_N} \exp(s_N t_N) q(t_N | c_N) \\ &= E[\exp(s_N T_N^{(k+1)}) | C_N^{(k)} = c_N] \\ &= E[\exp(s_N L_N^{Z_N^{(k+1)}}) | C_N^{(k)} = c_N] E[\exp(s_N W_N^{(k+1)})] \quad \text{by (2.3)} \\ &= E[\exp(Y_N^{(k)} \ln \beta_N(s_N)) | C_N^{(k)} = c_N] \omega_N(s_N) \quad \text{by (2.1) and (2.2)} \\ &= E[(\exp(M_N \ln \beta_N(s_N)))^{c_N}] \omega_N(s_N) \quad \text{by (2.5)} \\ &= \exp(c_N \ln \alpha_N(\ln \beta_N(s_N))) \omega_N(s_N) \\ &= \exp(c_N \hat{\alpha}_N(s_N)) \omega_N(s_N) \end{aligned} \quad (3.5)$$

where

$$\hat{\alpha}_N(s) = \ln \alpha_N(\ln \beta_N(s)).$$

Substitution of (3.5) into (3.4) gives

$$\begin{aligned}
\sum_{\vec{t}} \exp(\vec{s} \cdot \vec{t}) p(\vec{t} | \vec{u}) &= \prod_{n=1}^{N-1} \left[\sum_{t_n} \exp(s_n t_n) q(t_n | c_n) \right] \exp(c_N \hat{\alpha}_N(s_N)) \omega_N(s_N) \\
&= \prod_{n=1}^{N-2} \left[\sum_{t_n} \exp(s_n t_n) q(t_n | c_n) \right] \sum_{t_{N-1}} \exp(s_{N-1} t_{N-1}) q(t_{N-1} | c_{N-1}) \exp(c_N \hat{\alpha}_N(s_N)) \omega_N(s_N) \\
&= \prod_{n=1}^{N-2} \left[\sum_{t_n} \exp(s_n t_n) q(t_n | c_n) \right] \sum_{t_{N-1}} \exp((s_{N-1} + \hat{\alpha}_N(s_N)) t_{N-1}) q(t_{N-1} | c_{N-1}) \\
&\quad \cdot \exp((c_N - t_{N-1}) \hat{\alpha}_N(s_N)) \omega_N(s_N) \\
&= \prod_{n=1}^{N-2} \left[\sum_{t_n} \exp(s_n t_n) q(t_n | c_n) \right] \exp[c_{N-1} \hat{\alpha}_{N-1}(s_{N-1} + \hat{\alpha}_N(s_N))] \omega_{N-1}(s_{N-1} + \hat{\alpha}_N(s_N)) \\
&\quad \cdot \exp((c_N - t_{N-1}) \hat{\alpha}_N(s_N)) \omega_N(s_N) \tag{3.6}
\end{aligned}$$

The last step is done by noticing $(c_N - t_{N-1})$ is independent of t_{N-1} and by using the same summing technique used to derive (3.5). Continuing in this fashion, we can write

$$\begin{aligned}
\sum_{\vec{t}} \exp(\vec{s} \cdot \vec{t}) p(\vec{t} | \vec{u}) &= \prod_{n=1}^N \exp\left(\left(\sum_{j=n}^N u_j\right) G_n(\vec{s})\right) \prod_{n=1}^N \omega_n(s_n + H_n(\vec{s})) \\
&= \prod_{n=1}^N \exp(u_n F_n(\vec{s})) \prod_{n=1}^N \omega_n(s_n + H_n(\vec{s})) \\
&= \exp(\vec{u} \cdot \vec{F}(\vec{s})) \prod_{n=1}^N \omega_n(s_n + H_n(\vec{s})) \tag{3.7}
\end{aligned}$$

where

$$G_n(\vec{s}) = \hat{\alpha}_n(s_n + H_n(\vec{s})) \tag{3.7a}$$

$$H_n(\vec{s}) = \sum_{k=n+1}^N G_k(\vec{s}), \quad H_N(\vec{s}) = 0 \tag{3.7b}$$

$$F_n(\vec{s}) = \sum_{k=1}^n G_k(\vec{s}) = F_N(\vec{s}) - H_n(\vec{s}) \tag{3.7c}$$

$$\vec{F}(\vec{s}) = (F_1(\vec{s}), F_2(\vec{s}), \dots, F_N(\vec{s})).$$

Substitution of (3.7) into (3.2) gives

$$P(\vec{s}) = \left[\prod_{n=1}^N \omega_n(s_n + H_n(\vec{s})) \right] P(\vec{F}(\vec{s})) \quad (3.8)$$

Eqn. (3.8) is a recursive functional equation that allows us to solve for $P(\vec{s})$. We shall attempt to calculate the first and the second order moments of $\vec{T}^{(k)}$ with $k \rightarrow \infty$ in the next section.

Before leaving this section, we notice that exactly the same procedure can be used to derive the MGF of the joint terminal service times for the time continuous system with Poisson message arrivals. The results will be exactly the same except that all time quantities are expressed in seconds and the arrival MGFs $\{\alpha_i(s)\}$ are replaced by $\{\exp(\lambda_i(\exp(s)-1))\}$ where λ_i is the message arrival rate at queue i .

3.2 Moment calculations.

To calculate the moments of the joint terminal service times, it is convenient to take natural logarithm of both sides of (3.8) and write

$$\hat{P}(\vec{s}) = \sum_{n=1}^N \hat{\omega}_n(s_n + H_n(\vec{s})) + \hat{P}(\vec{F}(\vec{s})) \quad (3.9)$$

where

$$\hat{P}(\vec{s}) = \ln P(\vec{s}), \quad \hat{\omega}_n(s) = \ln \omega_n(s).$$

By differentiating (3.9), we can get the central moment (and hence the moments) of the joint terminal service times. Differentiation of (3.9) once and evaluation at $\vec{s} = 0$ gives

$$\frac{\partial \hat{P}}{\partial s_1} = \sum_{j=1}^N \frac{\partial \hat{P}}{\partial s_j} \frac{\partial F_j}{\partial s_1} + \sum_{j=1}^N \hat{\omega}'_j \frac{\partial (s_j + H_j)}{\partial s_1} \quad (3.10)$$

All derivatives will be evaluated at the origin unless stated otherwise. A second differentiation yields

$$\begin{aligned} \frac{\partial^2 \hat{P}}{\partial s_i \partial s_j} = & \sum_{k=1}^N \sum_{\ell=1}^N \frac{\partial^2 \hat{P}}{\partial s_k \partial s_\ell} \frac{\partial F_k}{\partial s_i} \frac{\partial F_\ell}{\partial s_j} + \sum_{k=1}^N \frac{\partial \hat{P}}{\partial s_k} \frac{\partial^2 F_k}{\partial s_i \partial s_j} \\ & + \sum_{k=1}^N \left[\hat{\omega}_k'' \frac{\partial(s_k + H_k)}{\partial s_i} \frac{\partial(s_k + H_k)}{\partial s_j} + \hat{\omega}_k' \frac{\partial^2(s_k + H_k)}{\partial s_i \partial s_j} \right] \end{aligned} \quad (3.11)$$

Eqns. (3.10) and (3.11) can be readily expressed in matrix form as follows:

$$\vec{v} = A\vec{v} + \vec{b} \quad (3.12)$$

$$V = AVA^T + B \quad (3.13)$$

where

$$\begin{aligned} \vec{v} &= (v_1, v_2, \dots, v_N)^T, \quad \vec{b} = (b_1, b_2, \dots, b_N)^T, \\ v_i &= \frac{\partial \hat{P}}{\partial s_i}, \quad b_i = \sum_{j=1}^N \hat{\omega}_j' \frac{\partial(s_j + H_j)}{\partial s_i}, \\ V &= [v_{ij}], \quad v_{ij} = \frac{\partial^2 \hat{P}}{\partial s_i \partial s_j}, \\ A &= [a_{ij}], \quad a_{ij} = \frac{\partial F_j}{\partial s_i}, \\ B &= [b_{ij}], \quad b_{ij} = \sum_{k=1}^N \frac{\partial \hat{P}}{\partial s_k} \frac{\partial^2 F_k}{\partial s_i \partial s_j} + \sum_{k=1}^N \left[\hat{\omega}_k'' \frac{\partial(s_k + H_k)}{\partial s_i} \frac{\partial(s_k + H_k)}{\partial s_j} \right. \\ &\quad \left. + \hat{\omega}_k' \frac{\partial^2(s_k + H_k)}{\partial s_i \partial s_j} \right]. \end{aligned}$$

3.3 Approaches to the Solutions of the Basic Equations.

There are two approaches to obtaining solutions of (3.12) and (3.13). The first approach is to recognize that (3.12) is a set of N linear equations

of N unknowns $\frac{\partial \hat{p}}{\partial s_i}$ with the forcing terms b_i . Then we notice that (3.13) represents another set of linear equations of $N(N+1)/2$ unknowns $\{v_{ij}\}$ ($v_{ij}=v_{ji}$) with forcing terms $\{b_{ij}\}$; these $\{b_{ij}\}$ are obtained from the solution of (3.12). Hence, we can solve (3.12) and (3.13) by standard techniques for solving systems of linear equations. Naturally, the questions regarding existence and uniqueness of the solutions to both set of equations arise. These questions can be answered by explicit series solutions using the second approach discussed later. The ultimate goal is to solve (3.12) and (3.13) explicitly in closed form. Unfortunately, no such solutions are available for (3.13) except for the symmetric case or for $N=2$.

The second approach is somewhat numerical in nature. We notice that both (3.12) and (3.13) represent recursive relations that allow iterative numerical solutions. A new estimate can be calculated from the immediately previous estimate using the right hand side of (3.12) and similarly for (3.13). If the iterative procedure converges, then we can obtain a solution. The main results obtained following this line of thought can be summarized in Theorem 3.1 in the next section.

3.4 Existence and Uniqueness of Solutions to the Basic Equations.

Theorem 3.1: If the utilization of the system satisfies $\rho < 1$, then unique solutions exist for both (3.12) and (3.13) and they are given by $\vec{v} = \sum_{k=0}^{\infty} A^k \vec{b}$

and $V = \sum_{k=0}^{\infty} A^k B (A^k)^T$ respectively.

Indeed, the theorem can be extended to all higher order derivatives of $\hat{P}(\vec{s})$ but we will restrict our attention to the first two orders of moments. The proof of Theorem 3.1 relies heavily on a result proved in the Appendix stating that if $\rho < 1$, then all the eigenvalues of A have magnitudes less than unity. We prove Theorem 3.1 using this result.

Proof of Theorem 3.1: We diagonalize A by

$$A = \Gamma \Lambda \Gamma^{-1} \quad (3.14)$$

where Λ is a diagonal matrix with elements equal to the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_N$ of A, and Γ contains the corresponding column eigenvectors. Using this diagonalization, we can write

$$\sum_{k=0}^{\infty} A^k \vec{b} = \sum_{k=0}^{\infty} \Gamma \Lambda^k \Gamma^{-1} \vec{b} = \Gamma \left(\sum_{k=0}^{\infty} \Lambda^k \right) \Gamma^{-1} \vec{b} \quad (3.15)$$

$$\begin{aligned} \text{and } \sum_{k=0}^{\infty} A^k B (A^k)^T &= \sum_{k=0}^{\infty} \Gamma \Lambda^k \Gamma^{-1} B (\Gamma^{-1})^T \Lambda^k \Gamma^T \\ &= \Gamma \left(\sum_{k=0}^{\infty} \Lambda^k C \Lambda^k \right) \Gamma^T \end{aligned} \quad (3.16)$$

where

$$C = \Gamma^{-1} B (\Gamma^{-1})^T = [c_{ij}].$$

$$\text{Now } \sum_{k=0}^{\infty} \Lambda^k = \left[\sum_{k=0}^{\infty} \lambda_i^k I(i=j) \right] \text{ and } \sum_{k=0}^{\infty} \Lambda^k C \Lambda^k = \left[\sum_{k=0}^{\infty} c_{ij} \lambda_i^k \lambda_j^k \right].$$

Both series exist if $|\lambda_i| < 1$ for all i and this is guaranteed if $\rho < 1$.

Therefore $\sum_{k=0}^{\infty} A^k \vec{b}$ and $\sum_{k=0}^{\infty} A^k B (A^k)^T$ both exist if $\rho < 1$ and substitution of these series into (3.12) and (3.13) will verify that they are indeed solutions to the equations. Hence the existence of the solutions to (3.12) and to (3.13) has been established.

To establish the uniqueness of those solutions, we let \vec{v} and \vec{u} be solutions to (3.12) and write

$$\vec{v} = A\vec{v} + \vec{b} \quad (3.17)$$

$$\vec{u} = A\vec{u} + \vec{b} \quad (3.18)$$

Subtraction gives

$$(\vec{v} - \vec{u}) = A(\vec{v} - \vec{u}) \quad (3.19)$$

which implies

$$(\vec{v} - \vec{u}) = A^k (\vec{v} - \vec{u}) \text{ for all } k \quad (3.20)$$

but $A^k (\vec{v} - \vec{u}) = \Gamma \Lambda^k \Gamma^{-1} (\vec{v} - \vec{u}) \rightarrow 0$ as $k \rightarrow \infty$ so $\vec{v} = \vec{u}$. Similar arguments establish the uniqueness of the solution to (3.13).

3.5 Comments Regarding Solutions of the Basic Equations.

The series solutions given in Theorem 3.1 are useful in establishing existence and uniqueness, but they do not shed light on the explicit closed-form solutions. Even for numerical solutions, one will follow the iterative procedures given by (3.12) and (3.13) instead of calculating the series in Theorem 3.1 term by term. In fact, a very nice numerical procedure was developed in [28-29]. However, it is desirable to have analytic results that

display the effects of changes in system parameters. In the next chapter, we shall derive a number of exact analytic results.

4. SOME EXACT ANALYTIC RESULTS.

4.1 Overview of Results.

A number of exact analytic results are available regarding the solutions to (3.12) and (3.13). First of all, (3.12) can be solved explicitly in closed form by considering the average cycle time. The solution to (3.12) allows us to simplify the forcing matrix B in (3.13). Next, we explore the structures of the matrices A and B in (3.13). A simple difference equation relating the unknowns $\{v_{ij}\}$ is derived by exploiting the structures of A and B. This difference equation has not been reported before and it will form the basis of the mean delay approximation to be presented in the next chapter.

4.2 Expected Cycle Time.

From equation (2.4), we observe that the expected cycle time at steady state is equal to the sum of all expected terminal service times. This means the average cycle time is independent of the queue. Using this fact, we can compute the average cycle time and obtain other results as well. We state these results precisely by the following Lemma.

Lemma 4.1:

The expected cycle time of queue i at steady state is independent of i and it is given by

$$EC = \sum_{j=1}^N \frac{\partial \hat{P}}{\partial s_j} = \frac{EW}{1-\rho} \quad (4.1)$$

Moreover,

$$\frac{\partial \hat{P}}{\partial s_i} = \rho_i EC + EW_i = \frac{\rho_i EW}{1-\rho} + EW_i \quad (4.2)$$

where

EC = average cycle time,

$EW_i = \hat{\omega}_i'(0)$ = expected overhead from queue i to the next queue,

$EW = \sum_{j=1}^N EW_j$ = total expected overhead,

$\rho_i = \hat{\alpha}_i'(0)$ = utilization of queue i ,

$\rho = \sum_{j=1}^N \rho_j$ = system utilization.

Proof of Lemma 4.1:

From (2.4), we can write

$$\begin{aligned} \lim_{k \rightarrow \infty} EC_i^{(k)} &= \lim_{k \rightarrow \infty} E \left[\sum_{j=i}^N T_j^{(k)} + \sum_{j=1}^{i-1} T_j^{(k+1)} \right] \\ &= \sum_{j=1}^N \lim_{k \rightarrow \infty} ET_j^{(k)} \\ &= \sum_{j=1}^N \frac{\partial \hat{P}}{\partial s_j} \end{aligned} \quad (4.3)$$

Equation (4.3) states that the average cycle time of queue i at steady state is independent of i and we denote this quantity by EC .

From (3.10), we have

$$\frac{\partial \hat{P}}{\partial s_i} = \sum_{j=1}^N \frac{\partial \hat{P}}{\partial s_j} \frac{\partial F_j}{\partial s_i} + \sum_{j=1}^N \hat{\omega}_j' \frac{\partial (s_j + H_j)}{\partial s_i}.$$

Next, we notice from (3.7a) and (3.7b) that G_2, G_3, \dots, G_N and H_1, H_2, \dots, H_N are all independent of s_1 and from (3.7c) that $F_1 = G_1 + G_2 + \dots + G_N$, therefore

$$\begin{aligned} \frac{\partial F_j}{\partial s_1} &= \frac{\partial G_1}{\partial s_1} = \hat{\alpha}_1'(0) \text{ and } \frac{\partial(s_j + H_j)}{\partial s_1} = I(j=1) \\ \Rightarrow \frac{\partial \hat{P}}{\partial s_1} &= \hat{\alpha}_1' \sum_{j=1}^N \frac{\partial \hat{P}}{\partial s_j} + \hat{\omega}_1' \end{aligned} \quad (4.4)$$

From the definitions of $\hat{\alpha}_i$ and $\hat{\omega}_i$, we have

$$\begin{aligned} \hat{\alpha}_i' &= \alpha_i' \beta_i' \\ &= (\text{average arrivals in a slot} \times \text{average message length in slots}) \text{ at} \\ &\quad \text{queue } i \\ &= \rho_i, \\ \hat{\omega}_i' &= \omega_i' = EW_i. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{\partial \hat{P}}{\partial s_1} &= \rho_1 \sum_{j=1}^N \frac{\partial \hat{P}}{\partial s_j} + EW_1 \\ &= \rho_1 EC + EW_1 \end{aligned} \quad (4.5)$$

Similar results can be obtained for all other queues so (4.5) can be rewritten as

$$\frac{\partial \hat{P}}{\partial s_i} = \rho_i EC + EW_i \quad (4.6)$$

Summation of (4.6) over all i gives

$$\begin{aligned} EC &= \rho EC + EW \\ \Rightarrow EC &= \frac{EW}{1-\rho} \end{aligned} \quad (4.7)$$

Combination of (4.3), (4.6) and (4.7) gives (4.1)-(4.2).

We remark here that (4.1)-(4.2) are very general results that can be obtained by considering equations (2.2)-(2.5) only without assuming any particular service discipline. One also notices that we did not solve (3.12) directly to obtain (4.2). Instead, we exploited the cyclic symmetry of the system to get (4.6) and (4.7), which are much simpler than (3.12). This readily demonstrates the importance of cyclic symmetry in our problem.

4.3 Simplification of the Forcing Matrix B.

The solution to (3.12) given by (4.2) can be used to simplify the elements in matrix B to a considerable extent. This is stated in Lemma 4.2.

Lemma 4.2

The elements $\{b_{ij}\}$ of the matrix B in (3.12) can be simplified as

$$b_{ij} = \left[\sum_{k=1}^N t_k \frac{\partial(s_k + H_k)}{\partial s_i} \frac{\partial(s_k + H_k)}{\partial s_j} \right] EC \quad (4.8)$$

where

$$\begin{aligned} t_i &= \hat{\alpha}_i'' + \frac{\hat{\omega}_i''}{EC} \\ &= \rho_i^2 \frac{\text{var}M_i}{(EM_i)^2} + \rho_i \frac{\text{var}L_i}{EL_i} + \frac{\text{var}W_i}{EC} \end{aligned} \quad (4.8a)$$

EM_i = average message arrivals at queue i in a slot,

$\text{var}M_i$ = variance of message arrivals at queue i in a slot,

EL_i = average message length at queue i,

$\text{var}L_i$ = variance of message length at queue i,

$\text{var}W_i$ = variance of switching overhead from queue i to the next queue;

for the time continuous system with Poisson arrivals,

$$\begin{aligned}
t_i &= \hat{\alpha}_i'' + \frac{\hat{\omega}_i''}{EC} \\
&= EM_i \quad EL_i^2 + \frac{\text{var}W_i}{EC}
\end{aligned} \tag{4.8b}$$

EL_i^2 = second moment of the message length at queue i.

Proof of Lemma 4.2:

From the definition of $\{b_{ij}\}$ given in (3.13), we have

$$b_{ij} = \sum_{k=1}^N \left[\frac{\hat{\alpha}_k'}{\partial s_k} \frac{\partial^2 F_k}{\partial s_i \partial s_j} + \hat{\omega}_k' \frac{\partial^2 (s_k + H_k)}{\partial s_i \partial s_j} + \hat{\omega}_k'' \frac{\partial (s_k + H_k)}{\partial s_i} \frac{\partial (s_k + H_k)}{\partial s_j} \right]$$

and substitution of (4.2) gives

$$\begin{aligned}
b_{ij} &= \sum_{k=1}^N \left[(\hat{\alpha}_k' EC + \hat{\omega}_k') \frac{\partial^2 F_k}{\partial s_i \partial s_j} + \hat{\omega}_k' \frac{\partial^2 (s_k + H_k)}{\partial s_i \partial s_j} \right. \\
&\quad \left. + \hat{\omega}_k'' \frac{\partial (s_k + H_k)}{\partial s_i} \frac{\partial (s_k + H_k)}{\partial s_j} \right]
\end{aligned} \tag{4.9}$$

From (3.7a)-(3.7c), we have

$$F_i + H_i = F_N = G_1 + G_2 + \dots + G_N$$

and therefore

$$\begin{aligned}
&(\hat{\alpha}_k' EC + \hat{\omega}_k') \frac{\partial^2 F_k}{\partial s_i \partial s_j} + \hat{\omega}_k' \frac{\partial^2 (s_k + H_k)}{\partial s_i \partial s_j} \\
&= \hat{\alpha}_k' EC \frac{\partial^2 F_k}{\partial s_i \partial s_j} + \hat{\omega}_k' \frac{\partial^2 (F_N - H_k + s_k + H_k)}{\partial s_i \partial s_j} \\
&= \hat{\alpha}_k' EC \frac{\partial^2 F_k}{\partial s_i \partial s_j} + \hat{\omega}_k' \frac{\partial^2 F_N}{\partial s_i \partial s_j} \quad (\because \frac{\partial^2 s_k}{\partial s_i \partial s_j} = 0) \\
&= (\hat{\alpha}_k' EC + \hat{\omega}_k') \frac{\partial^2 F_N}{\partial s_i \partial s_j} - \hat{\alpha}_k' EC \frac{\partial^2 H_k}{\partial s_i \partial s_j}
\end{aligned} \tag{4.10}$$

Summation of (4.10) over all k and substitution into (4.9) give

$$\begin{aligned}
 b_{ij} &= (\rho EC + EW) \frac{\partial^2 F_N}{\partial s_i \partial s_j} - \sum_{k=1}^N \hat{\alpha}_k' EC \frac{\partial^2 H_k}{\partial s_i \partial s_j} \\
 &\quad + \sum_{k=1}^N \hat{\omega}_k'' \frac{\partial(s_k + H_k)}{\partial s_i} \frac{\partial(s_k + H_k)}{\partial s_j} \\
 &= EC \sum_{k=1}^N \left[\frac{\partial^2 G_k}{\partial s_i \partial s_j} - \hat{\alpha}_k' \frac{\partial^2 H_k}{\partial s_i \partial s_j} \right] + \sum_{k=1}^N \hat{\omega}_k'' \frac{\partial(s_k + H_k)}{\partial s_i} \frac{\partial(s_k + H_k)}{\partial s_j} \quad (4.11)
 \end{aligned}$$

From (3.7 a), we have

$$\frac{\partial^2 G_k}{\partial s_i \partial s_j} = \hat{\alpha}_k'' \frac{\partial(s_k + H_k)}{\partial s_i} \frac{\partial(s_k + H_k)}{\partial s_j} + \hat{\alpha}_k' \frac{\partial^2(s_k + H_k)}{\partial s_i \partial s_j}$$

and substitution into (4.11) gives

$$\begin{aligned}
 b_{ij} &= EC \sum_{k=1}^N \left(\hat{\alpha}_k'' + \frac{\hat{\omega}_k''}{EC} \right) \frac{\partial(s_k + H_k)}{\partial s_i} \frac{\partial(s_k + H_k)}{\partial s_j} \\
 &= EC \sum_{k=1}^N t_k \frac{\partial(s_k + H_k)}{\partial s_i} \frac{\partial(s_k + H_k)}{\partial s_j} .
 \end{aligned}$$

4.4 The Jacobian Matrix A.

To study the Jacobian matrix A , we have to look into the properties of the functions $\{F_i\}$, $\{G_i\}$ and $\{H_i\}$. The functions $\{G_i\}$ and $\{H_i\}$ are defined iteratively by (3.7 a) and (3.7 b). We start with H_N and calculate G_N from H_N , then we obtain H_{N-1} from G_N , then G_{N-1} from H_{N-1} , then H_{N-1} from G_{N-1} and G_N , and the iteration goes on. After we have obtained G_1, G_2, \dots, G_N and H_1, H_2, \dots, H_N we compute F_N by summing $\{G_i\}$ over all i and then F_1 by $(F_N - H_1)$. These are the basic steps we follow to calculate the elements in A .

Lemma 4.3:

The Jacobian matrix A in (3.13) can be expressed as

$$A = [\vec{u}_0 - \vec{u}_1 | \vec{u}_0 - \vec{u}_2 | \vec{u}_0 - \vec{u}_3 | \dots | \vec{u}_0 - \vec{u}_N] \quad (4.12)$$

where

$$\begin{aligned} \vec{u}_0 &= \begin{bmatrix} \rho_1 \\ (1+\rho_1)\rho_2 \\ (1+\rho_1)(1+\rho_2)\rho_3 \\ \vdots \\ \rho_{N-1} \\ \left(\prod_{k=1}^{N-1} (1+\rho_k)\right)\rho_N \end{bmatrix}, \quad \vec{u}_1 = \begin{bmatrix} 0 \\ \rho_2 \\ (1+\rho_2)\rho_3 \\ (1+\rho_2)(1+\rho_3)\rho_4 \\ \vdots \\ \rho_{N-1} \\ \left(\prod_{k=2}^{N-1} (1+\rho_k)\right)\rho_N \end{bmatrix}, \quad \vec{u}_2 = \begin{bmatrix} 0 \\ 0 \\ \rho_3 \\ (1+\rho_3)\rho_4 \\ \vdots \\ \rho_{N-1} \\ \left(\prod_{k=3}^{N-1} (1+\rho_k)\right)\rho_N \end{bmatrix}, \\ \dots, \quad \vec{u}_{N-1} &= \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \rho_N \end{bmatrix}, \quad \vec{u}_N = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \end{aligned}$$

Proof of Lemma 4.3:

From (3.7a)-(3.7c), we have

$$\frac{\partial G_j}{\partial s_i} = \hat{a}_j, \quad \frac{\partial(s_j + H_j)}{\partial s_i} = \rho_j(I(i=j) + \frac{\partial H_j}{\partial s_i}) \quad (4.13)$$

$$\frac{\partial H_j}{\partial s_i} = \sum_{k=j+1}^N \frac{\partial G_k}{\partial s_i}, \quad \frac{\partial F_N}{\partial s_i} = \sum_{k=1}^N \frac{\partial G_k}{\partial s_i} \quad (4.14)$$

$$\frac{\partial F_j}{\partial s_i} = \frac{\partial F_N}{\partial s_i} - \frac{\partial H_j}{\partial s_i} \quad (4.15)$$

Eqns. (4.13) and (4.14) imply

$$\frac{\partial H_{j-1}}{\partial s_i} = (1 + \rho_j) \frac{\partial H_j}{\partial s_i} + \rho_j I(i=j) \quad j > 1 \quad (4.16a)$$

$$\text{and } \frac{\partial F_N}{\partial s_i} = (1 + \rho_1) \frac{\partial H_1}{\partial s_i} + \rho_1 I(i=1) \quad (4.16b)$$

It can be seen that

$$\vec{u}_{j-1} = (1 + \rho_j) \vec{u}_j + \rho_j \vec{e}_j \quad (4.17)$$

where \vec{e}_j is the j th unit column vector.

The iterative relations given by (4.16a) - (4.16b) and (4.17) are identical

and since $\vec{u}_N = 0 = \left[\frac{\partial H_N}{\partial s_i} \right]$, we conclude that

$$\vec{u}_j = \left[\frac{\partial H_j}{\partial s_i} \right] \text{ and } \vec{u}_0 = \left[\frac{\partial F_N}{\partial s_i} \right] \quad (4.18)$$

Clearly, (4.18) and (4.15) imply the result given in Lemma 4.3

4.5 An Auxilliary Result.

We derive an auxilliary result before the derivation of the difference equation relating the unknowns $\{v_{ij}\}$ in the next section.

Lemma 4.4:

Let $a_i = \frac{\partial F_N}{\partial s_i}$ and $I(\cdot)$ be the indicator function defined earlier in

Chapter 2. Then

$$(b_{i+1i+1} - \frac{a_{i+1}^2}{a_i^2} b_{ii}) - 2(\frac{a_{i+1}}{a_i} b_{1i+1} - \frac{a_{i+1}^2}{a_i^2} \frac{a_i}{a_1} b_{1i})$$

$$= EC \left[(t_{i+1} + (\rho_{i+1} - \frac{a_{i+1}^2}{a_i^2}) t_i) + 2t_1 \left(\frac{a_2}{a_1} \right)^2 I(i=1) \right] \quad i < N \quad (4.19)$$

Proof of Lemma 4.4:

From the structure of $\frac{\partial H_j}{\partial s_i}$ given by (4.18), we have

$$\begin{aligned} \frac{\partial H_j}{\partial s_{i+1}} &= \frac{(1+\rho_i)\rho_{i+1}}{\rho_i} \frac{\partial H_j}{\partial s_i} \\ &= \frac{a_{i+1}}{a_i} \frac{\partial H_j}{\partial s_i} \quad j < i < N \end{aligned} \quad (4.20)$$

From (4.8),

$$\begin{aligned} b_{ii} &= EC \sum_{k=1}^N t_k \left[\frac{\partial (s_k + H_k)}{\partial s_i} \right]^2 \\ &= EC \left[\sum_{k=1}^{i-1} t_k \left(\frac{\partial H_k}{\partial s_i} \right)^2 + t_i \right] \quad (\because \frac{\partial H_k}{\partial s_i} = 0 \text{ for } k > i) \end{aligned} \quad (4.21)$$

Thus,

$$\begin{aligned} &(b_{i+1i+1} - \frac{a_{i+1}^2}{a_i^2} b_{ii}) \\ &= EC \left[\sum_{k=1}^i t_k \left(\frac{\partial H_k}{\partial s_{i+1}} \right)^2 - \sum_{k=1}^{i-1} t_k \left(\frac{a_{i+1}}{a_i} \frac{\partial H_k}{\partial s_i} \right)^2 + t_{i+1} - \frac{a_{i+1}^2}{a_i^2} t_i \right] \\ &= EC \left[t_i \frac{\partial H_i}{\partial s_{i+1}} + t_{i+1} - \frac{a_{i+1}^2}{a_i^2} t_i \right] \\ &= EC \left[t_{i+1} + (\rho_{i+1} - \frac{a_{i+1}^2}{a_i^2}) t_i \right] \quad i < N \end{aligned} \quad (4.22)$$

Using the same results from (4.18) and (4.8), we have

$$b_{1i} = EC \, t_1 \frac{\partial(s_1 + H_1)}{\partial s_i} \quad (4.23)$$

Thus,

$$\begin{aligned} & \left(\frac{a_{i+1}}{a_1} b_{1i+1} - \frac{a_{i+1}^2}{a_1^2} \frac{a_i}{a_1} b_{1i} \right) \\ &= \frac{a_{i+1}}{a_1} \left(b_{1i+1} - \frac{a_{i+1}}{a_i} b_{1i} \right) \\ &= \frac{a_{i+1}}{a_1} EC \, t_1 \left[\frac{\partial s_1}{\partial s_{i+1}} - \frac{a_{i+1}}{a_i} \frac{\partial s_1}{\partial s_i} + \frac{\partial H_1}{\partial s_{i+1}} - \frac{a_{i+1}}{a_i} \frac{\partial H_1}{\partial s_i} \right] \\ &= -EC \, t_1 \left(\frac{a_2}{a_1} \right)^2 I(i=1) \quad i < N \end{aligned} \quad (4.24)$$

Combination of (4.24) and (4.22) give (4.19).

4.6 A Difference Equation for the Second Joint Moments of Terminal Service Times.

It is difficult, if not impossible, to solve (3.13) in closed form. Analytic solutions are available for the symmetric case and for $N=2$ only. For the symmetric case, (3.13) can be solved quite easily by arguments based on symmetry i.e. $v_{ii} = v_{jj}$ and $v_{ij} = v_{i+k, j+k}$. In fact, v_{ij} is independent of i and j so long as $i \neq j$ for the symmetric case. The algebra involved for the very simple case of $N=2$ is already very tedious and the solution in this case suggests that there is little hope in the search for a simple formula for the general asymmetric case. Numerical techniques and heuristic approximations were developed but their difficulties have already been discussed in Chapter 1. Here, we derive a simple difference equation

relating the unknowns $\{v_{ij}\}$. Although this equation does not allow us to solve (3.13) analytically, it represents a big improvement over existing analytic results.

Lemma 4.5:

The unknowns $\{\tilde{v}_{ij}\}$, modified from $\{v_{ij}\}$, satisfy the following relation:

$$\begin{aligned} & (D_H D_V + \rho_i D_V + \rho_j D_H)(\tilde{v}_{ij}) \\ &= -2 \frac{t_i}{\rho_i} I(j=i) + \frac{t_{i+1}}{\rho_{i+1}} I(j = i+1) \quad i < j < N \end{aligned} \quad (4.25)$$

where

$$\tilde{v}_{ij} = \begin{cases} \frac{v_{ij}}{\rho_i \rho_j^{EC}} & i \neq j \\ \frac{1}{\rho_i^2} \left(\frac{v_{ii}}{EC} - t_i \right) = \frac{\text{var } C_i}{EC} & i = j \end{cases} \quad (4.25a)$$

$\text{var } C_i$ = variance of cycle time of queue i at steady state,

$D_H \equiv$ the forward difference operator in the first index,

$D_V \equiv$ the forward difference operator in the second index;

that is

$$D_H(\tilde{v}_{ij}) = \tilde{v}_{ij} - \tilde{v}_{i+1j},$$

$$D_V(\tilde{v}_{ij}) = \tilde{v}_{ij} - \tilde{v}_{ij+1},$$

$$D_H D_V(\tilde{v}_{ij}) = D_V D_H(\tilde{v}_{ij}) = \tilde{v}_{ij} - \tilde{v}_{ij+1} - \tilde{v}_{i+1j} + \tilde{v}_{i+1j+1}.$$

Proof of Lemma 4.5:

We define \vec{r}_i^T to be the i th row vector of the matrix $[\vec{u}_1 | \vec{u}_2 | \vec{u}_3 | \dots | \vec{u}_N]$,

$a_i = \frac{\partial F}{\partial s_i}$ as in Lemma 4.4 and $\vec{l} = (1, 1, \dots, 1)^T$. Using these definitions, we can write the i th row vector of A as $(a_i \vec{l}^T - \vec{r}_i^T)$. Then from (3.13), we have

$$\begin{aligned} v_{ii} &= (a_i \vec{l}^T - \vec{r}_i^T) V (a_i \vec{l} - \vec{r}_i) + b_{ii} \\ &= a_i^2 \vec{l}^T V \vec{l} - 2a_i \vec{l}^T V \vec{r}_i + \vec{r}_i^T V \vec{r}_i + b_{ii} \end{aligned} \quad (4.26)$$

$$\begin{aligned} \text{and } v_{li} &= (a_l \vec{l}^T - \vec{r}_l^T) V (a_i \vec{l} - \vec{r}_i) + b_{li} \\ &= a_l a_i \vec{l}^T V \vec{l} - a_l \vec{l}^T V \vec{r}_i + b_{li} \quad (\because \vec{r}_l = 0) \end{aligned} \quad (4.27)$$

Elimination of $\vec{l}^T V \vec{r}_i$ using (4.26) and (4.27) gives

$$(v_{ii} - 2 \frac{a_i}{a_l} v_{li}) = -a_i^2 \vec{l}^T V \vec{l} + \vec{r}_i^T V \vec{r}_i + (b_{ii} - 2 \frac{a_i}{a_l} b_{li}) \quad (4.28)$$

Replacement of i by $i+1$ in (4.28) gives

$$(v_{i+li+1} - 2 \frac{a_{i+1}}{a_l} v_{li+1}) = -a_{i+1}^2 \vec{l}^T V \vec{l} + \vec{r}_{i+1}^T V \vec{r}_{i+1} + (b_{i+li+1} - 2 \frac{a_{i+1}}{a_l} b_{li+1}) \quad i < N \quad (4.29)$$

Elimination of $\vec{l}^T V \vec{l}$ using (4.28) and (4.29) gives

$$\begin{aligned} &(v_{i+li+1} - 2 \frac{a_{i+1}}{a_l} v_{li+1}) - \frac{a_{i+1}^2}{a_i^2} (v_{ii} - 2 \frac{a_i}{a_l} v_{li}) \\ &= \vec{r}_{i+1}^T V \vec{r}_{i+1} - \frac{a_{i+1}^2}{a_i^2} \vec{r}_i^T V \vec{r}_i + (b_{i+li+1} - 2 \frac{a_{i+1}}{a_l} b_{li+1}) - \frac{a_{i+1}^2}{a_i^2} (b_{ii} - 2 \frac{a_i}{a_l} b_{li}) \\ &= \vec{r}_{i+1}^T V \vec{r}_{i+1} - \frac{a_{i+1}^2}{a_i^2} \vec{r}_i^T V \vec{r}_i + \tilde{b}_{ii} \quad i < N \end{aligned} \quad (4.30)$$

where

$$\begin{aligned}
\tilde{b}_{ii} &= (b_{i+1 \ i+1} - 2 \frac{a_{i+1}}{a_1} b_{1i+1}) - \frac{a_{i+1}^2}{a_i^2} (b_{ii} - 2 \frac{a_i}{a_1} b_{1i}) \\
&= EC \left[t_{i+1} + (\rho_{i+1} - \frac{a_{i+1}^2}{a_1^2} t_i) + 2t_1 \left(\frac{a_2}{a_1} \right)^2 I(i=1) \right] \quad i < N \quad (4.30a)
\end{aligned}$$

Eqn. (4.30a) is obtained by Lemma 4.4.

Now,

$$\begin{aligned}
\vec{r}_{i+1}^T \vec{v}_{i+1} &- \frac{a_{i+1}^2}{a_i^2} \vec{r}_i^T \vec{v}_i \\
&= \sum_{k=1}^N \sum_{\ell=1}^N \left[\frac{\partial H_k}{\partial s_{i+1}} \frac{\partial H_\ell}{\partial s_{i+1}} - \frac{a_{i+1}^2}{a_i^2} \frac{\partial H_k}{\partial s_i} \frac{\partial H_\ell}{\partial s_i} \right] v_{k\ell} \\
&= \sum_{k=1}^i \sum_{\ell=1}^i \left[\frac{\partial H_k}{\partial s_{i+1}} \frac{\partial H_\ell}{\partial s_{i+1}} - \frac{a_{i+1}^2}{a_i^2} \frac{\partial H_k}{\partial s_i} \frac{\partial H_\ell}{\partial s_i} \right] v_{k\ell} \quad (\because \frac{\partial H_j}{\partial s_i} = 0 \text{ for } j > i) \\
&= \sum_{k=1}^{i-1} \frac{\partial H_k}{\partial s_{i+1}} \frac{\partial H_i}{\partial s_{i+1}} v_{ki} + \sum_{\ell=1}^{i-1} \frac{\partial H_i}{\partial s_{i+1}} \frac{\partial H_\ell}{\partial s_{i+1}} v_{i\ell} + \frac{\partial H_i}{\partial s_{i+1}} \frac{\partial H_i}{\partial s_{i+1}} v_{ii} \\
&\quad \text{by (4.20)} \\
&= 2 \frac{\partial H_i}{\partial s_{i+1}} \sum_{k=1}^{i-1} \frac{\partial H_k}{\partial s_{i+1}} v_{ki} + \left(\frac{\partial H_i}{\partial s_{i+1}} \right)^2 v_{ii} \\
&= 2 \rho_{i+1} \sum_{k=1}^{i-1} \frac{a_{i+1}}{a_i} \frac{\partial H_k}{\partial s_i} v_{ki} + \rho_{i+1}^2 v_{ii} \quad \text{by (4.20)} \\
&= 2 \frac{a_{i+1}}{a_i} \rho_{i+1} \vec{r}_i^T \vec{v}_i + \rho_{i+1}^2 v_{ii} \quad (4.31)
\end{aligned}$$

Substitution of (4.31) into (4.30) gives

$$\begin{aligned}
&(v_{i+1i+1} - 2 \frac{a_{i+1}}{a_1} v_{1i+1}) - \frac{a_{i+1}^2}{a_i^2} (v_{ii} - 2 \frac{a_i}{a_1} v_{1i}) \\
&= 2 \frac{a_{i+1}}{a_i} \rho_{i+1} \vec{r}_i^T \vec{v}_i + \rho_{i+1}^2 v_{ii} + \tilde{b}_{ii} \quad i < N \quad (4.32)
\end{aligned}$$

The next step is crucial in this proof. We notice that all the variables in (4.32) are functions of $\{\rho_i\}$ and $\{t_i\}$. By cyclic symmetry, eqn. (4.32) should hold if we perform a cyclic shift of indices i.e. replace ρ_i by ρ_{i+1} , ρ_N by ρ_1 and similarly for $\{t_i\}$. We denote this cyclic shift of indices by the operator $CS(\cdot)$. We further notice that

$$CS(v_{ij}) = v_{i+1j+1} \quad i, j < N \quad (4.33)$$

$$CS(\rho_i) = \rho_{i+1} \quad i < N \text{ and } CS(\rho_N) = \rho_1 \quad (4.34)$$

$$CS(t_i) = t_{i+1} \quad i < N \text{ and } CS(t_N) = t_1 \quad (4.35)$$

$$CS\left(\frac{a_{i+1}}{a_i}\right) = \frac{a_{i+2}}{a_{i+1}} \quad i < N-1 \quad (4.36)$$

With these in mind, we perform the cyclic shift of indices on (4.32) and get

$$\begin{aligned} & (v_{i+2i+2} - 2 CS\left(\frac{a_{i+1}}{a_1}\right) v_{2i+2}) - \frac{a_{i+2}^2}{a_{i+1}^2} (v_{i+1i+1} - 2 CS\left(\frac{a_i}{a_1}\right) v_{1i+1}) \\ &= 2 \frac{a_{i+2}}{a_{i+1}} \rho_{i+2} CS(\vec{r}_i^T \vec{v}_e)_i + \rho_{i+2}^2 v_{i+1i+1} + CS(\tilde{b}_{ii}) \quad i < N-1 \end{aligned} \quad (4.37)$$

Replacement of i by $i+1$ in (4.32) gives

$$\begin{aligned} & (v_{i+2i+2} - 2 \frac{a_{i+2}}{a_1} v_{1i+2}) - \frac{a_{i+2}^2}{a_{i+1}^2} (v_{i+1i+1} - 2 \frac{a_{i+1}}{a_1} v_{1i+1}) \\ &= 2 \frac{a_{i+2}}{a_{i+1}} \rho_{i+2} \vec{r}_{i+1}^T \vec{v}_e_{i+1} + \rho_{i+2}^2 v_{i+1i+1} + \tilde{b}_{i+1i+1} \quad i < N-1 \end{aligned} \quad (4.38)$$

Noticing that $CS(\tilde{b}_{ii}) = \tilde{b}_{i+1i+1}$ for $1 < i < N-1$, we can subtract (4.37) from

(4.38) and get

$$-2 \left(\frac{a_{i+2}}{a_1} v_{1i+2} - CS\left(\frac{a_{i+1}}{a_1}\right) v_{2i+2} \right) + 2 \frac{a_{i+2}^2}{a_{i+1}^2} \left(\frac{a_{i+1}}{a_1} v_{1i+1} - CS\left(\frac{a_i}{a_1}\right) v_{2i+1} \right)$$

$$= 2 \frac{a_{i+2}}{a_{i+1}} \rho_{i+2} (\vec{r}_{i+1}^T \vec{V}_{i+1} - CS(\vec{r}_i^T \vec{V}_i)) \quad 1 < i < N-1 \quad (4.39)$$

Now,

$$\begin{aligned} & \vec{r}_{i+1}^T \vec{V}_{i+1} - CS(\vec{r}_i^T \vec{V}_i) \\ &= \sum_{k=1}^i \frac{\partial H_k}{\partial s_{i+1}} v_{ki+1} - CS \left(\sum_{k=1}^{i-1} \frac{\partial H_k}{\partial s_i} v_{ki} \right) \\ &= \sum_{k=1}^i \frac{\partial H_k}{\partial s_{i+1}} v_{ki+1} - \sum_{k=1}^{i-1} \frac{\partial H_{k+1}}{\partial s_{i+1}} v_{k+1i+1} \\ &= \sum_{k=1}^i \frac{\partial H_k}{\partial s_{i+1}} v_{ki+1} - \sum_{k=2}^i \frac{\partial H_k}{\partial s_{i+1}} v_{ki+1} \\ &= \frac{\partial H_1}{\partial s_{i+1}} v_{1i+1} \\ &= \frac{a_{i+1}}{1+\rho_1} v_{1i+1} \quad i < N \end{aligned} \quad (4.40)$$

Moreover,

$$CS \left(\frac{a_i}{a_1} \right) = \frac{a_{i+1}}{a_2} \quad i < N \quad (4.41)$$

Substitution of (4.40) and (4.41) into (4.39) gives

$$\begin{aligned} & -2a_{i+2} \left(\frac{v_{1i+2}}{a_1} - \frac{v_{2i+2}}{a_2} \right) + 2a_{i+2} \frac{a_{i+2}}{a_{i+1}} \left(\frac{v_{1i+1}}{a_1} - \frac{v_{2i+1}}{a_2} \right) \\ &= 2a_{i+2} \frac{\rho_{i+2}}{1+\rho_1} v_{1i+1} \quad 1 < i < N-1 \end{aligned} \quad (4.42)$$

Simplification of (4.42) gives

$$(1+\rho_1+\rho_{i+1})\tilde{v}_{1i+1} - (1+\rho_{i+1})\tilde{v}_{2i+1} - (1+\rho_1)\tilde{v}_{1i+2} + \tilde{v}_{2i+2} = 0$$

$$1 < i < N-1 \quad (4.43)$$

$$\therefore (1+\rho_1+\rho_i)\tilde{v}_{1i} - (1+\rho_i)\tilde{v}_{2i} - (1+\rho_1)\tilde{v}_{1i+1} + \tilde{v}_{2i+1} = 0$$

$$2 < i < N \quad (4.44)$$

Similar calculations can be done for $i = 1$ and $i = 2$ and the results are very similar except for the residual terms on the right hand side of the equation.

The results are

$$(1+\rho_1+\rho_2)\tilde{v}_{12} - (1+\rho_2)\tilde{v}_{22} - (1+\rho_1)\tilde{v}_{13} + \tilde{v}_{23} = \frac{t_2}{\rho_2} \quad (4.45)$$

$$(1+2\rho_1)\tilde{v}_{11} - (1+\rho_1)\tilde{v}_{21} - (1+\rho_1)\tilde{v}_{12} + \tilde{v}_{22} = -2 \frac{t_1}{\rho_1} \quad (4.46)$$

Combination of (4.44), (4.45) and (4.46) gives

$$(D_H D_V + \rho_1 D_V + \rho_j D_H)\tilde{v}_{1j}$$

$$= -2 \frac{t_1}{\rho_1} I(j=1) + \frac{t_2}{\rho_2} I(j=2) \quad j < N \quad (4.47)$$

Cyclic shift of indices on (4.47) yields (4.25).

We would also like to show $\tilde{v}_{1i} = \frac{\text{var } C_1}{EC}$ given in eqn. (4.25a). From (4.26), we have

$$v_{11} = a_1 \tilde{f}^T V \tilde{f} + b_{11}$$

$$= \rho_1 \tilde{f}^T V \tilde{f} + t_1 EC.$$

$$\text{Thus } \tilde{v}_{11} EC = \tilde{f}^T V \tilde{f}$$

$$\begin{aligned}
&= \sum_{i=1}^N \sum_{j=1}^N \frac{\partial^2 P}{\partial s_i \partial s_j} \\
&= \sum_{i=1}^N \sum_{j=1}^N \left[\frac{\partial^2 P}{\partial s_i \partial s_j} - \frac{\partial P}{\partial s_i} \frac{\partial P}{\partial s_j} \right] \\
&= \lim_{k \rightarrow \infty} \sum_{i=1}^N \sum_{j=1}^N \{ E[T_i^{(k)} T_j^{(k)}] - E[T_i^{(k)}] E[T_j^{(k)}] \} \\
&= \lim_{k \rightarrow \infty} \{ E[(\sum_{i=1}^N T_i^{(k)})^2] - (E[\sum_{i=1}^N T_i^{(k)}])^2 \} \\
&= \lim_{k \rightarrow \infty} \text{var } C_1^{(k)} \\
\Rightarrow \quad \tilde{v}_{11} &= \frac{\text{var } C_1}{EC} \\
\Rightarrow \quad \tilde{v}_{ii} &= \frac{\text{var } C_i}{EC} \tag{4.48}
\end{aligned}$$

Lemma 4.5 is a major contribution of this thesis. No such simple result relating the joint moments of terminal service times has been reported before. Eqn. (4.25) is a two dimensional linear but spatially variant difference equation. It can be solved up to N arbitrary constants. To see this, we rewrite (4.25) as

$$\frac{1+p_i+p_j}{1+p_i} \tilde{v}_{ij} - \frac{1+p_j}{1+p_i} \tilde{v}_{i+1j} + \frac{1}{1+p_i} \tilde{v}_{i+1j+1} - \frac{1}{1+p_i} \frac{t_{i+1}}{\rho_{i+1}} I(j=i+1) \quad i < j \tag{4.49a}$$

$$\tilde{v}_{ij+1} = \left\{ \begin{aligned} &\frac{1}{2} \left[\frac{1+2\rho_i}{1+p_i} \tilde{v}_{ii} + \frac{1}{1+p_i} \tilde{v}_{i+1i+1} \right] + \frac{1}{1+p_i} \frac{t_i}{\rho_i} \quad i=j \end{aligned} \right. \tag{4.49b}$$

Using (4.49b), we can express \tilde{v}_{i1+1} including $CS(\tilde{v}_{N-1N})$ in terms of the diagonal elements. Then we can use (4.49a) to express other off-diagonal elements and their cyclic-shifted counterparts in terms of the diagonal elements in a successive manner. Thus, we can assign N arbitrary values to

the diagonal elements and generate a solution to (4.25). In other words, (4.25) is not a complete set of equations that characterize a unique solution.

However, we can use (4.48) to generate N equations relating \tilde{v}_{ii} to other off-diagonal elements. This suggests a new numerical algorithm for solving (3.13). We start with a guess of the diagonal elements, compute all other off-diagonal elements by (4.49a) and (4.49b), update the diagonal elements by (4.48) and repeat the same procedure. The criteria for convergence of this numerical procedure has not been established but experiments have shown that not only does the procedure converge for $\rho < 1$, it also converges much faster than the algorithm given in [29].

Eqn. (4.25) is not only useful for numerical solutions, it can also be used to solve (3.13) explicitly for special cases and to obtain approximate solution for the general case. In the next section, we use (4.25) to solve (3.13) for the symmetric case. An approximation based on (4.25) will be presented in the next chapter.

4.7 Solution for the Symmetric Case.

As remarked earlier, (3.13) can be solved easily for the symmetric case. We do this in the following Lemma.

Lemma 4.6

If ρ_i and t_i are independent of i , then the solution of (3.13) is given by (in terms of \tilde{v}_{ij})

$$\tilde{v}_{ij} = \frac{t}{(1+\rho_i)(1-\rho)} + \frac{1}{1+\rho_i} \frac{t_i}{\rho_i} I(i \neq j) \quad (4.50)$$

where

$$\rho = N\rho_i, \quad t = Nt_i.$$

Proof of Lemma 4.6:

By symmetry, $\tilde{v}_{11} = \tilde{v}_{22} = \dots = \tilde{v}_{NN}$.

From (4.49b), we have

$$\tilde{v}_{12} = \tilde{v}_{23} = \dots = \tilde{v}_{N-1N} = \tilde{v}_{ii} + \frac{1}{1+\rho_i} \frac{t_i}{\rho_i} \quad (4.51)$$

From (4.49a), we have

$$\tilde{v}_{13} = \tilde{v}_{24} = \dots = \tilde{v}_{N-2N} = \tilde{v}_{ii} + \frac{1}{1+\rho_i} \frac{t_i}{\rho_i} \quad (4.52)$$

Successive calculations using (4.49a) show that

$$\tilde{v}_{ij} = \tilde{v}_{ii} + \frac{1}{1+\rho_i} \frac{t_i}{\rho_i} \quad i \neq j \quad (4.53)$$

Furthermore,

$$\begin{aligned} \tilde{v}_{11} &= \frac{1}{EC} \vec{1}^T V \vec{1} \\ &= \sum_{i=1}^N \sum_{j=1}^N \rho_i \rho_j \tilde{v}_{ij} + \sum_{i=1}^N t_i \\ &= N(\rho_1 \tilde{v}_{11} + t_1) + N(N-1)\rho_1 \left(\tilde{v}_{11} + \frac{1}{1+\rho_1} \frac{t_1}{\rho_1} \right) \\ &= \rho_1^2 \tilde{v}_{11} + \frac{1+\rho_1}{1+\rho_1} t_1 \\ \therefore \tilde{v}_{11} &= \frac{t}{(1+\rho_1)(1-\rho)} \end{aligned} \quad (4.54)$$

Combination of (4.53) and (4.54) gives (4.50).

5. MEAN DELAY ANALYSIS.

5.1 Relationship between Mean Message Delays and Service Time Moments.

As remarked in Chapter 1, our goal is to analyze the mean message delays on a token-ring. In the last chapter, we concentrated on the solution for the second joint moments of the terminal service times, because there is an intimate relationship between the mean message delays and these moments. This relationship is stated in Lemma 5.1.

Lemma 5.1:

The mean queueing delay (excluding service) of a message at queue i is given by

$$ED_i^{(\text{slot})} = \frac{1}{2} (1+\rho_i)(\tilde{v}_{ii} + EC) + \frac{1}{2} \left(\frac{\text{var } M_i - EM_i}{EM_i} EL_i - 1 \right) \quad (5.1)$$

for the discrete time slot system and

$$ED_i^{(\text{Poisson})} = \frac{1}{2} (1+\rho_i)(\tilde{v}_{ii} + EC) \quad (5.2)$$

for the continuous time system with Poisson arrivals. All quantities in (5.1) and (5.2) are defined earlier in Chapter 4, EC in Lemma 4.1, $\text{var } M_i$, EM_i and EL_i in Lemma 4.2 and \tilde{v}_{ii} in Lemma 4.5.

Recall that the delay in a discrete time slot system is measured in slots whereas the delay in a time continuous system is measured in seconds. Thus, the quantities in (5.1) and (5.2) may have different numerical values even if they represent the same thing. This is important for system conversions such as treating the time continuous system as a limiting case of the discrete time slot system.

Proof of Lemma 5.1:

The derivation of eqn. (5.1) can be found in [27] but we shall include it here to make this thesis self-contained.

Physically, the average message delay at a queue on a token ring is the ratio of the sum of all message delays over many cycles to the number of messages served in the same period of time. Let N_c be the number of cycles we measure the message delays, ED_i^T be the expected total message delays in a cycle and EZ_i be the expected number of messages served in a cycle. Then the average message delay ED_i is given by

$$ED_i = \frac{N_c}{N_c} \frac{ED_i^T}{EZ_i} = \frac{ED_i^T}{EZ_i} \quad (5.3)$$

Quantities ED_i^T and EZ_i are independent of the cycle because we are only interested in the steady state. For this reason, all superscripts k denoting the k th cycle will be dropped.

From (2.1), (2.2) and (2.5), we have

$$EZ_i = EY_i = EC_i EM_i \quad (5.4)$$

The quantity ED_i^T is more difficult to calculate. We sum all the message in a typical cycle and let

C_i^P = previous cycle,

$M_{i,j}^P$ = number of messages arrived in the j th slot of the previous cycle,

Z_i^C = number of messages served in the current cycle,

$L_{i,j}^C$ = the length of the j th message being served in the current cycle.

The delay of a message consists of two parts: waiting time for the beginning of current service cycle and the sum of service times of messages buffered up front of the queue. By separating these components, we can see that the total delay is given by

$$D_i^T = \sum_{j=1}^{C_i^P} (C_i^P - j) M_{i,j}^P + \sum_{j=1}^{Z_i^C} \sum_{k=1}^{j-1} L_{i,k}^C \quad (5.5)$$

Taking expectation of (5.5) gives

$$\begin{aligned} ED_i^T &= E \left[\sum_{j=1}^{C_i^P} (C_i^P - j) EM_i + \sum_{j=1}^{Z_i^C} \sum_{k=1}^{j-1} EL_i \right] \\ &= E \left[\frac{1}{2} C_i^P (C_i^P - 1) EM_i + \sum_{j=1}^{Z_i^C} (j-1) EL_i \right] \\ &= E \left[\frac{1}{2} C_i^P (C_i^P - 1) EM_i + \frac{1}{2} Z_i^C (Z_i^C - 1) EL_i \right] \end{aligned} \quad (5.6)$$

Since $Z_i^C = \sum_{j=1}^{C_i^P} M_{i,j}^P$, therefore

$$\begin{aligned} E[Z_i^C(Z_i^C - 1)] &= E \left[\left(\sum_{j=1}^{C_i^P} M_{i,j}^P \right)^2 - \sum_{j=1}^{C_i^P} M_{i,j}^P \right] \\ &= E \left[C_i^P EM_i^2 + C_i^P (C_i^P - 1) (EM_i)^2 - C_i^P EM_i \right] \\ &= E \left[C_i^P (\text{var } M_i + C_i^P (EM_i)^2 - EM_i) \right] \end{aligned} \quad (5.7)$$

Substitution of (5.7) into (5.6) yields the following after some simplification:

$$ED_i^T = \frac{1}{2} (1 + EM_i EL_i) EC_i^2 EM_i + \frac{1}{2} EC_i (\text{var } M_i - EM_i) EL_i - \frac{1}{2} EC_i EM_i \quad (5.8)$$

Substitution of (5.8) and (5.4) into (5.3) gives (5.1).

Eqn. (5.2) can be derived from (5.1) as a special limiting case. To do this, we first let the number of message arrivals in a slot be Poisson. Then the term $(\text{var } M_i - EM_i)$ becomes zero. Next, we let the time slot go to zero while keeping the absolute value (in seconds) of the cycle time fixed. Thus, the constant $(-\frac{1}{2})$ effectively disappears from (5.1) and (5.1) is converted to (5.2) with the units changed from time slots to seconds.

5.2 A Fundamental Relation for $\{\tilde{v}_{ii}\}$.

Eqns. (5.1) and (5.2) show that \tilde{v}_{ii} is an important element of the mean message delays. In fact, it is the only unknown appearing in the mean delay expressions. Therefore, it is helpful to have as much information on $\{\tilde{v}_{ii}\}$ as possible. Based on Kleinrock's conservation law [15], a simple relation for $\{\tilde{v}_{ii}\}$ can be derived. This is stated in Lemma 5.2.

Lemma 5.2:

The unknowns $\{\tilde{v}_{ii}\}$ satisfy the following equation:

$$\sum_{i=1}^N \rho_i (1 + \rho_i) \tilde{v}_{ii} = \frac{\rho t}{1 - \rho} \quad (5.9)$$

where

$$\rho = \sum_{i=1}^N \rho_i = \text{total system utilization,}$$

$$t = \sum_{i=1}^N t_i .$$

Proof of Lemma 5.2:

Let us call the general time slot system we are considering System A and denote its parameters with superscript A. We now artificially construct System B with parameters superscripted by B from the following specifications:

(i) System B is time continuous with Poisson arrivals,

(ii) $\rho_i^B = \rho_i^A$, $t_i^B = t_i^A$ for all i ,

(iii) $\text{var } w_i^B = 0$, $EW_i^B = \left(\sum_{j=1}^N EW_j^A \right) I(i=N)$.

It is important to realize that System B is realizable i.e. no overspecifications or inconsistencies in the parameters.

Since the numerical values of $\{\tilde{v}_{ii}\}$ depend only on $\{\rho_i\}$ and $\{t_i\}$, we conclude that \tilde{v}_{ii}^A is equal to \tilde{v}_{ii}^B for all i numerically. Furthermore, \tilde{v}_{ii}^B is independent of the fixed overhead w_N^B , which is the only overhead in System B. Thus, if we let EW_N^B go to zero, the mean delay at each queue in System B will be given by $\frac{1}{2} (1+\rho_i^B) \tilde{v}_{ii}^B$. With $EW_N^B=0$, System B is now a work conserving system i.e. the server will not be idle so long as there is a customer in the system. We are now ready to apply the conservation law.

Kleinrock's conservation law states that in a work conserving multi-queue system, the intensity weighted mean of the average delays of all the queues is independent of the way the queues are served. The law also states that for any M/G/1 system with no preemption, the intensity weighted mean of the average delays is given by

$$\sum_{i=1}^N \frac{\rho_i}{\rho} ED_i = \frac{w_0}{1-\rho} \quad (5.10)$$

where

$$W_o = \sum_{i=1}^N \frac{1}{2} \rho_i \frac{EL_i^2}{EL_i} \quad (5.10a)$$

We notice from (4.8b) that

$$t_i^B = \rho_i^B \frac{E(L_i^B)^2}{EL_i^B} \quad (5.11)$$

Substitution of (5.11) into (5.10a) shows

$$W_o^B = \sum_{i=1}^N \frac{1}{2} t_i^B = \frac{1}{2} t^B \quad (5.12)$$

and therefore

$$\sum_{i=1}^N \frac{\rho_i^B}{\rho^B} ED_i^B = \frac{\frac{1}{2} t^B}{1-\rho^B} \quad (5.13)$$

Therefore,

$$\begin{aligned} \sum_{i=1}^N \frac{\rho_i^B}{\rho^B} \frac{1}{2} (1+\rho_i^B) \tilde{v}_{ii}^B &= \frac{\frac{1}{2} t^B}{1-\rho^B} \\ \Rightarrow \sum_{i=1}^N \rho_i^B (1+\rho_i^B) \tilde{v}_{ii}^B &= \frac{\rho^B t^B}{1-\rho^B} \end{aligned} \quad (5.14)$$

Replacing the superscripts B by A gives the desired result.

5.3 An Approximation for $\{\tilde{v}_{ii}\}$.

To obtain exact solutions for $\{\tilde{v}_{ii}\}$, (3.13) or equivalent equations must be solved. We cannot provide a closed-form analytic solution. However, a very good approximation based on Lemma 4.5 can be derived. We start by considering the difference $(\tilde{v}_{22} - \tilde{v}_{11})$. From (4.26) or (4.27), we have

$$\tilde{v}_{11} = \frac{1}{EC} \vec{I}^T V \vec{I} \quad (5.15)$$

and

$$\tilde{v}_{22} = \frac{1}{EC} CS(\vec{I}^T V \vec{I}) \quad (5.16)$$

where $CS(\cdot)$ is the cyclic shift operator defined in Sec. 4.6.

Subtraction gives

$$\begin{aligned} (\tilde{v}_{22} - \tilde{v}_{11}) &= \frac{1}{EC} (CS(\vec{I}^T V \vec{I}) - \vec{I}^T V \vec{I}) \\ &= \frac{2}{EC} \left[\sum_{j=1}^{N-1} CS(v_{jN}) - \sum_{j=2}^N v_{1j} \right] \\ &= 2 \left[\sum_{j=1}^{N-1} CS(\rho_j \rho_N \tilde{v}_{jN}) - \sum_{j=2}^N \rho_1 \rho_j \tilde{v}_{1j} \right] \\ &= 2 \left[\sum_{j=1}^{N-1} \rho_1 \rho_{j+1} CS(\tilde{v}_{jN}) - \sum_{j=2}^N \rho_1 \rho_j \tilde{v}_{1j} \right] \\ &= 2\rho_1 \sum_{j=1}^{N-1} \rho_{j+1} [CS(\tilde{v}_{jN}) - \tilde{v}_{1j+1}] \end{aligned} \quad (5.17)$$

Now, assuming that the total utilization is low, all the coefficients

appearing in (4.49a)-(4.49b) can be essentially replaced by unity and thus

$$\begin{aligned} \tilde{v}_{1j} - \tilde{v}_{i+1j} + \tilde{v}_{i+1j+1} - \frac{t_{j+1}}{\rho_{j+1}} I(j=i+1) \quad i < j \\ \tilde{v}_{ij+1} \approx \left\{ \begin{aligned} &\frac{1}{2} (\tilde{v}_{ii} + \tilde{v}_{i+1i+1}) + \frac{t_i}{\rho_i} \quad i = j \end{aligned} \right. \end{aligned} \quad (5.18)$$

Successive calculations of the off-diagonal elements using (5.18) gives

$$\tilde{v}_{ij} \approx \frac{1}{2} (\tilde{v}_{ii} + \tilde{v}_{jj}) + \frac{t_i}{\rho_i} \quad i < j \quad (5.19)$$

Substitution of (5.19) into (5.17) gives

$$\begin{aligned} (\tilde{v}_{22} - \tilde{v}_{11}) &\approx 2\rho_1 \sum_{j=1}^{N-1} \rho_{j+1} \left[CS\left(\frac{1}{2} (\tilde{v}_{jj} + \tilde{v}_{NN}) + \frac{t_j}{\rho_j}\right) - \frac{1}{2} (\tilde{v}_{11} + \tilde{v}_{j+1j+1}) - \frac{t_1}{\rho_1} \right] \\ &= 2\rho_1 \sum_{j=1}^{N-1} \rho_{j+1} \left(\frac{t_{j+1}}{\rho_{j+1}} - \frac{t_1}{\rho_1} \right) \end{aligned}$$

$$\begin{aligned}
&= 2\rho_1 \left[\sum_{j=1}^{N-1} t_{j+1} - \frac{t_1}{\rho_1} \sum_{j=1}^{N-1} \rho_{j+1} \right] \\
&= 2\rho_1 \left[(t - t_1) - \frac{t_1}{\rho_1} (\rho - \rho_1) \right] \\
&= 2\rho t \left(\frac{\rho_1}{\rho} - \frac{t_1}{t} \right)
\end{aligned} \tag{5.20}$$

By cyclic symmetry

$$\tilde{v}_{i+1i+1} - \tilde{v}_{ii} \approx 2\rho t \left(\frac{\rho_i}{\rho} - \frac{t_i}{t} \right) \tag{5.21}$$

Using (5.21), the left hand side of (5.9) can be computed as follows:

$$\begin{aligned}
\sum_{i=1}^N \rho_i(1+\rho_i) \tilde{v}_{ii} &\approx \sum_{i=1}^N \rho_i(1+\rho_i) \left[\tilde{v}_{11} + 2\rho t \sum_{j=1}^{i-1} \left(\frac{\rho_j}{\rho} - \frac{t_j}{t} \right) \right] \\
&= \sum_{i=1}^N \rho_i(1+\rho_i) \tilde{v}_{11} + 2\rho t \sum_{i=1}^N \sum_{j=1}^{i-1} \rho_i(1+\rho_i) \left(\frac{\rho_j}{\rho} - \frac{t_j}{t} \right)
\end{aligned} \tag{5.22}$$

Thus, (5.9) can be used to calculate \tilde{v}_{11} as

$$\tilde{v}_{11} \approx \left[\frac{\rho t}{1-\rho} - 2\rho t \sum_{i=1}^N \sum_{j=1}^{i-1} \rho_i(1+\rho_i) \left(\frac{\rho_j}{\rho} - \frac{t_j}{t} \right) \right] / \sum_{i=1}^N \rho_i(1+\rho_i) \tag{5.23}$$

By cyclic shifting of indices, we can compute $\tilde{v}_{22}, \tilde{v}_{33}, \dots, \tilde{v}_{NN}$.

Unfortunately, the double sum in (5.23) cannot be simplified for general cases so we have to leave it as it is.

5.4 Comments on the Approximation for $\{\tilde{v}_{ii}\}$.

It is important to establish or have some indication on the region of validity of our approximation given by (5.21) and (5.23). Clearly, our approximation is exact for the symmetric case. From the low-traffic assumption we made to establish (5.18), our approximation should also be very

good at low traffic levels. Although (5.21) is not exact, it suggests that the differences among $\{\tilde{v}_{ii}\}$ are small compared with the actual values of $\{\tilde{v}_{ii}\}$ at very high traffic levels ($\rho \rightarrow 1$). With this assumption and using Lemma 5.2, we can write

$$\tilde{v}_{ii} \approx \frac{\rho t}{1-\rho} / \sum_{i=1}^N \rho_i (1+\rho_i) \quad (5.24)$$

for ρ close to unity. This matches the dominant term in (5.23). Thus, we expect (5.23) to work well at high traffic levels as well. As a matter of fact, comparisons with exact numerical results have shown that the approximation given by (5.21) and (5.23) is very good over a very wide range of the parameters $\{\rho_i\}$ and $\{t_i\}$. These comparisons will be discussed in the next section.

A number of interesting properties of $\{\tilde{v}_{ii}\}$ can be deduced from (5.21) and (5.23). Eqn. (5.23) shows that \tilde{v}_{ii} consists of two parts. The first part is a hyperbolic function independent of i that grows unbounded as $\rho \rightarrow 1$. The other part is the position dependent part given by a double sum. Eqn. (5.23) is also linear in t_i , which conforms with the linearity of the exact equation (3.13) with respect to t_i . Eqn. (5.21) gives a somewhat different flavour about $\{\tilde{v}_{ii}\}$; it shows the relation of \tilde{v}_{ii} 's between adjacent queues. For example, (5.21) suggests that a queue 'suffers' if the preceding queue has a large fraction of the total utilization but 'benefits' if that preceding queue has a large t_i . If t_i is directly proportional to ρ_i , (5.21) predicts that \tilde{v}_{ii} is approximately independent of i . This special case corresponds to the time continuous system with Poisson message arrivals,

fixed overheads and a constant message length for the whole system.

5.5 Comparison of Approximation and Exact Numerical Results for $\{\tilde{v}_{ii}\}$.

In this section, we compare the approximations given by (5.21) and (5.23) with exact numerical results. The biggest problem in the comparison is the search for representative cases, because we cannot possibly do the comparison for all conceivable combinations of parameters. Besides the number of queues N in the system, the parameters needed to solve (3.13) are $\{\rho_i\}$ and $\{t_i\}$. These parameters cannot take on arbitrary values. First of all, they are all positive. Secondly, the total utilization ρ of the system should be less than unity. Even with these restrictions, the parameter space is still unmanageably large.

Since the approximation is exact for the symmetric case, we certainly should look into highly asymmetric cases. Furthermore, we notice from (4.8a) that

$$t_i = \rho_i^2 \frac{\text{var } M_i}{(EM_i)^2} + \rho_i \frac{\text{var } L_i}{EL_i} + \frac{\text{var } W_i}{EC}$$

This suggests that t_i can be very roughly regarded as a linear combination of ρ_i^2 , ρ_i , 1 if the normalized variances $\frac{\text{var } M_i}{(EM_i)^2}$, $\frac{\text{var } L_i}{EL_i}$ and $\frac{\text{var } W_i}{EC}$ do not vary appreciably among the queues. For this reason, we shall look into cases for $t_i \propto \rho_i^2$, $t_i \propto \rho_i$ and $t_i \propto 1$. With these considerations in mind, we do the comparisons for all possible combinations of the following parameters:

(i) $N = 8, 16$.

(ii) $t_i \propto \rho_i^2, \rho_i, 1, 1/\rho_i, 1/\rho_i^2$.

(iii) $0.01 < \rho < 0.99$ with steps of 0.01.

(iv) 1 large user with utilization 100 times that of other identical users,

1 small user with utilization 0.01 times that of other identical users,

2 clusters of identical users with utilization ratio of 1:5.

In each case, the relative errors (absolute value) between the approximation and the exact numerical solution are computed for all queues; then the maximum and the average of these relative errors among the queues are plotted against the total utilization. There are five curves, labelled from 1 to 5, in each graph. Curve 1 corresponds to the case of $t_i \propto \rho_i^2$, 2 to $t_i \propto \rho_i$, 3 to $t_i \propto 1$, 4 to $t_i \propto 1/\rho_i$, 5 to $t_i \propto 1/\rho_i^2$. The plots are shown at the end of this section. Some of the curves in these plots are too close to be labelled separately so they will be labelled as a group.

As expected, our approximation works well under very light and very heavy traffic, as shown in the plots. The plots also show that the approximation is very good for $t_i \propto \rho_i^2$, $t_i \propto \rho_i$, $t_i \propto 1$ as well as for all cases of a single large user. When $t_i \propto 1/\rho_i$ or $t_i \propto 1/\rho_i^2$, the maximum relative errors at medium system utilizations of 0.7-0.8 can be as high as 12%. Fortunately, these are rather unrealistic cases. For example, in the D-net protocol, $\text{var } W_i = 0$ and so from (4.8a)

$$t_i = \rho_i^2 \frac{\text{var } M_i}{(EM_i)^2} + \rho_i \frac{\text{var } L_i}{EL_i}.$$

Therefore, t_i tends to be small if ρ_i is small and this contradicts the inverse relations given by $t_i \propto 1/\rho_i$ and $t_i \propto 1/\rho_i^2$. Even with these

unrealistic system parameters, our approximation can still give reasonably accurate results (errors of about 10%). For more realistic parameters, our approximation is excellent (errors of 1 to 2%). Thus, it is not unreasonable to treat our approximation as though it were exact for many applications.

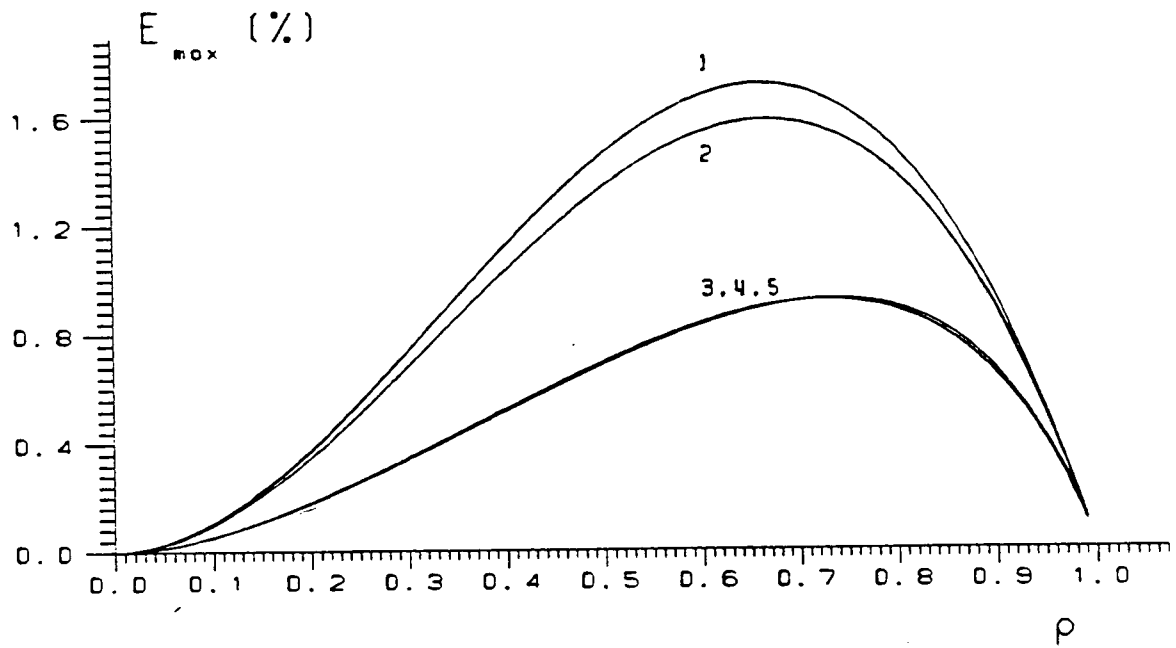


Fig. 5.1a E_{\max} versus ρ for $N=8$ and 1 single large user.

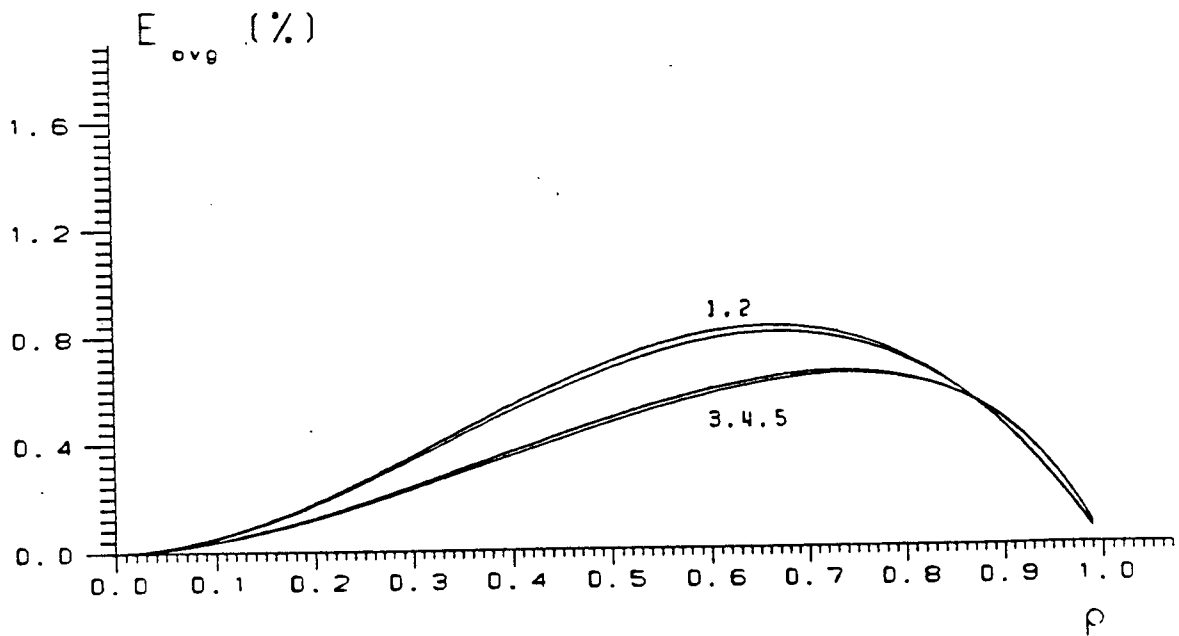


Fig. 5.1b E_{avg} versus ρ for $N=8$ and 1 single large user.

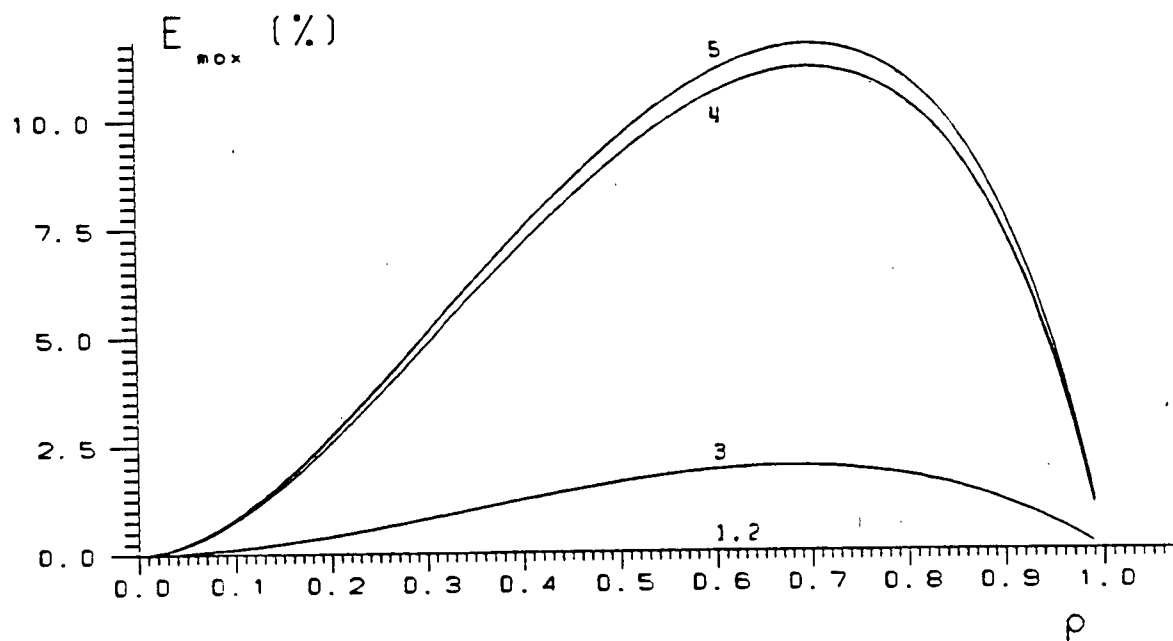


Fig. 5.2a E_{\max} versus ρ for $N=8$ and 1 single small user.

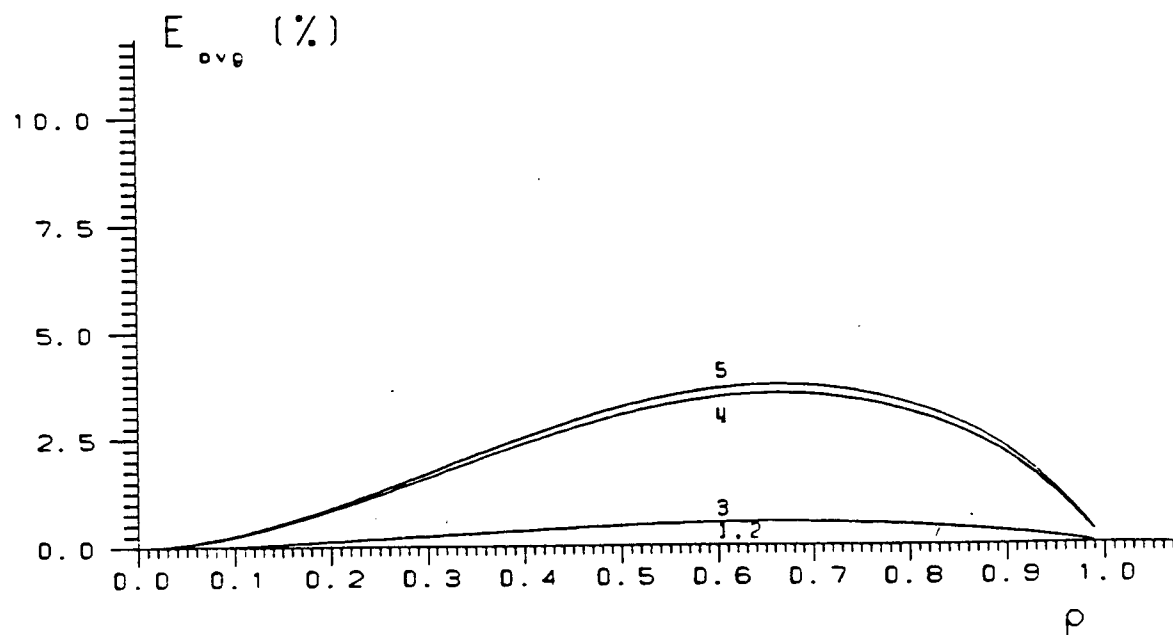


Fig. 5.2b E_{avg} versus ρ for $N=8$ and 1 single small user.

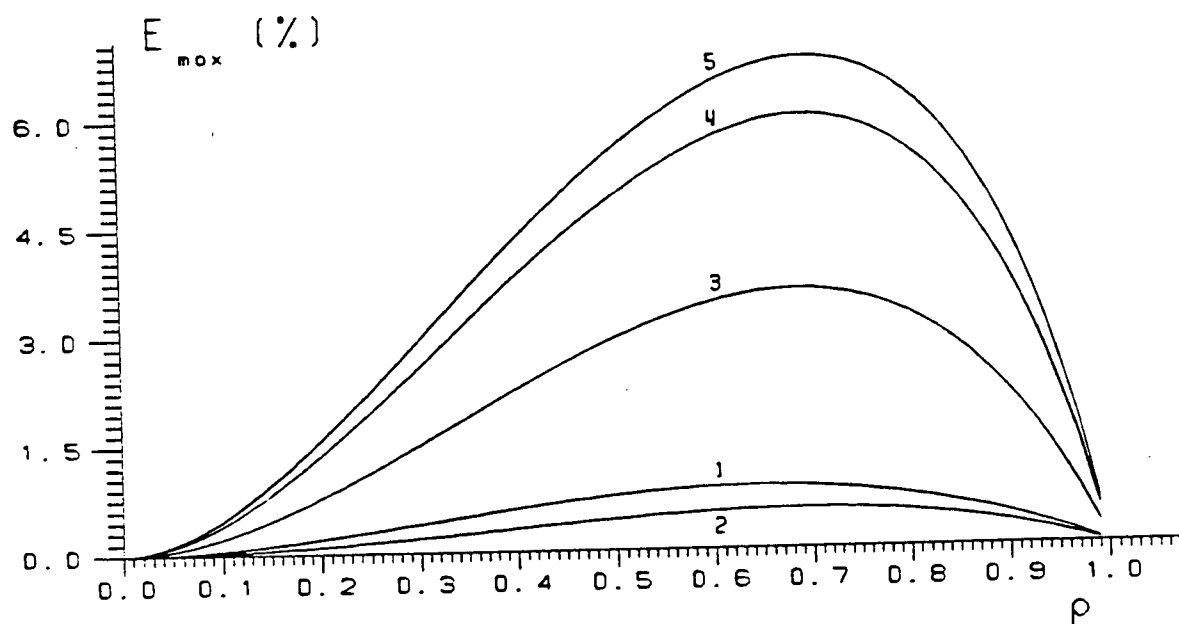


Fig. 5.3a E_{\max} versus ρ for $N=8$ and 2 clusters of identical users.

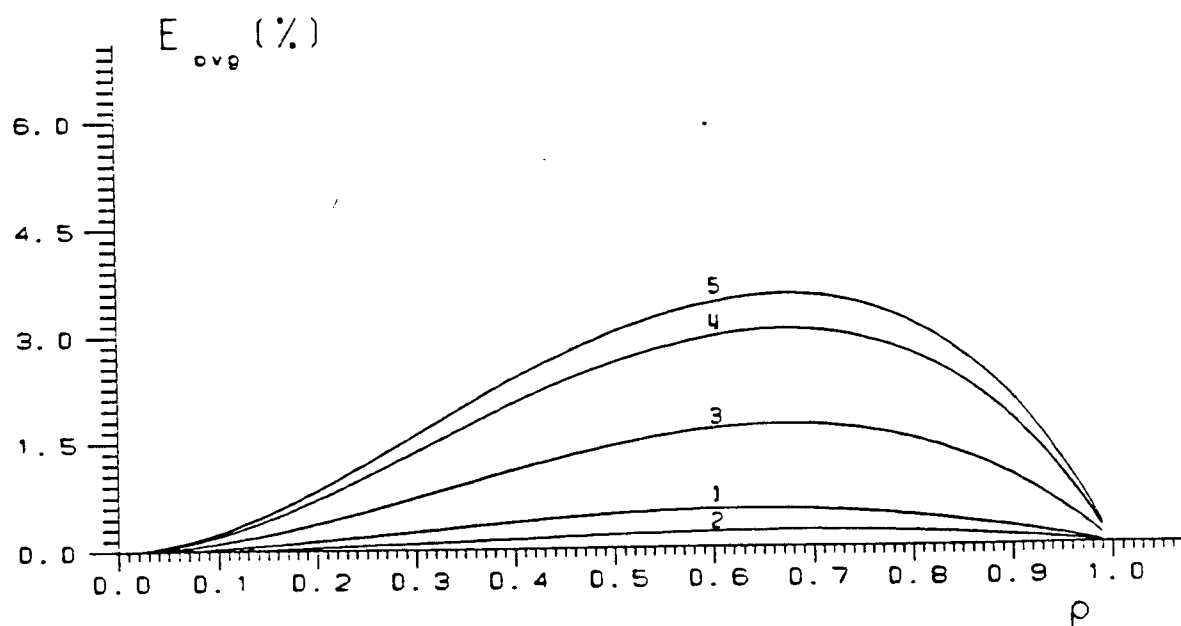


Fig. 5.3b E_{avg} versus ρ for $N=8$ and 2 clusters of identical users.

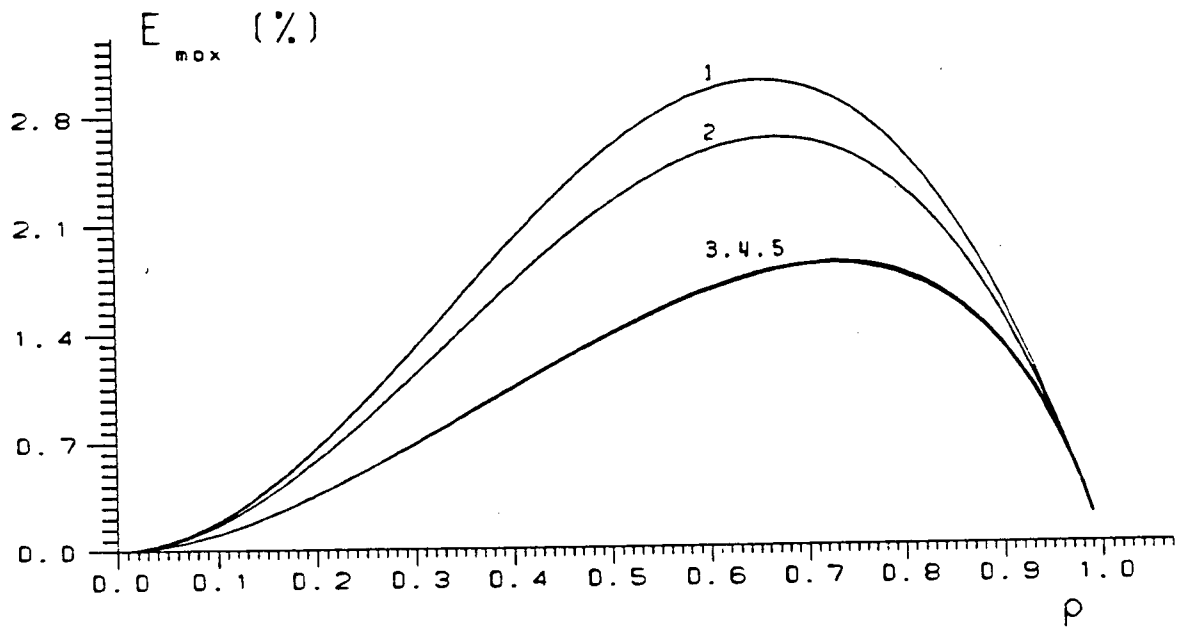


Fig. 5.4a E_{\max} versus ρ for $N=16$ and 1 single large user.

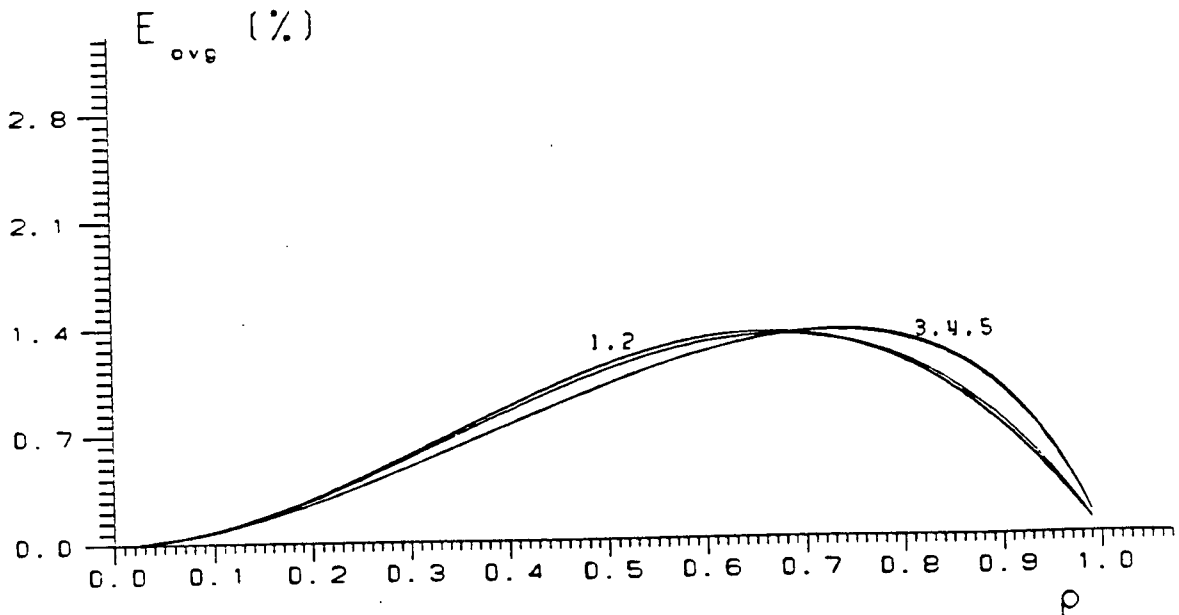


Fig. 5.4a E_{avg} versus ρ for $N=16$ and 1 single large user.

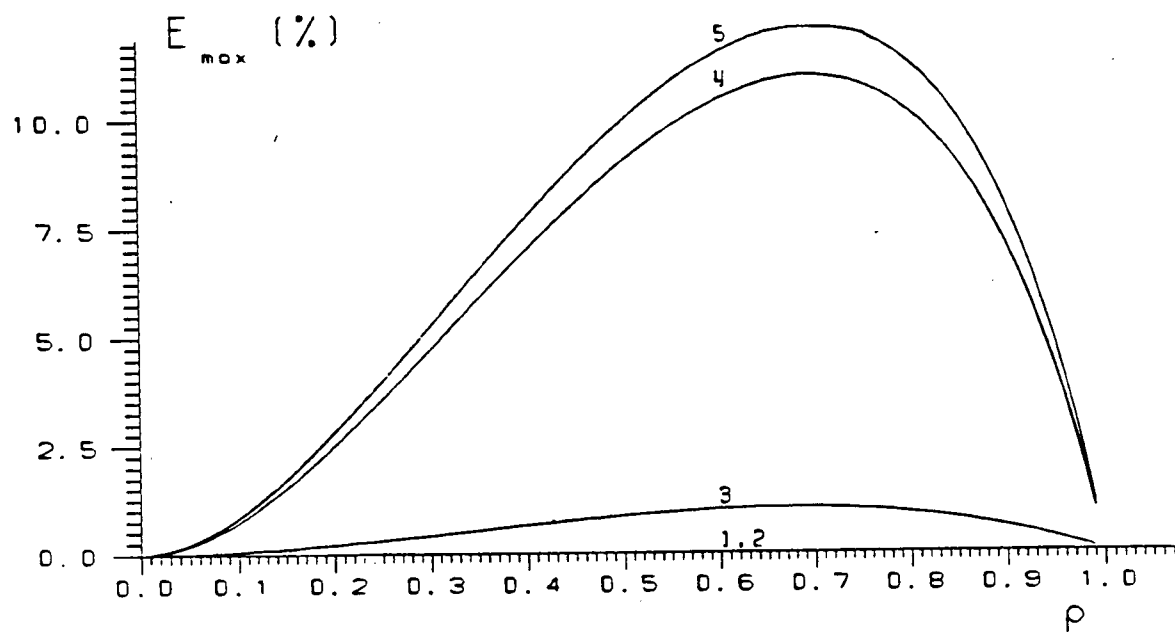


Fig. 5.5a E_{\max} versus ρ for $N=16$ and 1 single small user.

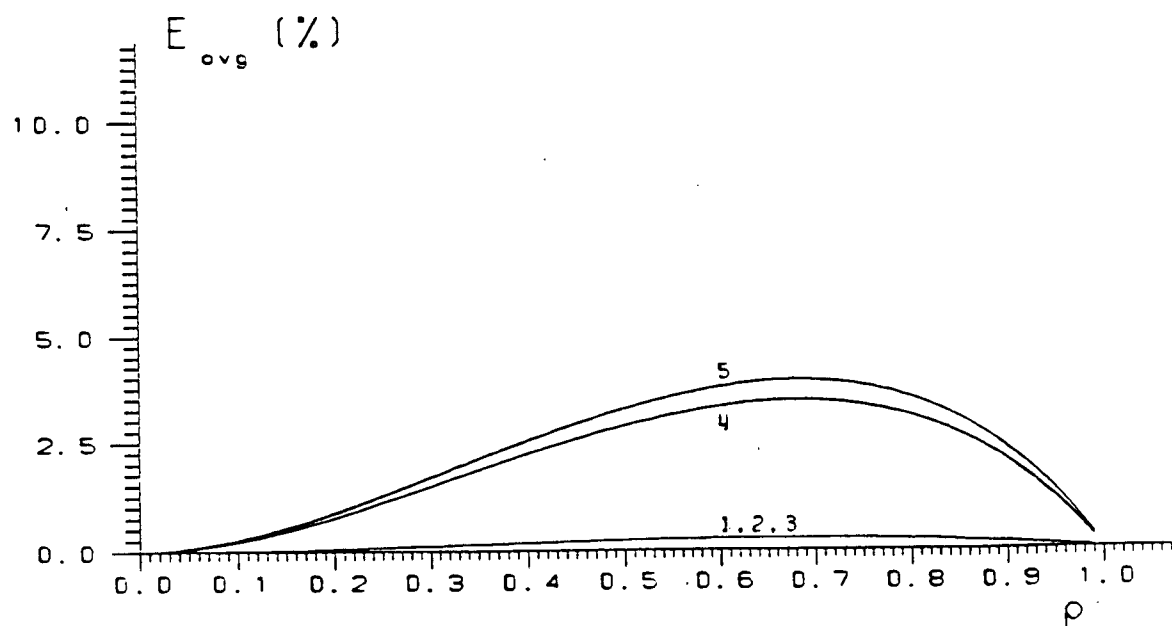


Fig. 5.5b E_{avg} versus ρ for $N=16$ and 1 single small user.

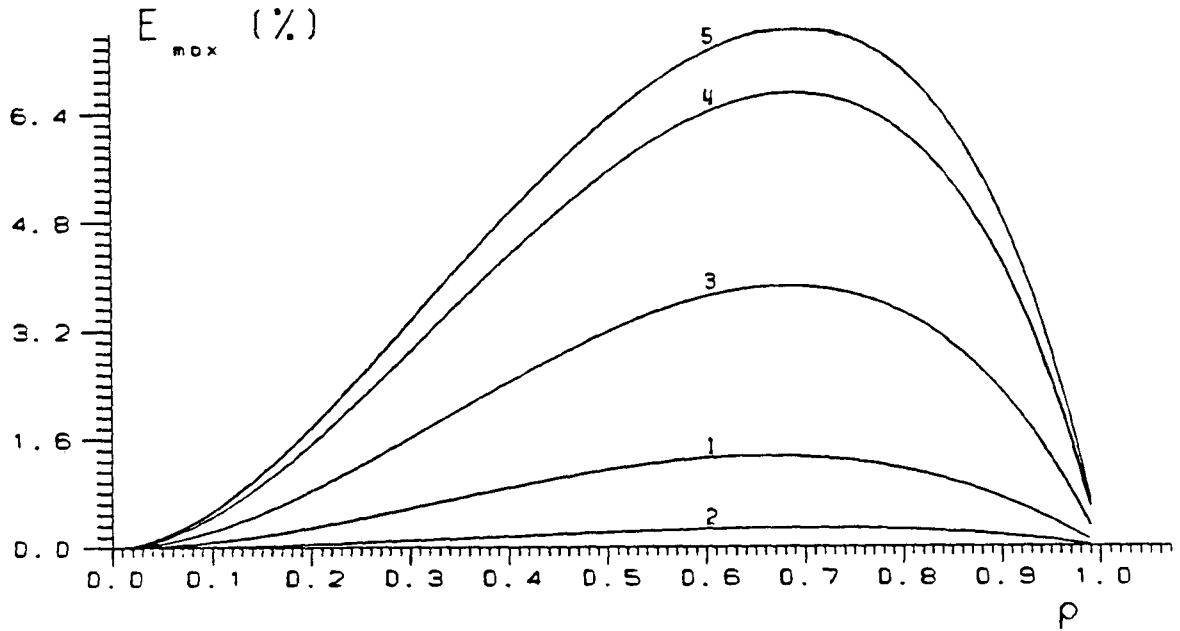


Fig. 5.6a E_{\max} versus ρ for $N=16$ and 2 clusters of identical users.

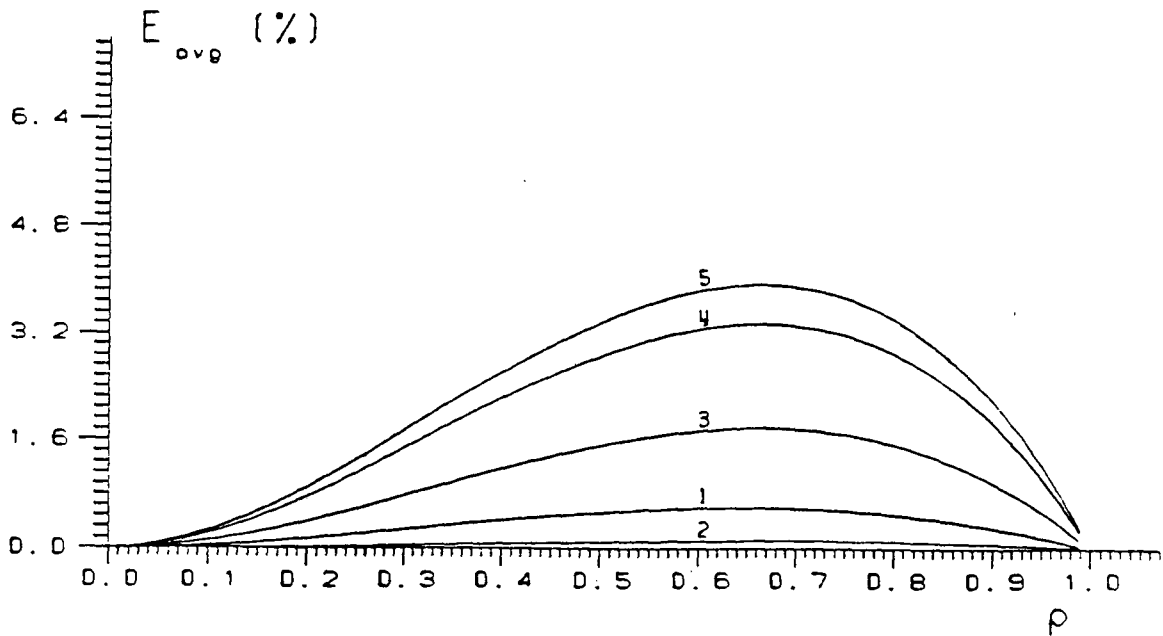


Fig. 5.6b E_{avg} versus ρ for $N=16$ and 2 clusters of identical users.

5.6 Delay-vs-Throughput Results.

Eqns. (5.1) and (5.2) show the relation between average message delays and the unknowns $\{\tilde{v}_{ii}\}$ while eqns. (5.21) and (5.23) give approximate expressions for $\{\tilde{v}_{ii}\}$. By combining all these equations, we can obtain an approximate solution for the average message delays. We copy (5.1), (5.2), (5.21) and (5.23) here for easy reference:

$$ED_i^{(\text{slot})} = \frac{1}{2} (1+\rho_i)(\tilde{v}_{ii} + EC) + \frac{1}{2} \left(\frac{\text{var } M_i - EM_i}{EM_i} EL_i - 1 \right) \quad (5.1)$$

$$ED_i^{(\text{Poisson})} = \frac{1}{2} (1+\rho_i)(\tilde{v}_{ii} + EC) \quad (5.2)$$

$$\tilde{v}_{i+1i+1} - \tilde{v}_{ii} \approx 2\rho t \left(\frac{\rho_i}{\rho} - \frac{t_i}{t} \right) \quad (5.21)$$

$$\tilde{v}_{11} \approx \left[\frac{\rho t}{1-\rho} - 2\rho t \sum_{i=1}^N \sum_{j=1}^{i-1} \rho_i (1+\rho_i) \left(\frac{\rho_j}{\rho} - \frac{t_j}{t} \right) \right] / \sum_{i=1}^N \rho_i (1+\rho_i) \quad (5.23)$$

There are a number of special cases worth attention.

We first consider the case where \tilde{v}_{ii} is independent or approximately independent of i . With this assumption and using Lemma 5.2, we get (5.24).

We can rewrite (5.24) as

$$\tilde{v}_{ii} \approx \frac{t}{1-\rho} / \left(1 + \sum_{i=1}^N \rho_i^2 / \rho \right) \quad (5.25)$$

Substitution of (5.25) into (5.1) and (5.2) gives

$$ED_i^{(\text{slot})} \approx \frac{1}{2} \frac{\eta_i t_{\text{new}}}{1-\rho} + \frac{1}{2} \left(\frac{\text{var } M_i - EM_i}{EM_i} EL_i - 1 \right) \quad (5.26)$$

and

$$ED_i^{(\text{Poisson})} \approx \frac{1}{2} \frac{\eta_i t_{\text{new}}}{1-\rho} \quad (5.27)$$

where

$$\eta_i = (1+\rho_i) / (1 + \sum_{i=1}^N \rho_i^2/\rho),$$

$$t_{\text{new}} = t + EW(1 + \sum_{i=1}^N \rho_i^2/\rho).$$

The first term on the right hand side of eqn. (5.26) is the queueing delay of an M/G/1 single-queue system with appropriate parameters. The second term is due to the non-Poisson nature of the arrivals, which can be positive or negative. The third term arises because all message arrivals are registered at the end of a time slot. For the time continuous system with Poisson arrivals, the last two terms vanish, as noted earlier.

For very heavy system traffic ($\rho \rightarrow 1$), \tilde{v}_{ii} is approximately independent of i , as remarked earlier. Thus, (5.26) and (5.27) are valid for this case. Moreover, the hyperbolic term dominates the right hand side of (5.26).

Therefore

$$ED_i(\text{slot}) \approx \frac{1}{2} \frac{\eta_i t_{\text{new}}}{1-\rho} \quad (5.28)$$

and

$$ED_i(\text{Poisson}) \approx \frac{1}{2} \frac{\eta_i t_{\text{new}}}{1-\rho} \quad (5.29)$$

Eqns. (5.28) and (5.29) show that the average delays for both slot and Poisson arrivals are approximately proportional to $(1+\rho_i)$ as $\rho \rightarrow 1$. This in turn implies that if a system has a user with utilization close to unity, the delay suffered by this user is roughly twice the delay suffered by other users.

For very low system traffic ($\rho \rightarrow 0$), (5.21) implies \tilde{v}_{ii} is

approximately independent of i . Thus, (5.26) and (5.27) are again valid in this case. Furthermore,

$$\frac{1}{2} \frac{\eta_i t_{\text{new}}}{1-\rho} \approx \frac{1}{2} (t + EW) \quad (5.30)$$

By making the reasonable assumptions that $\text{var } M_i \rightarrow 0$ as $EM_i \rightarrow 0$ and $\text{var } L_i \rightarrow 0$ as $EL_i \rightarrow 0$, we can further approximate t for $\rho \rightarrow 0$ by

$$t \approx \frac{1}{EC} \sum_{i=1}^N \text{var } W_i \approx \frac{\text{var } W}{EW} \quad (5.31)$$

where

$$EW = \sum_{i=1}^N EW_i,$$

$$\text{var } W = \sum_{i=1}^N \text{var } W_i.$$

Therefore,

$$ED_i(\text{slot}) \approx \frac{1}{2} \left(\frac{\text{var } W}{EW} + EW \right) + \frac{1}{2} \left(\frac{\text{var } M_i - EM_i}{EM_i} EL_i - 1 \right) \quad (5.32)$$

and

$$ED_i(\text{Poisson}) \approx \frac{1}{2} \left(\frac{\text{var } W}{EW} + EW \right) \quad (5.33)$$

If $\text{var } W = 0$, then

$$ED_i(\text{Poisson}) \approx \frac{1}{2} EW \quad (5.34)$$

Eqn. (5.34) is intuitively clear but (5.33) shows a more subtle result if $\text{var } W \neq 0$. For the time slot system, more information is needed on the arrival statistics in order to simplify (5.32) any further.

For the system with a single dominant user, that is, ρ_i and t_i for the dominant user are much greater than ρ_i 's and t_i 's for the other users, eqn.

(5.21) again implies that \tilde{v}_{ii} are approximately independent of i . Thus, (5.26) and (5.27) are valid in this case.

Furthermore,

$$\frac{1}{2} \frac{\eta_i t_{\text{new}}}{1-\rho} \approx \frac{1}{2} \frac{t + EW(1+\rho)}{1-\rho} \quad (5.35)$$

for the dominant user and

$$\frac{1}{2} \frac{\eta_i t_{\text{new}}}{1-\rho} \approx \frac{1}{2} \frac{1+\rho_i}{1+\rho} \frac{t + EW(1+\rho)}{1-\rho} \quad (5.36)$$

for all other users.

For the system with $N-1$ identical users and a single small user, that is, ρ_i and t_i for the small user are much less than ρ_i 's and t_i 's for the other users, eqn. (5.21) implies once more that \tilde{v}_{ii} is approximately independent of i . As a result, (5.26) and (5.27) are valid. Furthermore

$$\frac{1}{2} \frac{\eta_i t_{\text{new}}}{1-\rho} \approx \frac{1}{2} \frac{1}{1+\rho/(N-1)} \frac{t+EW(1+\rho/(N-1))}{1-\rho} \quad (5.37)$$

for the small user and

$$\frac{1}{2} \frac{\eta_i t_{\text{new}}}{1-\rho} \approx \frac{1}{2} \frac{t+EW(1+\rho/(N-1))}{1-\rho} \quad (5.38)$$

for the other identical users.

For $t_i \propto \rho_i$, eqn. (5.21) again implies that \tilde{v}_{ii} is approximately independent of i so eqn. (5.26) and (5.27) are again valid.

For the system with clusters of identical users, eqn. (5.21) shows that \tilde{v}_{ii} 's obey a linear relationship within each cluster. This linear relationship in turn shows that the delays given by (5.1) and (5.2) are linear within each cluster of identical users.

In most of the special cases considered so far, \tilde{v}_{ii} is approximately independent of i . This is verified for the following combinations of parameters:

(i) $N=8$.

(ii) $\tau_1 = \rho_1^2, \rho_1, 1$.

(iii) $0.01 < \rho < 0.99$ with steps of 0.01.

(iv) 1 large user with utilization 100 times that of other identical users,
1 small user with utilization 0.01 times that of other identical users.

2 clusters of identical users with utilization ratio of 1:5.

There are two plots for all combinations of (iii) and (iv). The first plot consists of the exact numerical solution for $\{\tilde{v}_{ii}\}$ versus ρ (8 curves) and the second plot consists of the normalized variance of $\{\tilde{v}_{ii}\}$ versus ρ . The normalized variance of $\{\tilde{v}_{ii}\}$ is defined as

$$\text{var}(\tilde{v}_{ii}) = \frac{1}{Nm^2} \sum_{i=1}^N (\tilde{v}_{ii} - m)^2 \quad (5.39)$$

where

$$m = \frac{1}{N} \sum_{i=1}^N \tilde{v}_{ii}.$$

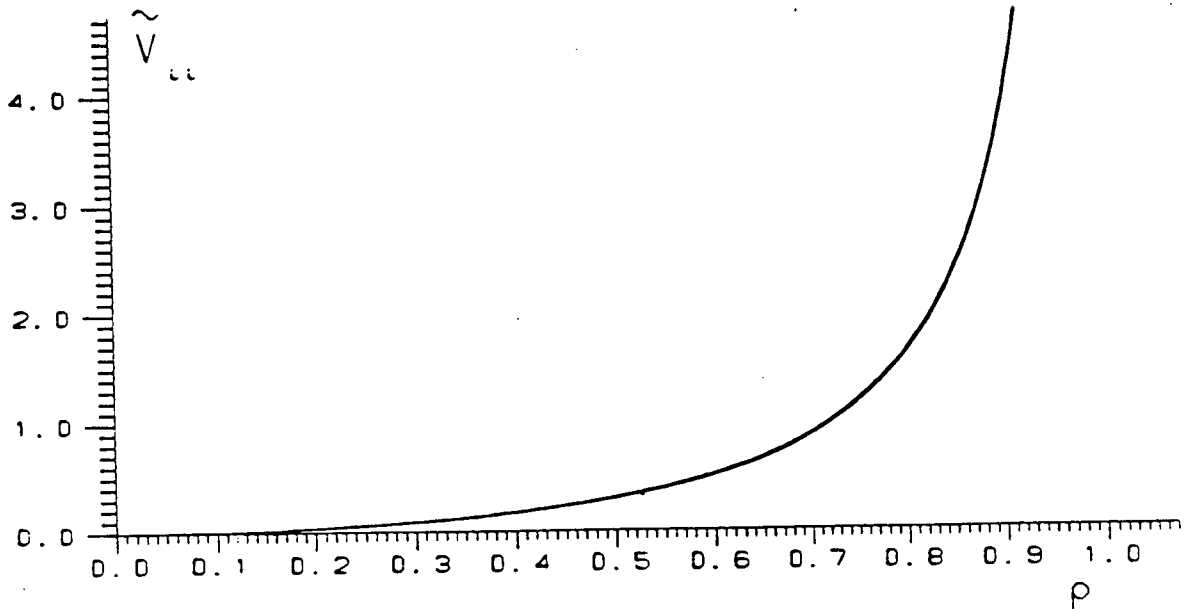


Fig. 5.7a \tilde{V}_{ii} versus ρ for $N=8$, 1 single large user and $t_i = \rho_i^2$.

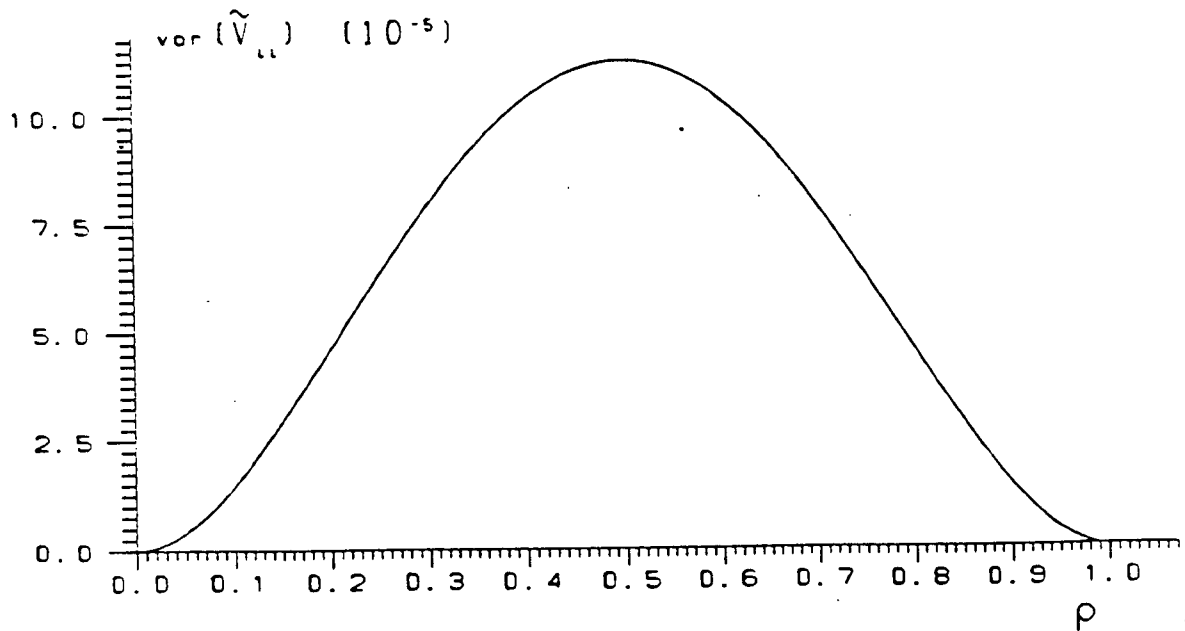


Fig. 5.7b $\text{var}(\tilde{V}_{ii})$ versus ρ for $N=8$, 1 single large user and $t_i = \rho_i^2$.

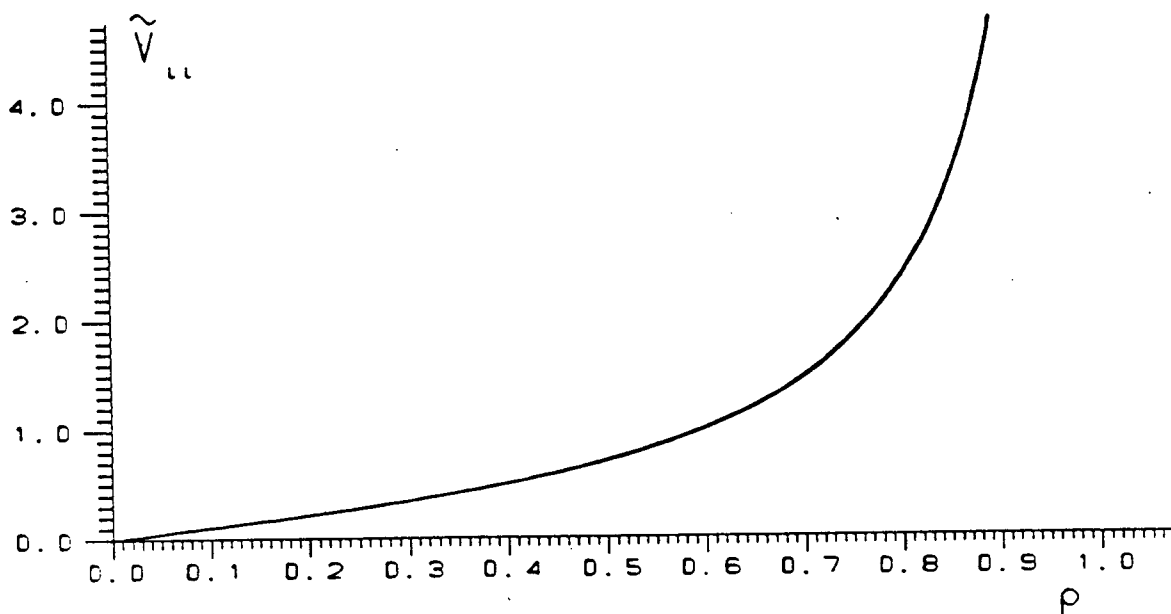


Fig. 5.8a \tilde{V}_{ii} versus ρ for $N=8$, 1 single large user and $t_i = \rho_i$.

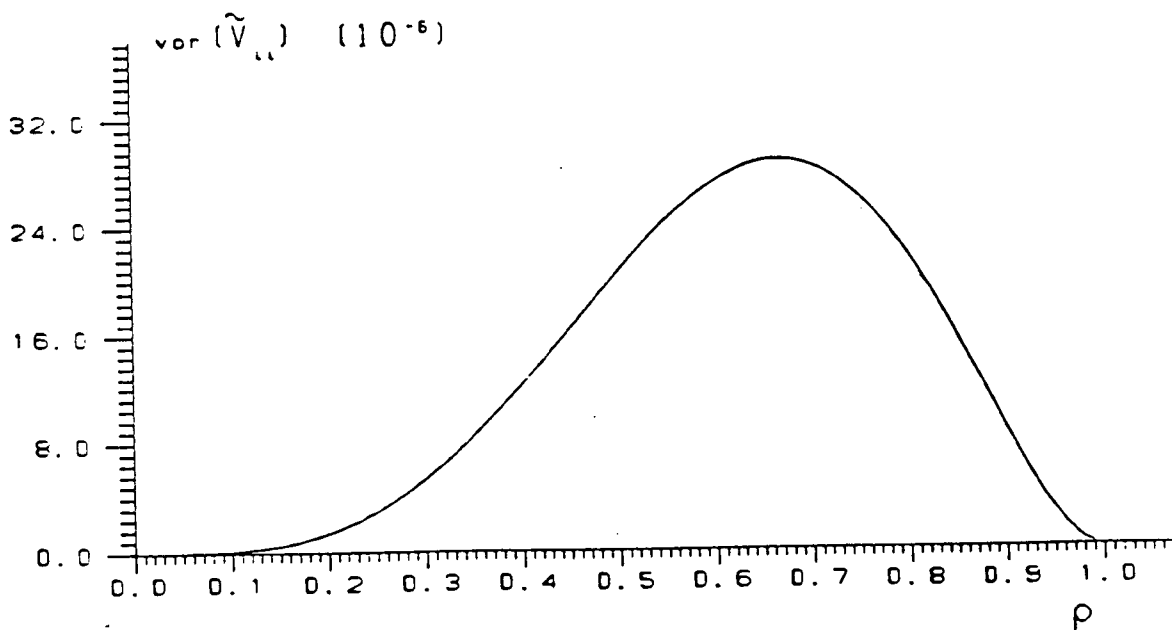


Fig. 5.8b $\text{var}(\tilde{V}_{ii})$ versus ρ for $N=8$, 1 single large user and $t_i = \rho_i$.

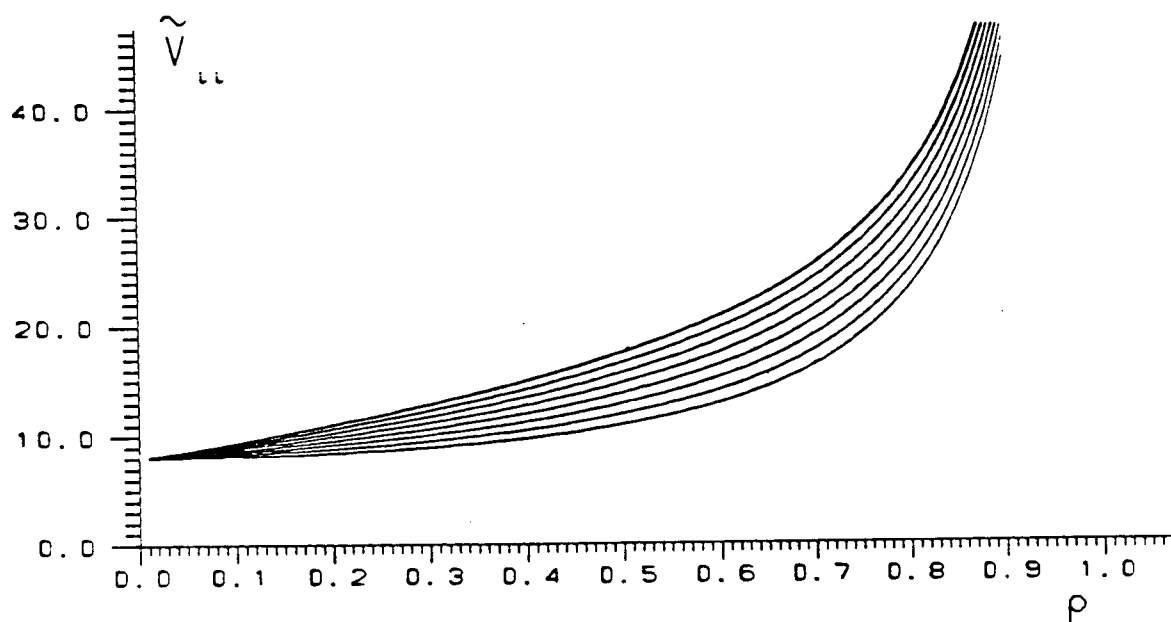


Fig. 5.9a \tilde{V}_{ii} versus ρ for $N=8$, 1 single large user and $t_1=1$.

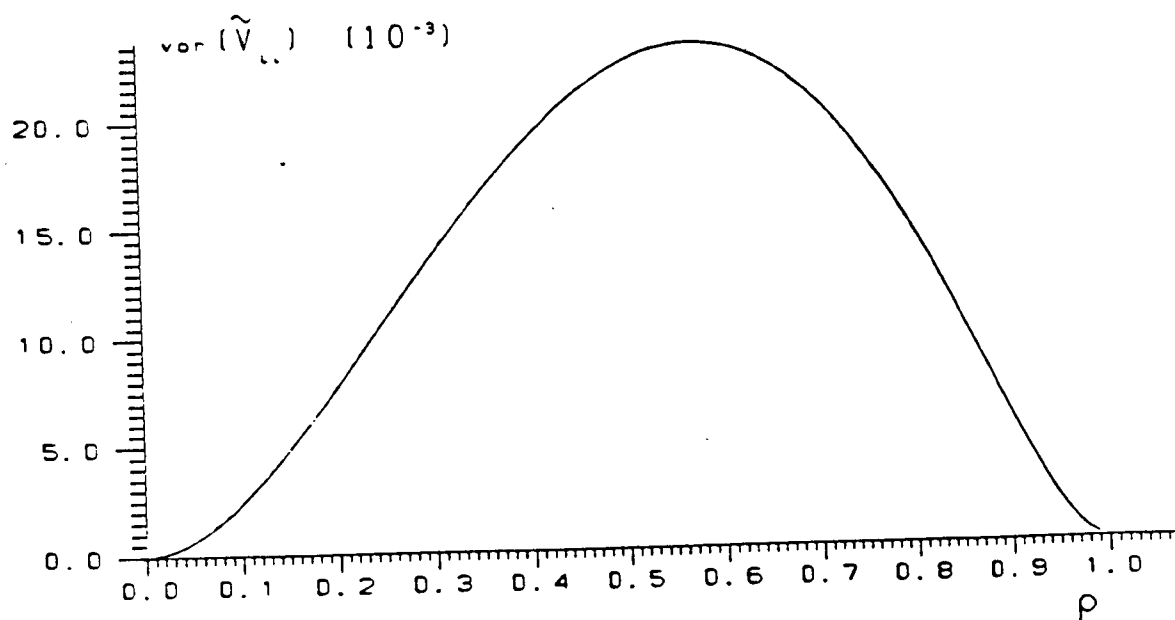


Fig. 5.9b $\text{var}(\tilde{V}_{ii})$ versus ρ for $N=8$, 1 single large user and $t_1=1$.

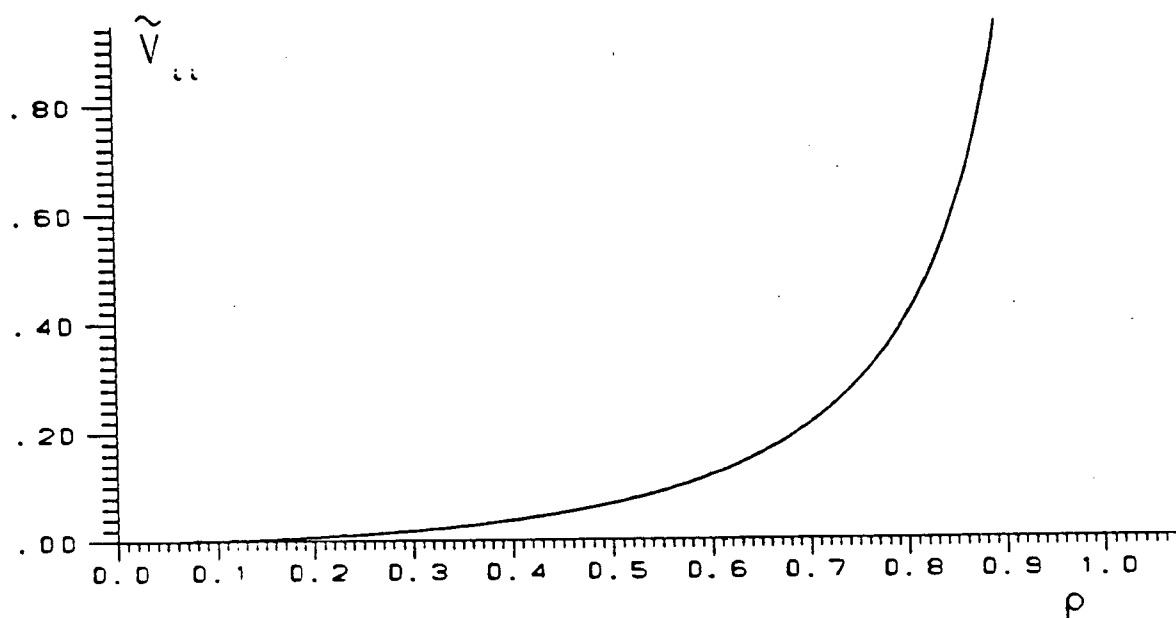


Fig. 5.10a \tilde{V}_{ii} versus ρ for $N=8$, 1 single small user and $t_i = \rho_i^2$.

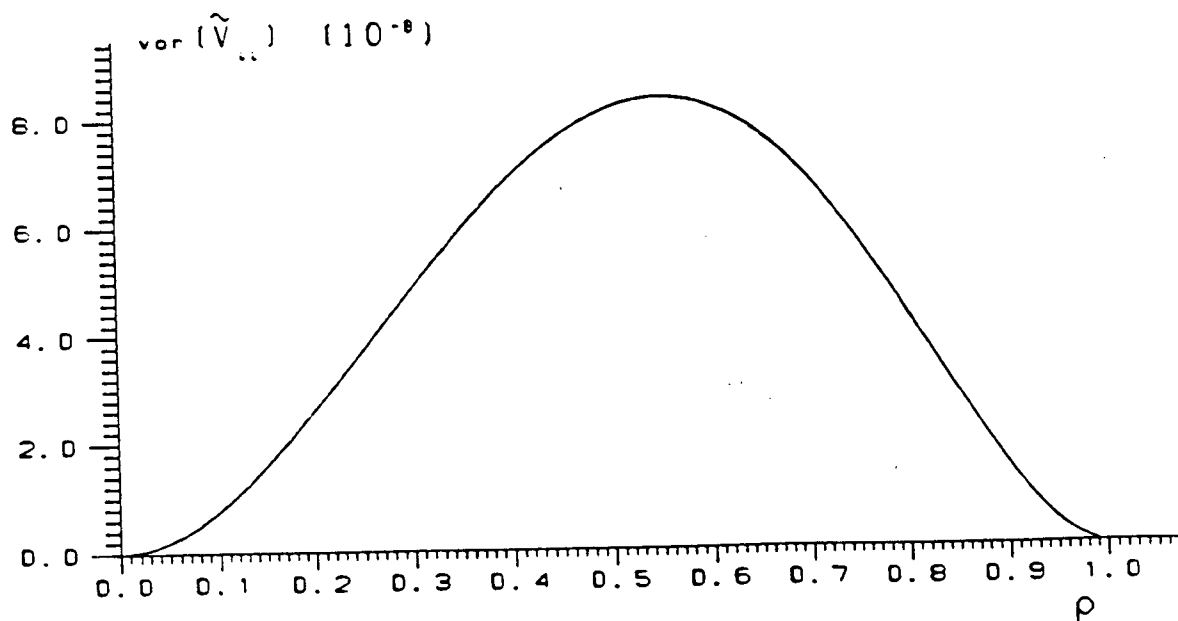


Fig. 5.10b $\text{var}(\tilde{V}_{ii})$ versus ρ for $N=8$, 1 single small user and $t_i = \rho_i^2$.

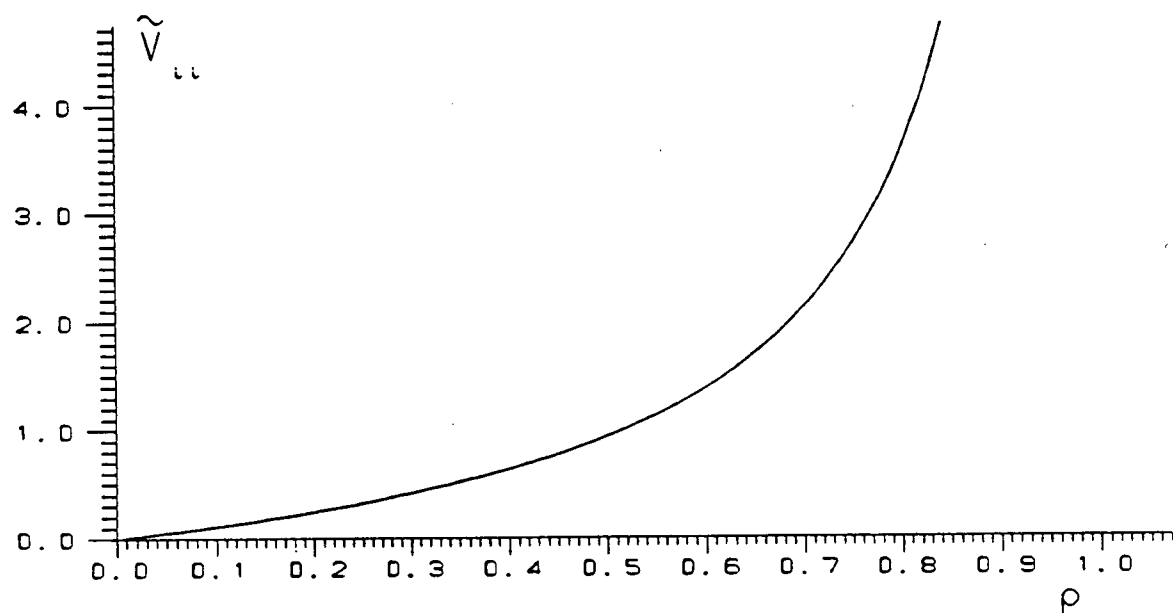


Fig. 5.11a \tilde{V}_{11} versus ρ for $N=8$, 1 single small user and $t_1=\rho_1$.

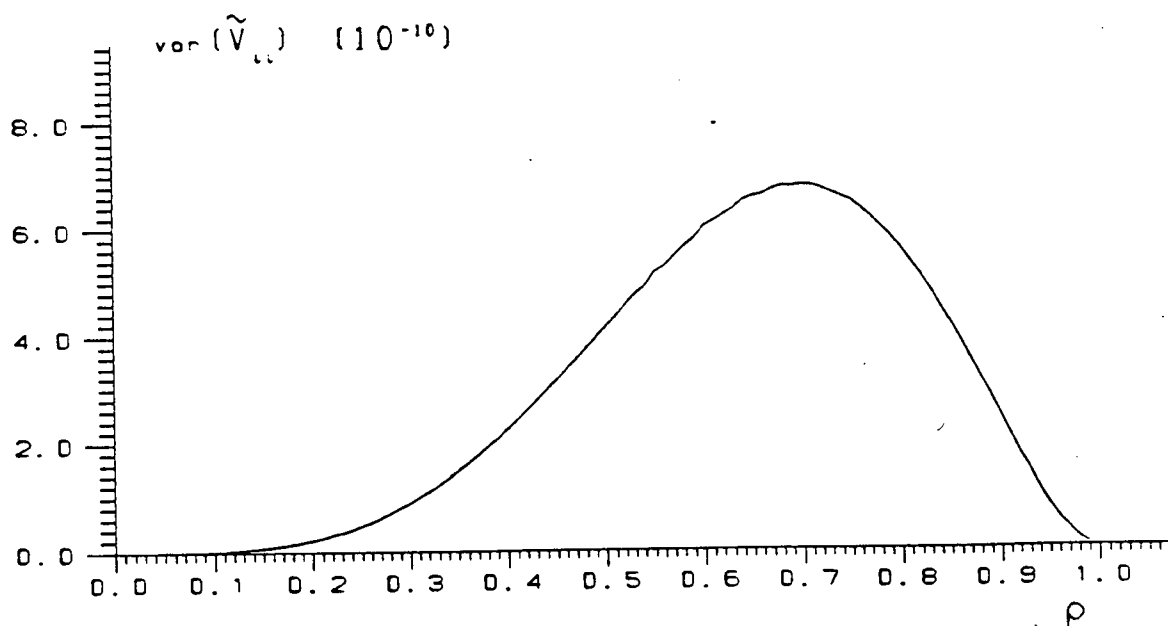


Fig. 5.11b $\text{var}(\tilde{V}_{11})$ versus ρ for $N=8$, 1 single small user and $t_1=\rho_1$.

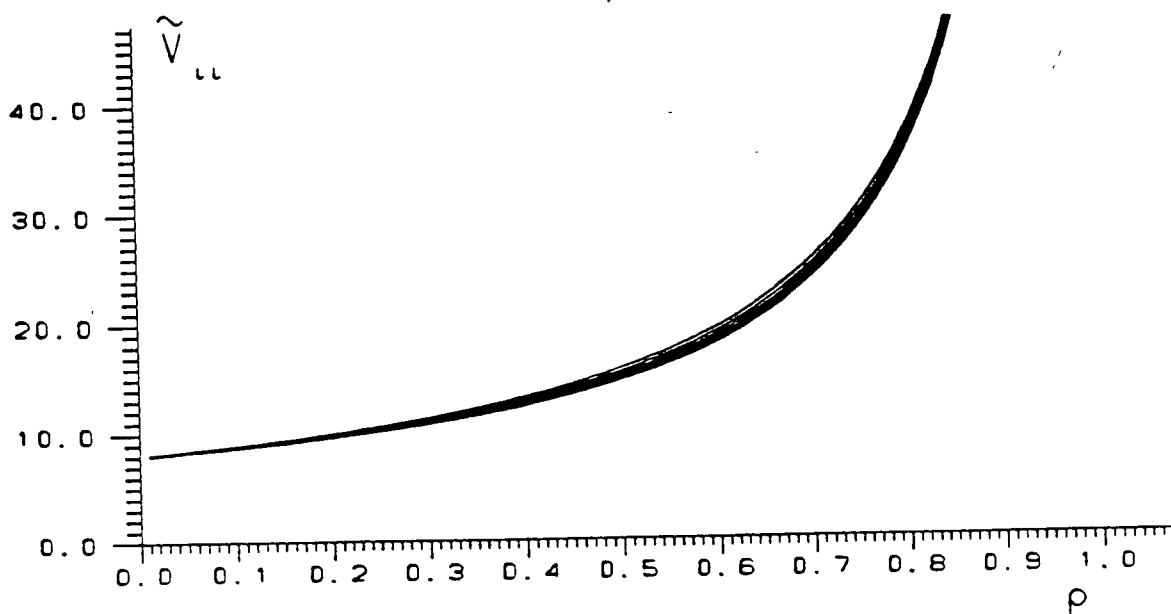


Fig. 5.12a \tilde{V}_{ii} versus ρ for $N=8$, 1 single small user and $t_i=1$.

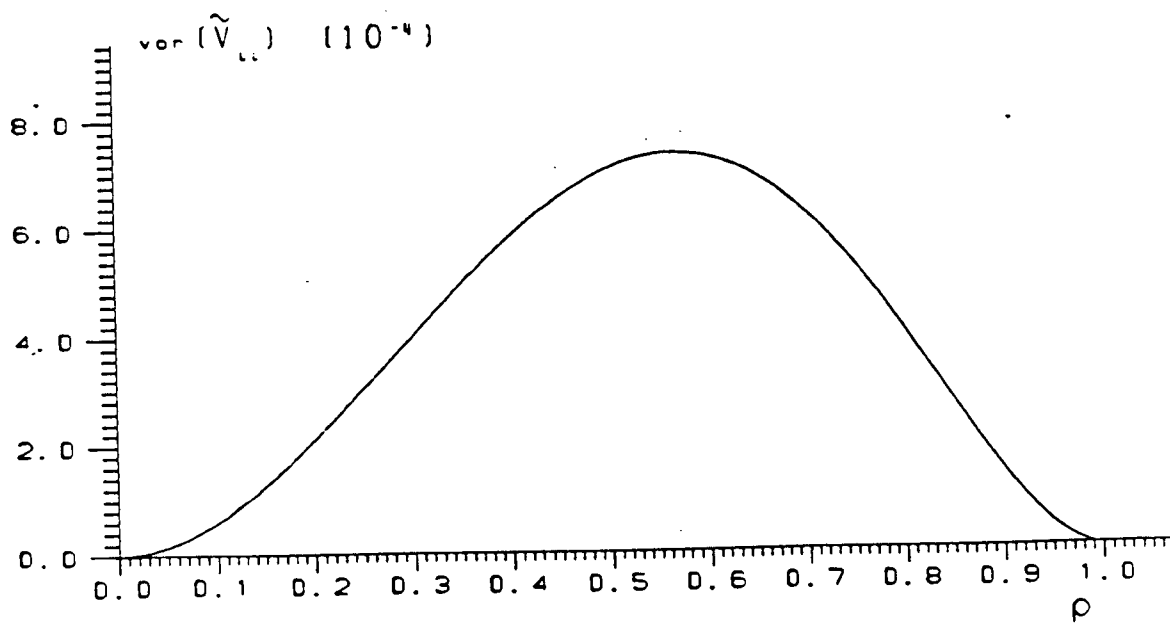


Fig. 5.12b $\text{var}(\tilde{V}_{ii})$ versus ρ for $N=8$, 1 single small user and $t_i=1$.

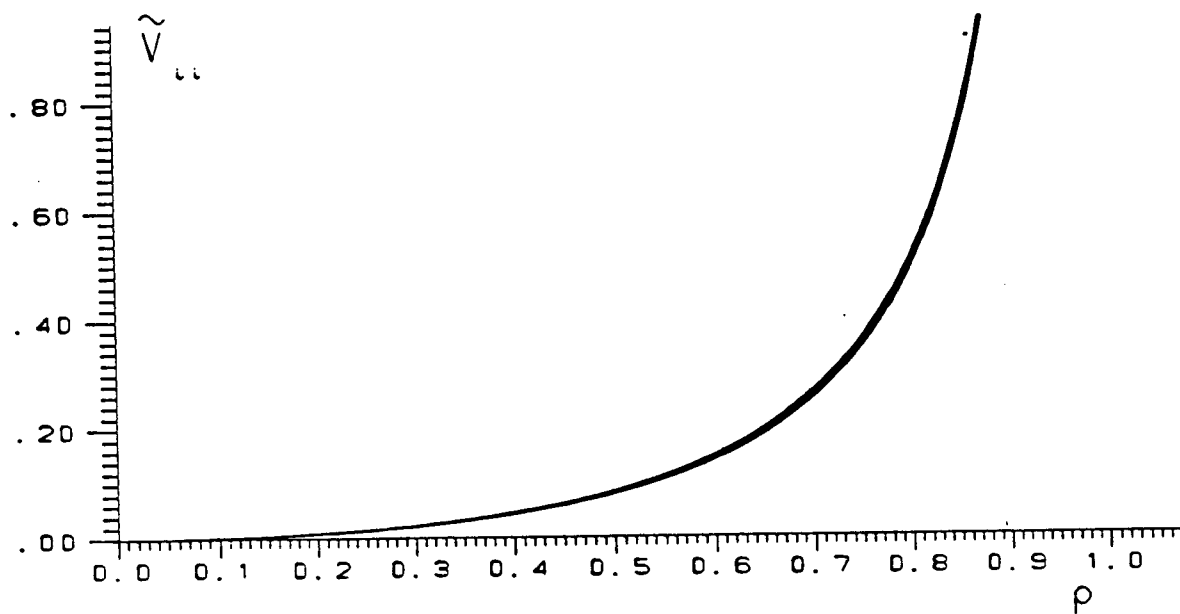


Fig. 5.13a \tilde{V}_{ii} versus ρ for $N=8$, 2 clusters of identical users and $t_i = \rho_i^2$.

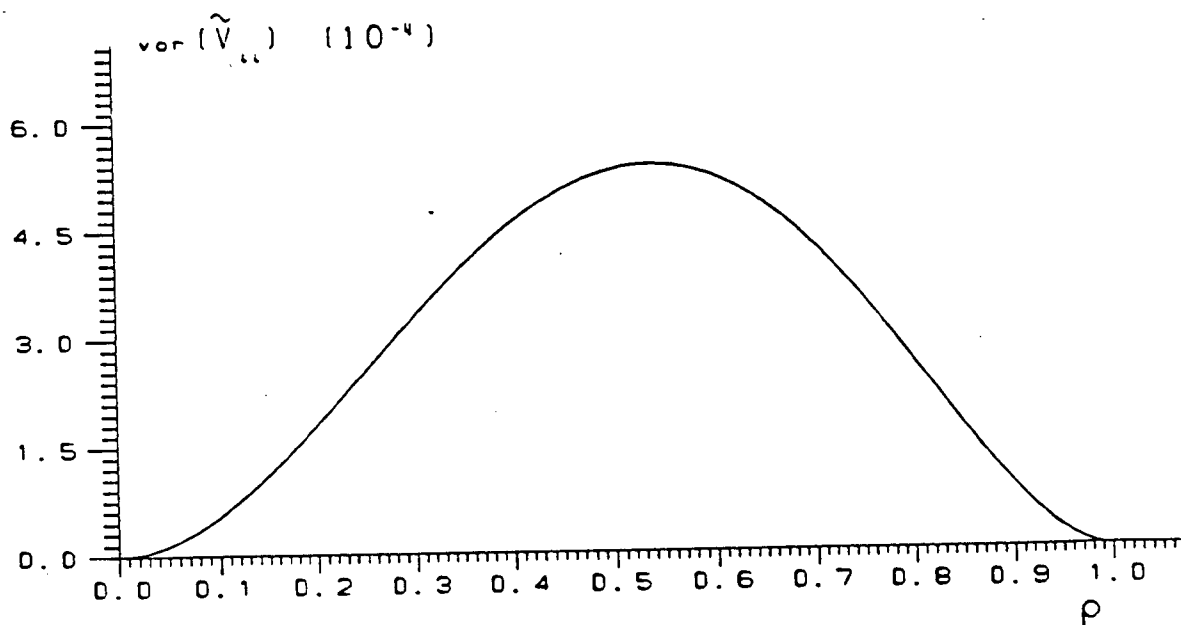


Fig. 5.13b $\text{var}(\tilde{V}_{ii})$ versus ρ for $N=8$, 2 clusters of identical users and $t_i = \rho_i^2$.

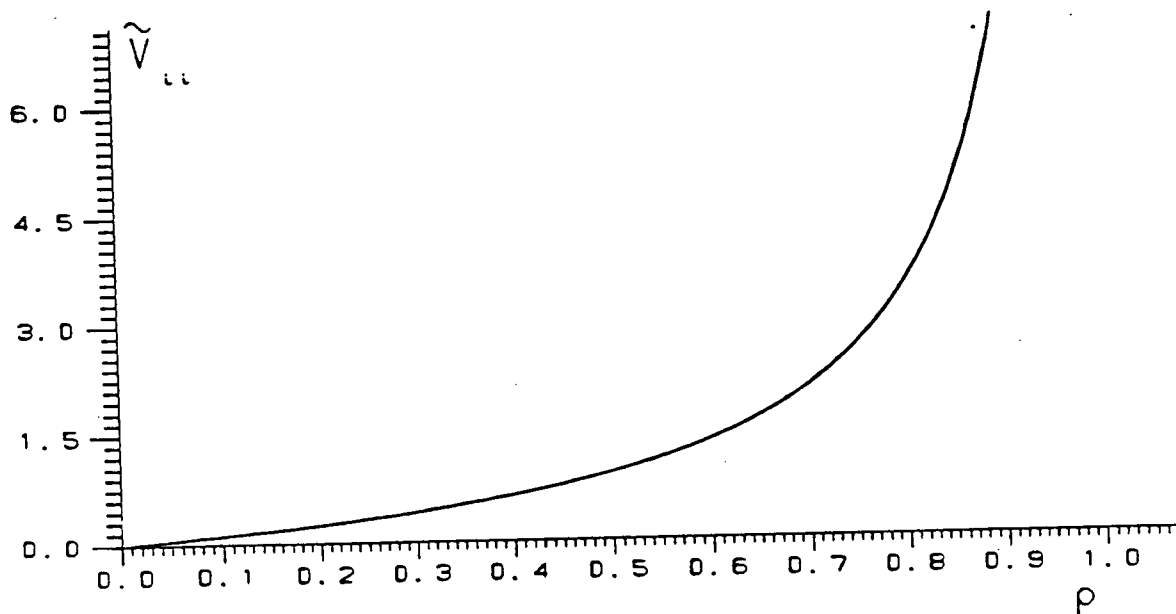


Fig. 5.14a \tilde{V}_{ii} versus ρ for $N=8$, 2 clusters of identical users and $t_i = \rho_i$.

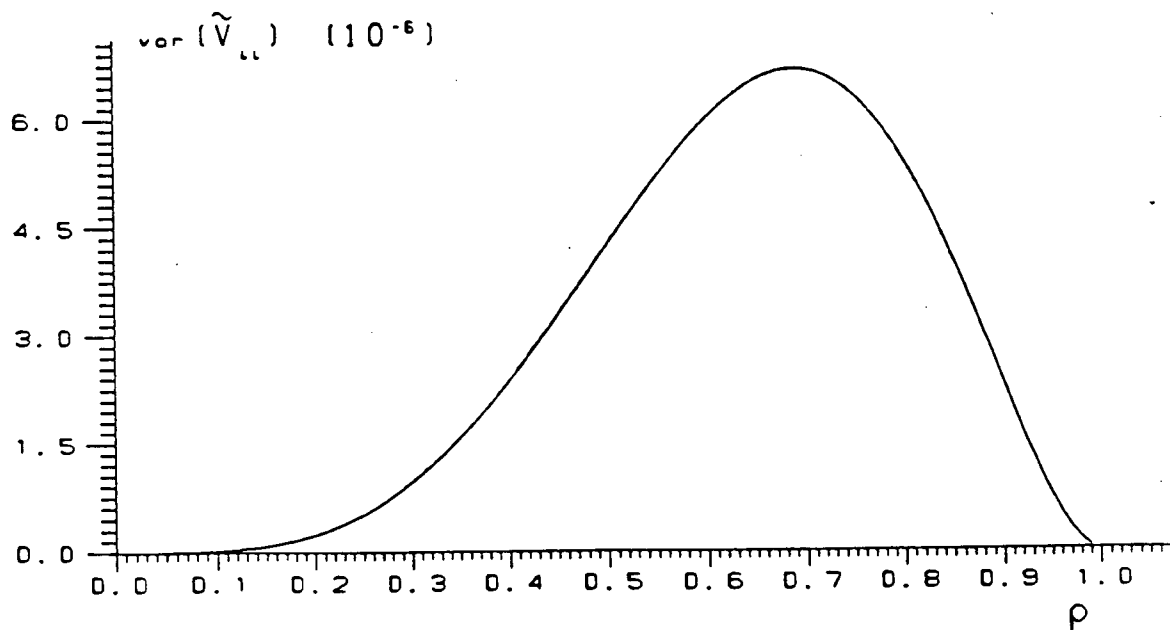


Fig. 5.14b $\text{var}(\tilde{V}_{ii})$ versus ρ for $N=8$, 2 clusters of identical users and $t_i = \rho_i$.

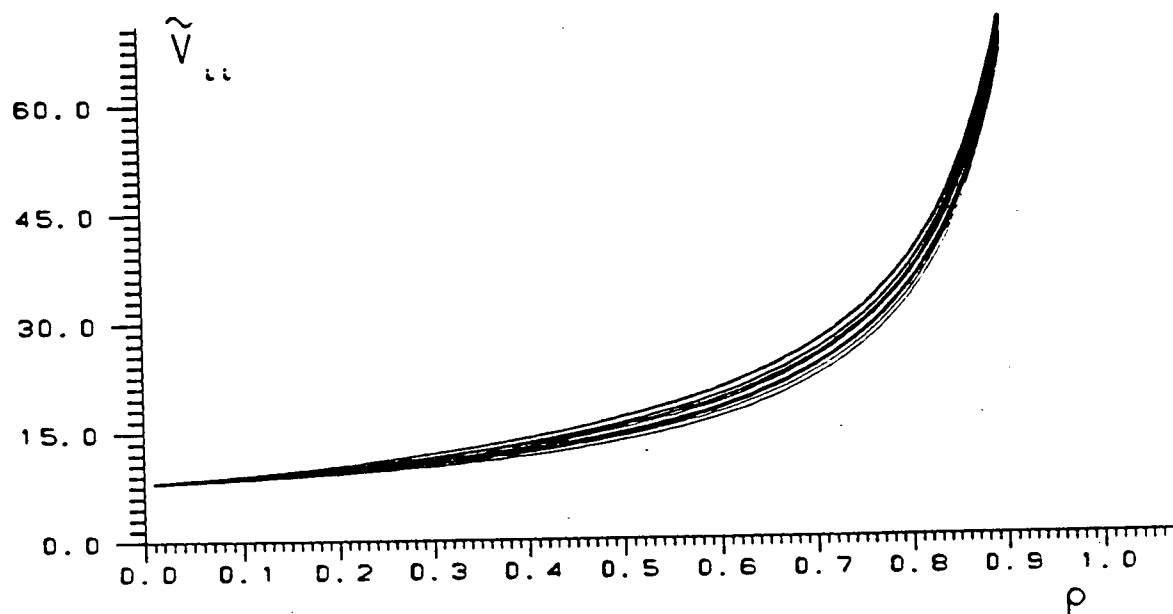


Fig. 5.15a \tilde{V}_{ii} versus ρ for $N=8$, 2 clusters of identical users and $t_1=1$.

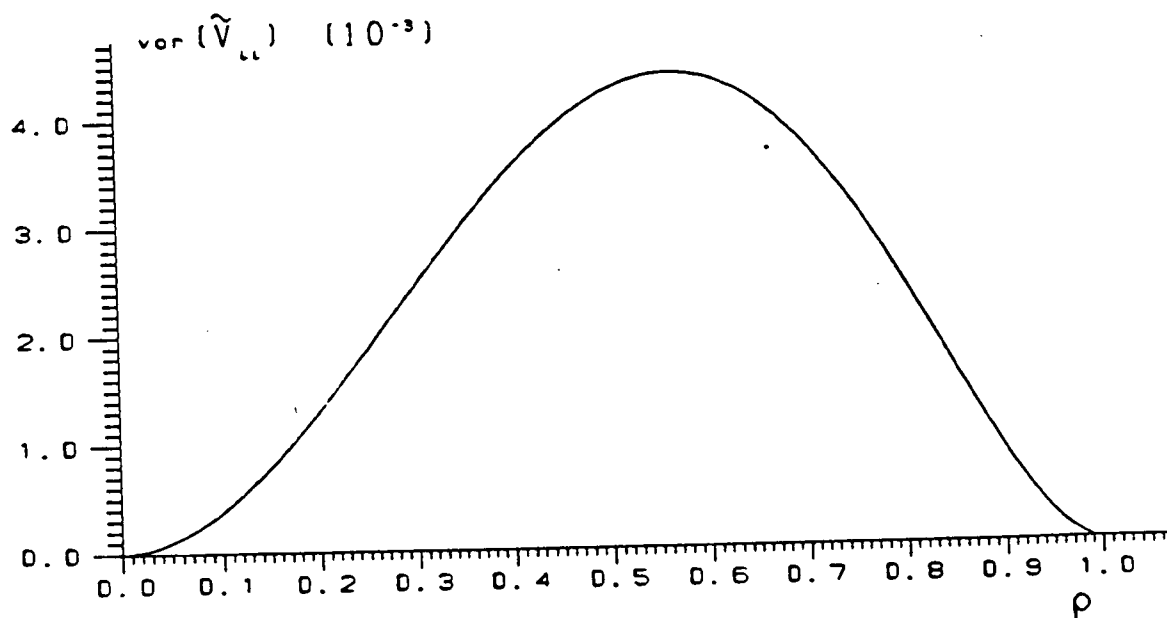


Fig. 5.15b $\text{var}(\tilde{V}_{ii})$ versus ρ for $N=8$, 2 clusters of identical users and $t_1=1$.

5.7 Finite Buffering and Non-exhaustive Services.

We have assumed infinite buffering and gated exhaustive service throughout this thesis. Nevertheless, we can use our results to approximate the more realistic cases of finite buffering and non-exhaustive services.

We start by considering the finite buffering case. Two kinds of buffer designs are possible, message buffering and slot buffering. As their names imply, a message is the fundamental unit in message buffering and a slot of a message is the basic unit in slot buffering. While slot buffering appears to be more efficient in terms of memory requirements, message buffering allows for hierarchical buffer management. We shall consider message buffering here. Let us modify the infinite buffering system we have considered to a finite buffering system with n_i message buffers at queue i . As long as the blocking probability is small, we can approximate the finite buffering system by the infinite buffering system. We further approximate the blocking probability at queue i by

$$\mu_i^{(\text{block})} \approx \lim_{k \rightarrow \infty} \text{Prob}(X_i^{(k)} > n_i) \quad (5.40)$$

This approximation is based on the assumption that the queue length is a local maximum at the scan instants. From (2.1) - (2.5) and previous analysis, we have

$$EX_i = EM_i EC \quad (5.41)$$

and

$$\begin{aligned}
\text{var } X_i &= \text{var } Y_i \\
&= \text{var } M_i \text{ EC} + (EM_i)^2 \text{ EC}_i^2 \\
&= \text{var } M_i \text{ EC} + \text{var } C_i (EM_i)^2 + (EM_i)^2 (EC)^2 \\
&= [\text{var } M_i + (EM_i)^2 (\tilde{v}_{ii} + EC)] \text{ EC}
\end{aligned} \tag{5.42}$$

where all variables are defined earlier. The blocking probability given by (5.40) can be bounded using (5.41), (5.42) and Chebyshev's inequality which states that

$$\text{Prob}(|X - m_X| > \delta) < \frac{\sigma_X^2}{\delta^2} \tag{5.43}$$

where

X is any random variable,

m_X is the mean of X and

σ_X^2 is the variance of X .

By choosing n_i large enough, we can limit $\mu_i^{(\text{block})}$ to an acceptable level. Notice that $\{n_i\}$ depends on $\{\rho_i\}$ and $\{t_i\}$. Thus, we cannot choose a fixed set $\{n_i\}$ to keep $\{\mu_i^{(\text{block})}\}$ small for all loading conditions. In fact, n_i grows unbounded as $\rho \rightarrow 1$.

The problem of approximating non-exhaustive services by the gated exhaustive service scheme is similar to the approximation for finite buffering. Let us consider the general gated non-exhaustive service where up to p_i messages are served in each cyclic service of queue i . Let v_i be the probability that the scan queue length (number of messages buffered at the scan instant) at queue i is greater than p_i . As long as v_i is small for all i , we can approximate the gated non-exhaustive service reasonably well by the gated exhaustive service. We further assume that v_i can be approximated by

$$v_i \approx \lim_{k \rightarrow \infty} \text{Prob} (X_i^{(k)} > p_i) \quad (5.44)$$

Again, v_i can be bounded using (5.41), (5.42) and (5.43). By choosing the appropriate $\{p_i\}$ and $\{t_i\}$, we can limit v_i within the desired level. Thus, we can obtain a region of loading conditions under which the gated non-exhaustive service can be approximated by the gated exhaustive service. In general, for a sufficiently small ρ , the gated non-exhaustive service can always be approximated by the gated exhaustive service. For heavy traffic, one can use various heavy traffic approximations; see for example [15]. For intermediate traffic, interpolation techniques could be used.

6. CONCLUDING REMARKS.

6.1 Summary of the Results.

Recently, applications of fiber optics to LANs and the development of UBSs have intensified interest in the performance analysis of token-ring systems. As a first step, the average message delays in a general asymmetric token-ring under gated exhaustive service discipline have been investigated in this thesis. Although the gated exhaustive service discipline is a special case, it provides insights and low-traffic approximations for most non-exhaustive service schemes.

The approach used in this thesis is to formulate and solve the imbedded Markov chain of the joint terminal service times. Recursive linear equations are derived for the first and second joint moments of the imbedded chain. The existence and uniqueness of solutions to these equations are also established. A number of results concerning the solutions are also derived. Among these, the simple difference equation relating the second joint moments of the terminal service times is a significant contribution of this thesis. Such simple results have never been reported before. Based on equation (4.25), a very good approximation given by (5.21) and (5.23) is derived. This approximation is exact for the symmetric case and it works very well over a wide range of system parameters for the asymmetric case. Even under very unusual circumstances, the relative errors in the approximation are still acceptable ($\sim 10\%$). Some delay-vs-throughput results based on the approximation are also discussed.

6.2 Suggestions for Further Work.

The performance analysis of a general token-ring is a very difficult problem. In this thesis, a special case has been investigated but the ultimate goal is to solve the general problem. Thus, future work in this area should concentrate on the generalizations of the work in this thesis.

Emphasis has been placed on the average message delay as a performance measure. This can be generalized to include the delay variance. The delay variance is of considerable interest in many applications such as packetized voice and interactive terminal conversations. In our model, the delay variances depend on the third joint moments of the terminal service times. The analysis will involve a third differentiation of the basic functional equation (3.9) and the solutions of (3.12) and (3.13) to set up the necessary equations for the third joint moments. The problem is to examine the solution of those moments. Generalizations to even higher moments are possible but they tend to be difficult and less useful for practical purposes.

Although the arrival model used in this thesis appears to be very general, it is still inadequate for many practical situations. For example, our arrival model does not accommodate deterministic arrivals in voice applications and dependent arrivals in functional distributed computing. Thus, generalizations on the arrival model is desirable.

Perhaps the most interesting generalizations are associated with the service discipline. First of all, we have assumed the cyclic service schedule in this thesis. In other words, each queue has a cyclic priority which cannot be interrupted by other queues. In a general polling system,

many possible priority assignments are possible. Even in a token-ring, the basic protocol can be modified to accommodate various priority schedulings. The assignment of priorities, static or dynamic, is a design parameter which deserves further attention.

Secondly, generalization of the gated exhaustive service discipline to the more general non-exhaustive schemes would be very useful. This is because non-exhaustive service schemes can be implemented very easily on a token-ring. Unfortunately, the rigorous analysis of non-exhaustive service schemes appears to be very difficult. Very often, one should contend with approximate approaches. One approximate approach is to note that most non-exhaustive schemes can be approximated by the gated exhaustive service discipline at low traffic levels; for heavy traffic, one can use various heavy-traffic approximations, see for example [15]. Thus, the difficult case occurs when traffic is between the two extremes.

One further generalization can be made on the service discipline. We notice that in all 'static' service disciplines, $Z_i^{(k)}$ can be derived directly from $X_i^{(k)}$ (e.g. the relation given by eqn. (2.1)). One can generalize this dependence. For example, in the D-net protocol, each station can detect the length of a train fairly easily and estimate the current network loading, and then the station can determine $Z_i^{(k)}$ not only from $X_i^{(k)}$ but also from previous network loading estimates. Such elaborate service disciplines may not be justified for many LAN applications but we cannot be sure until we have done the analysis.

There are some relatively less urgent matters, for example, finite buffering, dependent switching overheads, and time-variant system parameters whose effects could also be examined.

References

- [1] A.S. Tanenbaum, Computer Networks, Prentice-Hall, Inc., 1981.
- [2] W. Stallings, Local Networks: An Introduction, New York: MacMillan, 1984.
- [3] D.D. Clark, K.T. Pogran, and D.P. Reed, "An Introduction to Local Area Networks," Proc. IEEE, vol. 66, no. 11, pp. 1497-1517, Nov. 1978.
- [4] R.M. Metcalfe and D.R. Boggs, "Ethernet: Distributed Packet Switching for Local Computer Networks," Commun. Ass. Comput. Mach., vol. 19, pp. 395-404, 1976.
- [5] S.D. Personick, "Review of Fundamentals of Optical Fiber Systems," IEEE J. Select. Areas Commun., SAC-1, no. 3, pp. 373-380, Apr. 1983.
- [6] N.L. Rhodes, "Interaction of Network Design and Fiber Optic Component Design in Local Area Networks," IEEE J. Select. Areas Commun., SAC-1, no. 3, pp. 489-492, Apr. 1983.
- [7] M.R. Finley, Jr., "Optical Fibers in Local Area Networks," IEEE Communications Magazine, vol. 22, no. 8, pp. 22-35, Aug. 1984.
- [8] E.G. Rawson and R.M. Metcalfe, "Fibernet: Multimode Optical Fibers for Local Computer Networks," IEEE Trans. Commun., COM. 26, no. 7, pp. 983-990, July 1978.
- [9] R.V. Schmidt, E.G. Rawson, R.E. Norton, Jr., S.B. Jackson, and M.D. Bailey, "Fibernet II: A Fiber Optic Ethernet," IEEE J. Select. Areas Commun., SAC-1, no. 5, pp. 702-710, Nov. 1983.
- [10] E.S. Lee and P.I.P. Boulton, "The Principles and Performance of Hubnet: A 50Mbit/s Glass Fiber Local Area Network," IEEE J. Select. Areas Commun. SAC-1, no. 5, pp. 711-720, Nov. 1983.
- [11] F.A. Tobagi, F. Borgonovo, and L. Fratta, "Expressnet: A High-Performance Integrated-Services Local Area Network," IEEE J. Select. Areas Commun., SAC-1, no. 5, pp. 898-912, Nov. 1983.
- [12] C.-W. Tseng and B.-U. Chen, "D-Net, A New Scheme for High Data Rate Optical Local Area Networks," IEEE J. Select. Areas Commun., SAC-1, no. 53, pp. 493-499, Apr. 1983.
- [13] J.O. Limb and C. Flores, "Description of Fasnet: A Unidirectional Local Area Communications Network," Bell Syst. Tech. J. vol. 61, no. 7, Part I, pp. 1413-1440, Sep. 1982.
- [14] L. Kleinrock, Queueing Systems, Vol. 1: Theory, New York: Wiley, 1975.

- [15] L. Kleinrock, Queueing Systems, Vol. 2: Computer Applications, New York: Wiley, 1976.
- [16] H. Kobayashi and A.G. Konheim, "Queueing Models for Computer Communications System Analysis," IEEE Trans. Commun., COM-25, no. 1, pp. 2-28, Jan. 1977.
- [17] W. Bux, "Local-Area Subnetworks: A Performance Comparison," IEEE Trans. Commun., COM-29, no. 10, pp. 1465-1473, Oct. 1981.
- [18] W. Stallings, "Local Network Performance," IEEE Communications Magazine, vol. 22, no. 2, pp. 27-36, Feb. 1984.
- [19] L. Takacs, "Two Queues Attended by a Single Server," Oper. Res., vol. 16, no. 3, pp. 639-650, 1968.
- [20] B. Avi-Itzhak, W.L. Maxwell, and L.W. Miller, "Queueing with Alternating Priorities," Oper. Res., vol. 13, no. 2, pp. 306-318, 1965.
- [21] M. Eisenberg, "Two Queues with Changeover Times," Oper. Res., vol. 19, pp. 386-401, 1971.
- [22] R.B. Cooper and G. Murray, "Queues Served in Cyclic Order," Bell Syst. Tech. J., vol. 48, no. 3, pp. 675-689, Mar. 1969.
- [23] R.B. Cooper, "Queues Served in Cyclic Order: Waiting Times," Bell Syst. Tech. J., vol. 49, no. 3, pp. 399-413, Mar. 1970.
- [24] M. Eisenberg, "Queues with Periodic Service and Changeover Time," Oper. Res., vol. 20, pp. 440-451, 1972.
- [25] A.G. Konheim and B. Meister, "Waiting Lines and Times in a System with Polling," J. Ass. Comput. Mach., vol. 21, pp. 470-490, 1974.
- [26] G.B. Swartz, "Polling in a Loop System," J. Ass. Comput. Mach., vol. 27, pp. 42-59, 1980.
- [27] I. Rubin and L.F.M. DeMoraes, "Message Delay Analysis for Polling and Token Multiple-Access Schemes for Local Communication Networks," IEEE J. Selec. Areas Commun., SAC-1, no. 5, pp. 935-946, Nov. 1983.
- [28] Y. Aminetzah, "An Exact Approach to the Polling System," Ph.D. dissertation, McGill Univ., Montreal, P.Q., Canada, Mar. 1975.
- [29] M.J. Ferguson and Y.J. Aminetzah, "Exact Results for Nonsymmetric Token Ring Systems," IEEE Trans. Commun., COM-33, no. 3, pp. 223-231, Mar. 1985.

- [30] R.T. Carsten, E.E. Newhall, and M.J.M. Posner, "A Simplified Analysis of Scan Times in an Asymmetric Newhall Loop with Exhaustive Service," IEEE Trans. Commun., COM-25, no. 9, pp. 951-957, Sep. 1977.
- [31] W. Bux and H.L. Truong, "Token-ring Performance: Mean Delay Approximation," in Proc. 10th Int. Teletraffic Cong., June 1983, pp. 3.3.3.1-3.3.3-7.
- [32] M. Leibowitz, "An Approximate Method for Treating a Class of Multiqueue Problems," IBM J. Res. Develop., vol. 5, pp. 204-209, 1961.
- [33] S. Halfin, "An Approximate Method for Calculating Delays for a Family of Cyclic Type Queues," Bell Syst. Tech. J., vol. 54, no. 10, pp. 1733-1754, Dec. 1975.
- [34] P.J. Kuehn, "Multiqueue Systems with Nonexhaustive Cyclic Service," Bell Syst. Tech. J., vol. 58, no. 3, pp. 671-698, Mar. 1979.
- [35] S.S. Nair, "Alternating Priority Queues with Non-Zero Switch Rule," Comp. and Opns. Res., vol. 3, pp. 337-346, 1976.
- [36] C. Mack, T. Murphy, and N.L. Webb, "The Efficiency of N Machines Unidirectionally Patrolled by One Operative when Walking Time and Repair Times are Constants," J. Royal Stat. Soc., ser. B, no. 19, pp. 166-172, 1957.
- [37] A.R. Kaye, "Analysis of a distributed control loop for data transmission," in Proc. Symp. Comput. Commun. Networks Teletraffic, Polytech. Inst., Brooklyn, NY, Apr. 1972.
- [38] F.A. Tobagi and M. Fine, "Performance of Unidirectional Broadcast Local Area Networks: Expressnet and Fasnet," IEEE J. Select. Areas Commun., SAC-1, no. 5, pp. 913-925, Nov. 1983.

APPENDIX: Eigenvalues of the Jacobian Matrix A.

A number of results regarding the eigenvalues of the Jacobian Matrix A are proved in this Appendix.

Lemma A.1:

The characteristic equation $|A - rI| = 0$ of the matrix A is given by

$$Q(r) = \left\{ \prod_{i=1}^N [(1+\rho_i)r - \rho_i] - r^{N+1} \right\} / (1-r) = 0 \quad (A.1)$$

Proof of Lemma A.1:

From the structure of A given by Lemma 4.3, we claim that A can be written as follows:

$$A = T_N T_{N-1} \dots T_1 D R \quad (A.2)$$

where

$$T_1 = \left[\begin{array}{c|c|c} I_{i-1} & 0 & 0 \\ \hline \rho_1 \rho_i \dots \rho_i & 1 & 0 \\ \hline 0 & 0 & I_{N-i} \end{array} \right],$$

$I_k = k \times k$ identity matrix,

$$D = \left[\begin{array}{cccc} \rho_1 & & & \\ & \rho_2 & & 0 \\ & 0 & \ddots & \\ & & & \rho_N \end{array} \right],$$

$$R = \left[\begin{array}{cccc} 1 & 1 & \dots & 1 \\ & 1 & & \vdots \\ & 0 & \ddots & 1 \end{array} \right],$$

To see this, we successively compute $T_1 D$, $T_2 T_1 D$, ..., $T_N T_{N-1} \dots T_1 D$ and get

$$T_N T_{N-1} \dots T_1 D = [\rho_1(\vec{u}_1 + \vec{e}_1) | \rho_2(\vec{u}_2 + \vec{e}_2) | \dots | \rho_N(\vec{u}_N + \vec{e}_N)] \quad (A.3)$$

where \vec{u}_1 and \vec{e}_1 are defined in Lemma 4.3. Then we notice that

$$T_N T_{N-1} \dots T_1 D R = [\rho_1(\vec{u}_1 + \vec{e}_1) | \sum_{k=1}^2 \rho_k(\vec{u}_k + \vec{e}_k) | \dots | \sum_{k=1}^N \rho_k(\vec{u}_k + \vec{e}_k)] \quad (A.4)$$

$$\text{and } \vec{u}_{i-1} = \sum_{k=i}^N \rho_k(\vec{u}_k + \vec{e}_k) \quad (A.5)$$

Thus,

$$T_N T_{N-1} \dots T_1 D R = [\vec{u}_0 - \vec{u}_1 | \vec{u}_0 - \vec{u}_2 | \dots | \vec{u}_0 - \vec{u}_N] = A.$$

The determinant $|A - rI|$ can now be computed as follows:

$$\begin{aligned} |A - rI| &= |TDR - rI| \\ &= |T| |D - rT^{-1}R^{-1}| |R| \end{aligned} \quad (A.6)$$

where $T = T_N T_{N-1} \dots T_1$.

We notice that $|T| = |R| = 1$ so

$$\begin{aligned} |A - rI| &= |D - rT^{-1}R^{-1}| \\ &= |D - rT_1^{-1}T_2^{-1} \dots T_N^{-1}R^{-1}| \end{aligned} \quad (A.7)$$

The inverses of T_k 's and R can be expressed by

$$R^{-1} = \begin{bmatrix} 1 & & -1 & & 0 \\ & 1 & & -1 & \\ & & \ddots & \ddots & \\ 0 & & & -1 & \\ & & & & 1 \end{bmatrix}, \quad T_k^{-1} = \left[\begin{array}{c|c|c} I_{k-1} & 0 & 0 \\ \hline -\rho_k - \rho_k \dots - \rho_k & 1 & 0 \\ \hline 0 & 0 & I_{N-k} \end{array} \right]$$

Using these to calculate $T_N^{-1}R^{-1}$, $T_{N-1}^{-1}T_N^{-1}R^{-1}$, ..., $T_1^{-1}T_2^{-1} \dots T_N^{-1}R^{-1}$ successively, we have

$$T_1^{-1} T_2^{-1} \dots T_N^{-1} R^{-1} = \begin{bmatrix} -1 & -1 & & \\ -\rho_2 & 1+\rho_2 & -1 & 0 \\ -\rho_3 & & 1+\rho_3 & \ddots \\ \vdots & & 0 & \ddots & -1 \\ -\rho_N & & & & 1+\rho_N \end{bmatrix} \quad (A.8)$$

$$\begin{aligned} \therefore D &= r T^{-1} R^{-1} \\ &= \begin{bmatrix} \rho_1 r + \rho_1^{-(1+\rho_1)} r & r & & & \\ \rho_2 r & \rho_2^{-(1+\rho_2)} r & r & & 0 \\ \rho_3 r & & \rho_3^{-(1+\rho_3)} r & \ddots & \\ \vdots & 0 & & \ddots & \\ \rho_N r & & & & \rho_N^{-(1+\rho_N)} r \end{bmatrix} \end{aligned} \quad (A.9)$$

The determinant of the above matrix can be split into a sum of determinants of two matrices U_1 and U_2 where

$$U_1 = \begin{bmatrix} \rho_1 r & r & & & \\ \rho_2 r & \rho_2^{-(1+\rho_2)} r & r & & 0 \\ \vdots & & \rho_3^{-(1+\rho_3)} r & \ddots & \\ \vdots & & & \ddots & \\ \rho_N r & & 0 & & \rho_N^{-(1+\rho_N)} r \end{bmatrix}$$

and

$$U_2 = \begin{bmatrix} \rho_1^{-(1+\rho_1)} r & & & & \\ & r & & & \\ & \rho_2^{-(1+\rho_2)} r & & & 0 \\ & & \ddots & & \\ & & & \ddots & \\ 0 & & & & \rho_N^{-(1+\rho_N)} r \end{bmatrix}.$$

Clearly,

$$|U_2| = \prod_{i=1}^N [\rho_i - (1+\rho_i)r] \quad (\text{A.10})$$

Factorizing ρ_i from each row of U_1 and then $\frac{r}{1-r}$ from the first column, we have

$$|U_1| = \left(\prod_{i=1}^N \rho_i \right) \left(\frac{r}{1-r} \right) |U_3| \quad (\text{A.11})$$

where

$$U_3 = \begin{bmatrix} 1-r & \frac{r}{\rho_1} & & & 0 \\ 1-r & 1-(1+\frac{1}{\rho_2})r & \frac{r}{\rho_2} & & \\ \cdot & & 1-(1+\frac{1}{\rho_3})r & \cdot & \\ \cdot & & & \cdot & \\ \cdot & 0 & & & \frac{r}{\rho_{N-1}} \\ 1-r & & & & 1-(1+\frac{1}{\rho_N})r \end{bmatrix}$$

Replacement of the first column by subtracting from it the sum of all other columns yield

$$U_4 = \begin{bmatrix} 1-(1+\frac{1}{\rho_1})r & \frac{r}{\rho_1} & & & 0 \\ 0 & 1-(1+\frac{1}{\rho_2})r & \frac{r}{\rho_2} & & \\ 0 & & 1-(1+\frac{1}{\rho_3})r & \cdot & \\ \cdot & & & \cdot & \\ \cdot & 0 & & & \frac{r}{\rho_{N-1}} \\ \frac{r}{\rho_N} & & & & 1-(1+\frac{1}{\rho_N})r \end{bmatrix} \quad (\text{A.12})$$

where $|U_3| = |U_4|$.

Clearly,

$$|U_4| = \prod_{i=1}^N \left[1 - \left(1 + \frac{1}{\rho_i} \right) r \right] + (-1)^{N+1} \prod_{i=1}^N \left(\frac{r}{\rho_i} \right) \quad (\text{A.13})$$

Using (A.10)-(A.13) and after some algebra, we have

$$\begin{aligned} |A-rI| &= |U_1| + |U_2| \\ &= \frac{1}{1-r} (-1)^N \left\{ \prod_{i=1}^N \left[(1+\rho_i)r - \rho_i \right] - r^{N+1} \right\} \end{aligned} \quad (\text{A.13})$$

Since the factor $(-1)^N$ is unimportant in the characteristic equation, we arrive at (A.1)

Notice that the characteristic polynomial $Q(r)$ is indeed of degree N because $\prod_{i=1}^N [(1+\rho_i)r - \rho_i] - r^{N+1}$ is divisible by $(1-r)$. Also, $Q(r)$ does not depend on the specific order of the queues, that is, shuffle of the queue positions will not affect $Q(r)$.

Lemma A.2:

Let $\rho = \sum_{i=1}^N \rho_i$ be the system utilization and $\rho < 1$. Then there exists a positive real root of the characteristic equation $Q(r) = 0$ and it is bounded by ρ .

Proof of Lemma A.2:

Let $f(r) = \prod_{i=1}^N [(1+\rho_i)r - \rho_i]$ and $g(r) = r^{N+1}$

Clearly,

$$f(1) = 1 = g(1)$$

$$\text{and } f'(1) = \sum_{i=1}^N \rho_i + N < N+1 = g'(1).$$

Therefore, there exist $\epsilon > 0$, which can be arbitrarily small, such that $f(1-\epsilon) > g(1-\epsilon)$.

Next, we notice that $f(r) = 0$ has exactly N real roots at $r = \frac{\rho_i}{1+\rho_i}$, all lying between zero and one. Let t_{\max} be the maximum of $\frac{\rho_i}{1+\rho_i}$, then

$$f(t_{\max}) = 0 < g(t_{\max}).$$

Since $t_{\max} < 1$, we can choose ϵ such that $t_{\max} < (1-\epsilon)$. From the inequalities shown above, there must exist a real root of

$Q(r) = f(r) - g(r) = 0$ between t_{\max} and $(1-\epsilon)$. In fact, there is only one real root of $Q(r) = 0$ between t_{\max} and $(1-\epsilon)$ because $f(r)$ and $g(r)$ are both monotonic in that region. We shall call this root r_{\max} .

To show $r_{\max} < \rho$, we look at $(\ln f(r) - \ln g(r))$ on the interval $[t_{\max}, 1)$.

$$\begin{aligned} \ln f(r) - \ln g(r) &= \sum_{i=1}^N \ln[(1+\rho_i)r - \rho_i] - (N+1) \ln r \\ &= \sum_{i=1}^N \ln[1 - (1+\rho_i)(1-r)] - (N+1) \ln(1-(1-r)). \end{aligned}$$

Let $\delta = 1-r$ and expand the logarithmic function in Taylor's series at $r=1$ to obtain

$$\begin{aligned} \ln f(r) - \ln g(r) &= - \sum_{i=1}^N \sum_{k=1}^{\infty} \frac{(1+\rho_i)^k \delta^k}{k} + (N+1) \sum_{k=1}^{\infty} \frac{\delta^k}{k} \end{aligned}$$

$$\begin{aligned}
&= \sum_{k=1}^{\infty} \left[(N+1) - \sum_{i=1}^N (1+\rho_i)^k \right] \frac{\delta^k}{k} \\
&= \sum_{k=1}^{\infty} \left[(N+1) - \sum_{i=1}^N \left(1 + \sum_{j=1}^k \binom{k}{j} \rho_i^j \right) \right] \frac{\delta^k}{k} \\
&= \sum_{k=1}^{\infty} \left[1 - \sum_{i=1}^N \sum_{j=1}^k \binom{k}{j} \rho_i^j \right] \frac{\delta^k}{k} \\
&= \sum_{k=1}^{\infty} \left[1 - \sum_{j=1}^k \binom{k}{j} \sum_{i=1}^N \rho_i^j \right] \frac{\delta^k}{k} \\
&> \sum_{k=1}^{\infty} \left[1 - \sum_{j=1}^k \binom{k}{j} \left(\sum_{i=1}^N \rho_i \right)^j \right] \frac{\delta^k}{k} \\
&= \sum_{k=1}^{\infty} \left[1 - \left(1 + \sum_{i=1}^N \rho_i \right)^k + 1 \right] \frac{\delta^k}{k} \\
&= \sum_{k=1}^{\infty} 2 \frac{\delta^k}{k} - \sum_{k=1}^{\infty} \frac{(1+\rho)^k \delta^k}{k} \\
&= -2 \ln(1-\delta) + \ln(1-(1+\rho)\delta).
\end{aligned}$$

Setting $r = 1 - \delta = \rho$, we have

$$\ln f(\rho) - \ln g(\rho) > -2 \ln \rho + \ln \rho^2 = 0.$$

Therefore,

$$f(\rho) > g(\rho).$$

Since $\rho > t_{\max}$, we can deduce that $t_{\max} < r_{\max} < \rho$.

Lemma A.3

All the eigenvalues of A , real or complex, have magnitudes less than or equal to r_{\max} if $\rho < 1$.

Proof of Lemma A.3:

We prove Lemma A.3 using the famous Rouché's Theorem in complex analysis which states the following:

If $f(z)$ and $g(z)$ are analytic in a closed region on the complex plane with a simply connected boundary C and if $|f(z)| > |g(z)|$ on C , then $f(z)$, $f(z) + g(z)$, $f(z) - g(z)$ all have the same number of zeros inside C .

Let us consider a semicircle on the right half complex plane with centre at the origin and radius $R = r_{\max} + \epsilon < 1$ with $\epsilon > 0$ shown in Fig. A.1.

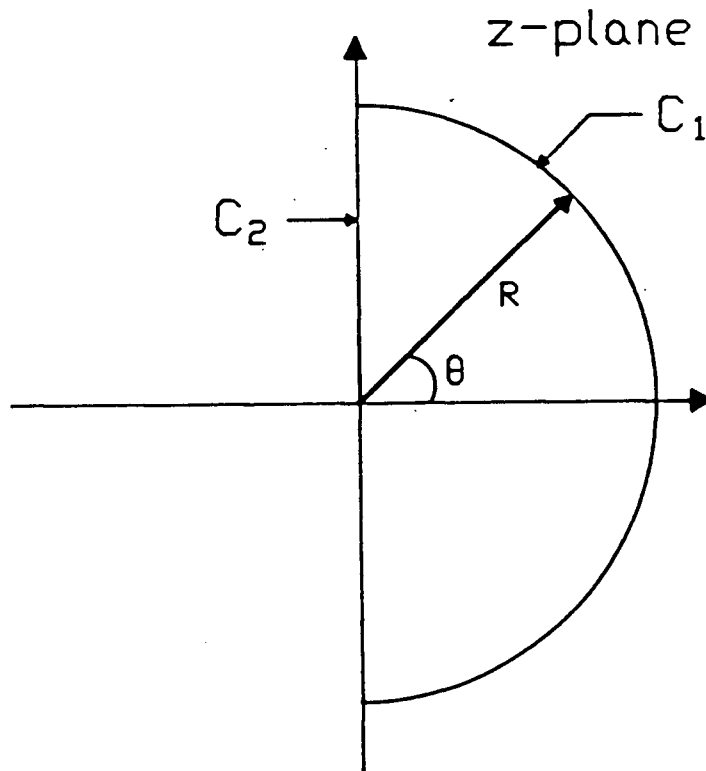


Fig. A.1 Right-half plane semicircle.

We let $f(z) = \prod_{i=1}^N [(1+\rho_i)z - \rho_i]$ and $g(z) = z^{N+1}$. On C_1 ,

$$\begin{aligned}
 |f(z)| &= \prod_{i=1}^N |(1+\rho_i)z - \rho_i| \\
 &= \prod_{i=1}^N |(1+\rho_i)R(\cos\theta + j\sin\theta) - \rho_i| \quad (j^2 = -1) \\
 &= \prod_{i=1}^N [(1+\rho_i)^2 R^2 + \rho_i^2 - 2(1+\rho_i)\rho_i R \cos\theta]^{1/2} \\
 &= \prod_{i=1}^N [((1+\rho_i)R - \rho_i)^2 + 2(1+\rho_i)\rho_i R(1 - \cos\theta)]^{1/2} \\
 &> \prod_{i=1}^N |(1+\rho_i)R - \rho_i| \\
 &= f(R) \\
 &> g(R) \\
 &= |g(z)|.
 \end{aligned}$$

On C_2 , $z = jy$ and

$$|f(z)| = \prod_{i=1}^N [(1+\rho_i)^2 y^2 + \rho_i^2]^{1/2} > y^N > y^{N+1} = |g(z)|$$

∴ By Rouché's Theorem, $f(z)$ and $f(z) - g(z)$ have the same number of zeros inside C . Clearly, $f(z)$ has exactly N real zeros inside C so $f(z) - g(z)$ should also have N zeros inside C . We already know that the $(N+1)$ th zero of $f(z) - g(z)$ is at $z = 1$. Thus $Q(z) = (f(z) - g(z))/(1-z) = 0$ has all N zeros inside C .

Therefore, all eigenvalues of A have magnitudes less than $R = r_{\max} + \varepsilon$. Since ε can be arbitrarily small, we can deduce that all eigenvalues of A have magnitudes less than or equal to r_{\max} . Moreover, $r_{\max} < \rho < 1$ shows that A has eigenvalues of magnitudes less than unity.