To the Graduate Council:

I am submitting herewith a dissertation written by Robert Allen McLaren entitled "A Study of the Behavior of Erlang Queuing Systems in Networks." I recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Management Science.

Ronald W. Boling, Major Professor

We have read this dissertation
and recommend its acceptance:

Gordon R. Sherman

D H Hutchinson

Robert A. McLean

Accepted for the Council:

Hilton A. Smith
Vice Chancellor
Graduate Studies and Research

A STUDY OF THE BEHAVIOR OF ERLANG QUEUING SYSTEMS IN NETWORKS

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee

Robert Allen McLaren

December 1974

ACKNOWLEDGMENTS

The author wishes to express his appreciation to the many individuals who contributed encouragement and assistance in this research: especially to his advisor, Dr. R. W. Boling, and to committee members, D. H. Hutchinson, R. A. McLean, and G. R. Sherman; to colleagues R. G. McMillan and J. N. Tunstall; and to Mrs. Marie Cuthbert and Mrs. Shirley Gentry for their patience in typing and editing the report.

This work was performed at United States Atomic Energy Commission installations operated by Union Carbide Corporation, Nuclear Division, under United States Government contract W-7405-eng-26.

1175477

ABSTRACT

Two major problems associated with the analysis of Erlang queuing systems in networks are studied. First, the superposition of heterogeneous Erlang renewal processes is considered. The distribution of the interarrival intervals of the resultant process is derived and the dependence introduced among the intervals is studied experimentally. It is found that this dependence has a significant effect on the performance of a service system to which the superposed input is presented. Methods for estimating queue behavior subject to this input dependence are suggested.

The interdeparture time distribution is approximated for one- and two-server Erlang queues. Dependence introduced in the output process is examined but is generally found to be of little significance to subsequent queue behavior.

The transmission of dependency through a network is also considered. For example, the superposition studies mentioned above involved source processes with independent arrival intervals. When there is mutual dependency among these intervals, however, superposition results in a process with different characteristics. Subsequent queue behavior may be considerably altered if the source correlation is high.

The results of these studies are then used to analyze two sample problems: a series of queues and a limited network. The mean queue lengths obtained by the analysis are compared to those obtained by simulation with encouraging results. Suggestions for extensions conclude the study.

TABLE OF CONTENTS

CHAPTER		PAGE
I.	INTRODUCTION	1
II.	LITERATURE REVIEW	6
	Erlang Queuing Systems	6
	Queuing Network Development	9
	Simulation of Queuing Networks	15
	Specialized Queuing Network Research	16
III.	SUPERPOSITION OF ERLANG PROCESSES	19
	The Distribution of Superposed Erlang Processes	21
	Dependency Introduced by Superposition	29
	The Limit Distribution of Merged Processes	38
	The Superposition of Interarrival Processes with Small Variances	39
	Estimation of Mean Queue Length for Queues with Superposed Input Processes	43
	The Superposition of Processes with Gamma Distributed Interarrival Times	45
	Summary	46
IV.	INTERDEPARTURE TIMES FOR ERLANG QUEUING SYSTEMS WITH ONE OR TWO SERVERS	47
	The Output of an $E_m/E_k/1$ System	48
	The Output of an $E_m/E_k/2$ System	57
	The Effects of Dependency in Interdeparture Times	64
	Summary	72

CHAPTER PAGE

V. SOME SECONDARY EFFECTS OF DEPENDENCY IN ARRIVAL AND DEPARTURE PROCESSES 73

Superposition of Sources with Dependent Interarrival Times 73

The Effect of Interarrival Time Dependency on Interdeparture Time Dependency 78

VI. ANALYSIS OF QUEUING NETWORKS 82

Series Queue Example 83

Queuing Network Example 84

VII. CONCLUSIONS 89

Summary of Results 89

Extensions 94

Outlook for the Future 96

REFERENCES 98

APPENDICES 105

VITA 132

LIST OF TABLES

TABLE PAGE

1. Coefficient of Variation of Interarrival Times for Homogeneous Source Streams and Resultant Processes 28
2. Mean Queue Length of Dependent and Independent Cases--Single Parameter Variations 32
3. Mean Queue Length of Dependent and Independent Cases--Variation of More Than One Parameter 34
4. Mean Queue Length of Dependent and Independent Cases--Comparison of Source Processes with Equal Variance in the Resultant Process 36
5. Effect of Dependency on Standard Deviation of Queue Length 37
6. Effect of Dependency in Interdeparture Times, Simulation Results 71
7. Effect of Dependency in Source Processes on the Dependency Resulting from Superposition 76
8. Interdeparture Time Dependency as Affected by Interarrival Time Dependency 81
9. Series Queue Example Results 86
10. Queuing Network Example Results 88
11. Comparison of Mean Queue Lengths for Erlang and Non-Erlang Input Processes 92
12. Results of Simple Queue Simulation 127
13. Analysis of Two Series Queue Experiment 129
14. Analysis of Superposition Results 130

LIST OF FIGURES

FIGURE PAGE

1. Superposition of Stochastic Processes 19
2. Density Functions of Superposed Interarrival Times for M Sources, m=2 24
3. Density Functions of Superposed Interarrival Times for M Sources, m=15 25
4. The Effect of Source Shape Parameter on Two Superposed Processes 26
5. An Arrival Pattern for Two Source Processes 29
6. Effect of Dependency Introduced by Superposition of M Sources with Shape Parameter 2 40
7. Effect of Dependency Introduced by Superposition of M Sources with Shape Parameter 15 40
8. Shape Parameter of Resultant Process Versus Shape Parameter of M Sources 42
9. Density Functions of Interdeparture Times of $M/E_k/1$ Queues with Traffic Intensity of 0.9 54
10. Density Function of Interdeparture Times of $E_m/E_2/1$ Systems with Traffic Intensity of 0.5 55
11. Coefficient of Variation of Interdeparture Times of $E_m/E_k/1$ Queues 55
12. Comparison of Density Functions of Interdeparture Times to Gamma Densities of Similar Means and Variances, Single-Server Queues 56

FIGURE PAGE

13. Coefficient of Variation of Interdeparture Times Versus Traffic Intensity for Single-Server Queues 58

14. Density Functions of Interdeparture Times from $E_2/E_k/2$ Queues with Traffic Intensity 0.9 65

15. Density Functions of Interdeparture Times from $E_m/E_2/2$ Queues with Traffic Intensity 0.5 65

16. Comparison of Interdeparture Time Density Functions of One- and Two-Server Queues 66

17. Comparison of Interdeparture Time and Gamma Density Functions with Similar Means and Variances for Two-Server Queues 67

18. Coefficient of Variation of Interdeparture Times Versus Traffic Intensity for Two-Server Queues. 68

19. Diagram of Simulations Used to Test Dependency Among Interdeparture Times 70

20. The Model Used to Test the Effect of Interarrival Dependency on the Interdeparture Dependency 74

21. Density Functions of Interdeparture Times from Queues 1 and 2 77

22. Systems Used for Comparing Effect of Input Dependency on Output Dependency 79

23. Equivalent Shape Parameter of Output Process as a Function of Input Process Shape 85

24. Configuration of Sample Network 87

CHAPTER I

INTRODUCTION

The abundance of waiting lines encountered in social, industrial, and mechanical environments has motivated researchers for most of this century. An extraordinary body of literature on queuing theory has resulted, much of the work being done by the ablest of mathematicians. The objective has usually been to develop a fundamental understanding of the behavior systems. If reliable conclusions of this type are to be reached, the variables of a system model and the relationships among them must be precisely defined and rigorously analyzed. Although necessary for much fundamental research, this approach has seriously constrained the study of queues to a relatively few tractable configurations. But practical problems are not so limited. The impact of queuing theory on real problems has thus been disappointing.

Researchers have not been unaware of this shortcoming--it has been pointed out repeatedly by some of the most prominent names in the field. Saaty in 1966 and Bhat in 1969 both called for a narrowing of the gap between queuing theory and practical problems (see Bhat [7, 1969]). Nor have these calls been completely unanswered. In the next chapter, some of the trends toward analysis of more general systems will be evident.

Another problem which soon becomes intractable when addressed rigorously is that of analyzing networks of queues. Certain specialized networks, such as tandem queues, have been the subject of

considerable work, with good approximations of system performance a common goal.

It is against this background that the objectives of the current paper were conceived. The basic subject, the analysis of Erlang queuing systems in networks, developed from an industrial problem. A production control system for a large manufacturing facility was to be analyzed; management had requested an independent evaluation of the effectiveness of the system. As a part of this evaluation, it was at one point thought useful to model the production facility as a network of queues. Most of the assumptions required by the theory were acceptable--one was not. Theoretical consideration of networks of queues required the assumption of exponentially distributed service times and this was considered unrealistic in the facility being analyzed. The queuing approach was thus abandoned.

Shortly after this experience, the results of work on the $E_m/E_k/s$ queuing system were published by Hillier and Lo [41, 1971] and Yu [69, 1972]. It thus became feasible to analyze, in steady-state, a system with a considerable range of variances in both the interarrival and service times. If networks of such queues could also be analyzed, then problems of the type described above might become tractable. The basic objective for this paper was thus set.

This objective, however, requires refining to obtain a problem of manageable size. The queuing system to be considered is first limited to queues with a single server or two homogeneous parallel servers with Erlang-distributed interarrival and service times. Typical assumptions apply: interarrival times are mutually independent, arrival

and service rates are independent of the state of the system, the system is in equilibrium, all processes are stationary, and queue lengths are unlimited. The way in which these queues may be interconnected also requires constraining. First, no feedback is permitted. Further, a single stochastic process may not be split into more than one process. Several arrival or intermediate processes may, however, be merged into a single input stream.

The equilibrium behavior of systems of queues constructed in this manner will be considered. The mean length of the queues which make up the system will be the primary measure of performance. The well-known relationship between expected values of queue length and waiting time can then be invoked (see, for example Maxwell [52, 1970]). Also of interest, but to be less vigorously pursued, is the variance of queue behavior in networks.

The approach to the analysis of networks begins by considering the merging, or superposition, of independent Erlang processes. The objective is to determine the characteristics of the resultant stochastic process which is then used as the input to a service system. The first task addressed is, therefore, the development of the distribution function of a process resulting from the superposition of Erlang processes.

Also required for the analysis of networks of queues is the interdeparture time distribution. The development of sufficiently accurate approximations of these functions is thus the second basic task.

It is well known that superposition of processes, each of which have mutually independent interarrival times, will generally result in a process with correlated interarrival intervals. Also, the inter-departure intervals of most queuing systems are correlated. The third major task of this paper is to analyze the effect of this dependency on the behavior of the queues within the network. Further, it is necessary to determine to what extent the dependency and its effects tend to accumulate in larger systems.

There is a further objective implicit throughout this study: the observation of the sensitivity of queue behavior to deviations from the assumed distribution functions. It is conjectured that the mean and variance of interarrival or service times are generally sufficient to uniquely determine the expected behavior of the queue. This conjecture is tested repeatedly, insofar as the interarrival times are concerned, throughout this paper. The process resulting from superposition of Erlang sources and the output process of an Erlang queue are not Erlang processes. Estimates of subsequent queue behavior are made, however, by using values (often interpolated) from Hillier and Lo's tables for Erlang systems. Comparisons of these estimates to results of simulations which use the actual distributions will be seen to provide supporting evidence to the conjecture.

Simulation is used throughout this study for verifying results and for the evaluation of the role dependency plays in system performance. In the process, both the power and the limitation of this tool are illustrated. On the one hand, simulation models of queuing networks are easily developed; further, they can generally be subjected to

several tests for validity. The price paid in computer time is high, however, and can be seen to increase directly with the size (number of queues) of the problem being analyzed. For very large problems, perhaps involving hundreds of queues, the cost could easily be prohibitive. It is this fact which gives a final impetus to the search for an alternative approach--the need for such encouragement will be evident as the difficulties of developing this alternative are revealed.

A review of the literature follows in the next chapter, beginning with a discussion of Erlang queues. Work on networks of queues and simulation of queuing systems follow. In Chapter III, the topic of superposition of Erlang renewal processes is developed, concluding with an examination of the dependency introduced among the interarrival intervals in the resultant process. The output of an Erlang Queuing system with one or two parallel servers is treated in Chapter IV. Approximations to the interdeparture time distribution are developed and dependency in the departure stream examined experimentally. Details in Chapters III and IV are included as appendices.

The way in which dependencies among arrival intervals are transferred through a network of queues is briefly examined in Chapter V. First, there are experiments involving the superposition of non-independent processes. Also considered is the effect of dependency in the input process upon the output process of a queue.

All of these findings are brought together in Chapter VI for the analysis of two sample problems: a system with four tandem queues and a four-queue network. Conclusions and extensions are presented in Chapter VII.

CHAPTER II

LITERATURE REVIEW

Two major topics will be covered in this chapter: the development of equilibrium state probabilities of $E_m/E_k/s$ queuing systems and work specifically related to networks of queues. Literature concerning superposition and the output processes of queues will be examined in Chapters III and IV.

Erlang Queuing Systems

Queuing systems with Poisson input and exponentially distributed service times have provided useful approximations to many queuing situations. It is widely recognized, however, that generalization of this model will broaden the potential applications of the theory (see, for example, Bhat [7, 1969]). Analysis of models with completely general (independent) input and general service time distributions, the GI/G/s system, is unlikely to be realized (see Kingman [48, 1966, p. 37] and Wishart [67, 1956]). Some progress has been made, though, in approaching this ideal, culminating in the recent work of Hillier and others with the $E_m/E_k/s$ system. The development of this, and other related results, will now be traced.

Some of the more tractable systems were analyzed early in queuing theory history (see Kendall [46, 1953]). The M/M/s and $M/E_k/1$ systems were developed by A. K. Erlang. The $E_k/G/1$ system was independently analyzed by F. Pollaczek in 1930 and by O. Volberg in 1939. The M/G/1 queue was first considered by Khintchine in 1932. Kendall also worked on this system in 1951. The GI/G/1 queue was independently

developed by D. V. Lindley, F. Pollacznek, and W. L. Smith, all publishing in 1952. Wishart provided system state probabilities for the $GI/E_k/1$ system [67, 1956]. Useful expressions summarizing much of this work were given by Saaty [61, 1957]. Saaty also reported in his book the work by R. R. P. Jackson and D. G. Nickols on the $E_k/M/1$ systems [62, 1961]. Restropo determined the mean queue length and waiting time for single-server systems with Erlang service and group Poisson arrivals [60, 1965].

Key to much of the more advanced work in queuing theory is the imbedded Markov chain, introduced by Kendall [46, 1953]. This technique is discussed in most subsequent books on queuing theory. (See, for example, Bharucha-Reid [6, 1960, Chapter 9] and Cooper [19, 1972, Chapter 5].)

Foundations for work with multiple-server systems were laid by Kiefer and Wolfowitz [47, 1955]. They proved that, under very general assumptions, the waiting time distribution approaches a limit as the number of arrivals approaches infinity. The traffic intensity must be strictly less than one for this distribution to exist. Further, the limiting distribution is independent of the initial state of the system.

More specific treatment of multiple-server systems began with Shapiro's work with the $M/E_2/s$ system [65, 1966]. Using the fact that Erlang distributed random variables are equal to sums of exponentially distributed random variables, the service time was represented as two "stages" of exponential service. This approach, appropriately called staging, is basic to the analysis of Erlang systems. A Markov chain formulation can then be applied to the staged system.

Using this same approach Mayhugh and McCormick [54, 1968] and Heffer [37, 1969] developed state probabilities for the $M/E_k/s$ system in equilibrium. Simplifications to the former technique, which the authors reported was tractable only for small values of k, were given subsequently by Mayhugh [53, 1970].

To overcome some of the complexities in the analysis of the $M/E_k/s$ system, Maaloe has developed relatively simple expressions for computing bounds on the expected waiting time [51, 1973]. Similarly, Brumelle shows that the mean wait time in a GI/M/s system is bounded above by that in an associated GI/M/1 system with service time proportionately reduced [9, 1973].

A variation to the major thrust of work on Erlang systems was presented by Love [50, 1970]. He developed the expected waiting time for an $E_k/M/s$ system with bulk service.

Finally, the more general $E_m/E_k/s$ system has been analyzed, using an approach similar to that of Heffer's. The difficulty of this analysis results from the very large number of system states generated by the process's stages. Simply keeping track of these states is non-trivial. Furthermore, the solution of the steady-state equations derived from the resultant Markov chain is difficult. A procedure for their solution was developed by Yu [69, 1972] and computations have been carried out by Hillier and Lo [41, 1971]. Solutions are feasible only for rather limited values of shape parameters and number of servers. Nevertheless, this work constitutes a very significant advancement in analyzing more general queuing systems and the results are of fundamental importance to this paper.

Queuing Network Development

In this section, the development of theory relating to general queuing networks will be examined. Included in this will be the important special network of queues in series. More specialized networks will be covered in a later section.

R. R. P. Jackson provided an early contribution to the theory of series queues [45, 1956]. He investigated a system with Poisson imput, unlimited interstage storage, and multiple servers. Service times were exponential and identically distributed for all servers at a stage. He obtained expressions for equilibrium state probabilities and, for a series of single-server queues, the mean queue length and waiting time for each stage.

Of particular importance in the study of queues in networks is the finding by Burke [10, 1956] regarding the distribution of interdeparture times from an M/M/s system. He proved that this distribution was exponential with the same mean as the interarrival time. Since the Poisson process was preserved by the system, a series of queues with exponential service could be analyzed independently. Furthermore, the additive property of the Poisson process suggested that more complex networks might be treated.

J. R. Jackson [43, 1957] proved that two additional assumptions would permit rather general networks of multiple-server Markovian queues to be analyzed independently. First, the routing of customers through the network must be defined by a transition matrix. The proportion of customers departing the system following service is also defined in this matrix. A second requirement is that the queue

discipline must be independent of both the routing and the service time. When such a system is in equilibrium, the individual queues will behave independently and each can be analyzed as a simple M/M/s queue.

Jackson later studied a series of single-server queues which represented a jobshop [44, 1963]. Poisson arrival and exponential service were assumed. Parts did not need to go through all service facilities, but could skip ahead according to a predefined routine. The unique feature in this study was that the arrival rate could be a function of the number of jobs already in the shop and the service rate of each station was a function of the queue length at the station. The joint probability distribution of all queue lengths was obtained. Contemporary with this work was a model by Evans in which the order of service in a series was unimportant [28, 1964]. A unique feature of this model was that a customer could be simultaneously processed by more than one server.

Also using Poisson input and exponential service, Nelson [55, 1958] developed the cumulative probability distribution function of waiting times for each queue in a series of multiple-server systems. Hunt [42, 1956] found the state probabilities and utilizations of a series of M/M/1 queues under a variety of assumptions about intermediate storage.

Concurrent with these early successes was a discouraging demonstration by Reich [56, 1957]. He showed that Burke's and Jackson's results could not be extended to systems with even minor departures from the Poisson and exponential distributions. Using identical Erlang distributions for interarrival and service times, he proved that, in

general, the output of the system is not described by Erlang inter-departure times.

A further limitation on the generality of Burke's result was revealed by Chang [14, 1963]. He showed that the interdeparture times of a GI/G/1 system are mutually independent only for the special case of M/M/1.

Further results did begin to accumulate around the Markovian systems, however. Reich [57, 1963] considered a series of single-server queues with Poisson input to the first queue. He showed that under equilibrium conditions the waiting times of a customer in each queue are mutually independent. The waiting time includes time spent in service. A somewhat surprising sequel to this result, provided by Burke [11, 1964], showed that the waiting times in the queues, exclusive of service, are not mutually independent in such a system.

Expanding further on Reich's 1963 work, Burke [12, 1968] also showed that independence of waiting times continued to hold if the first and last queue in the series were multiple-server Markovian queues. In a subsequent publication [13, 1969] he demonstrated that single-server queues in the intermediate locations were necessary to the independence of waiting times, apparently concluding this avenue of research.

Applications in production lines provided strong motivation for continued work in series queues, however. The results reviewed thus far are limited to queues with exponential service and, for the most part, unlimited interstage storage. Subsequent research is characterized by a progression to less restrictive assumptions about the service

distribution and to models with finite storage. Ghosal made one of the initial contributions in this direction by analyzing a series of single-server queues, the first service time of which could be gamma-distributed. Waiting time distributions were derived [32, 1962].

Two single-server queues in series were investigated by Avi-Itzhak and Yadin [5, 1965]. Finite or no interstage storage, Poisson input, and general service time distributions were considered, with particular attention to exponential and regular service. They demonstrated that interstage storage increased the capacity of the system when service was exponential but was of no benefit with regular service. Also, the order of the two service stations was found to be unimportant.

Avi-Itzhak also investigated a series of multiple-server queues, each station of which has the same number of homogeneous servers [3, 1965]. Input is arbitrary, service is regular, and intermediate queues of finite length are permitted. He showed that the time spent in the system by a customer is independent of both the order of the stations and the allowable size of the interstage queues. Also, the waiting time of a customer, excluding service, is the same as would have been obtained in a single stage system with regular service equal to the longest service time in the system. A similar problem, but with unlimited interstage storage, was considered by Friedman [31, 1965].

Hillier and Boling [40, 1967] developed a numerical method for analyzing a system made up of a series of single-server queues with exponential or Erlang service time distributions. Finite interstage storage and an infinite supply of customers to the first queue were

assumed. The mean output rate and the mean number of customers in the system were found. Efficient computational procedures were stressed. With service time distributions limited to the exponential case, Hillier and Boling then used these methods to analyze the effect of unbalancing a production line [39, 1966]. They found that for a series of at least three stations, the expected production rate could be increased by reducing the mean service time of the middle stations of the line. Boling later used a similar model to study sequential crew assignments [8, 1970]. He showed that when mean service rates of crews are unbalanced, the order of service does not significantly affect the production rate. Order does, however, influence the expected number of units being serviced by all crews.

The practicality of performing the calculations in such models is often determined by the number of feasible system states. A method of enumerating the feasible system states of a series of finite, single-server queues is given by Haydon [36, 1972]. Two subsystems of queues, the feasible system states of which are known, can be combined into a longer series of queues. Using Haydon's approach, the feasible states of the expanded system can be obtained. Service times may be Erlang distributed. Haydon cites work in progress by I. E. Scales as handling the multiple-server case.

The references cited thus far have concentrated on more general storage and service time distribution assumptions. Hildebrand focused on the arrival process [38, 1968]. He permitted general independent arrivals to a series of single-server queues with exponentially

distributed service times. Given finite interstage storage restrictions, he developed expressions for the capacity of the system.

In contrast to the rather abundant literature on series queues, more general network research is sparse. At the 1965 Symposium on Congestion Theory, T. L. Saaty suggested that the concepts of graph theory be adopted to provide a common approach to the study of queuing networks [63, 1965]. He expressed the opinion that queuing networks should be treated as a whole, instead of being decomposed into separate queues. The objective would be to eventually formulate general laws which govern the behavior of complex systems. There is little evidence that practical work has advanced along these lines, however.

Not all participants at the symposium agreed that decomposition was the less favorable approach. Chang considered decomposition as necessary. In his discussion of Saaty's paper, Evans noted some of the complexities which are introduced in networks. In particular, the dependencies introduced in the output intervals was disturbing. He suggested that relatively simple systems be used to approximate the more complex behavior.

Another problem fundamental to general networks is that of splitting a stochastic process into two or more processes. The term "decomposition" is also used to describe this phenomenon. Little work has appeared on this subject. Disney [23, 1967] considered three rules for deciding which branch an event in the original stochastic process would take. First, a Markov discipline was examined, in which the probability of an arrival taking a particular path would depend only upon the path taken by the previous arrival. Secondly, the path might

be determined by some attribute of the arrival. Finally, switching based upon the state of subsequent service facilities was considered. Disney also cited the need for research on the problems of "recomposition" (superposition) and the departure processes from queues.

Simulation of Queuing Networks

Simulation has been the most effective tool in dealing with general queuing networks and examples of its application are abundant. The development of simulation languages such as the General Purpose Simulation System (GPSS) and SIMSCRIPT have contributed to this success.

An early, and rather unique, simulation approach was reported by Dunn, Flagle, and Hicks [26, 1956]. An electromechanical device was constructed to analyze networks of single-server queues. Arrivals and service times were input on punched paper tape and could represent arbitrary distributions. Furthermore, the usual independence assumption of arrival intervals was not required. Stepping switches, representing queue lengths, were actuated by arrivals and by service completions. A plot of line lengths was recorded and an interval counter accumulated the time spent in each system state. Multiple inputs and a variety of queue disciplines could be handled.

Conway and Maxwell used simulation extensively in their investigation of jobshops [18, 1967]. An example of simulation of networks was their study of the effect of various queue disciplines on shop performance [17, 1962]. A queue network simulation language developed for the Burroughs 220 computer was used.

A review of discrete event simulation and several applications are presented in a recent book by Reitman [59, 1971]. The evaluation

of a computer system (Chapter II) is a particularly appropriate example. Using GPSS, the performances of different communications and computer systems were compared under varying system loads.

It is possible that analog techniques will again find favor with researchers. Saaty reported on the study of traffic flow problems using equations similar to those found in fluid mechanics [63, 1965]. The application of these methods to network representation has apparently not been attempted, however.

Specialized Queuing Network Research

In addition to series queues, there are several queue configurations which form simple networks which have received attention in the literature. In this section cyclic queues, queues with feedback or overflow, and some special network studies will be briefly examined.

Cyclic queuing systems are typified by several series queues with the output of the last queue becoming the input to the first. These are "closed" systems in that there are a fixed number of customers in the cycle. A good example is provided by Koenigsberg [49, 1960]. Consider A machinists who operate N identical machines. There may be more machines than machinists and excess operable machines form a queue awaiting use. Machines fail at intervals described by an exponential distribution and then enter a queue to be repaired by M repairmen. Repair time is also exponential and repaired machines enter the excess queue. The state probabilities of this system were derived.

More general closed systems were analyzed by Gordon and Newell ([33, 1967] and [34, 1967]). Several stages of multiple server queues with homogeneous exponential service were permitted. A fixed number

of customers progressed through the network in accordance with some fixed probability, p_{ij}, that a customer leaving stage i will go to stage j. The states of the system are the number of customers at each stage. Using a Markov process representation, the equilibrium state probabilities were determined. This closed system was also shown to be equivalent to an open system with a single input point.

Another type of multiple-queue system is the two-server queue with overflow. Poisson input to a single-server queue with exponential service time is assumed. If the line length exceeds a fixed number, arrivals are diverted to another facility. If the line length of the second system exceeds a quantity, N, the customer is lost. Cinlar and Disney [16, 1967] showed that this overflow system is GI/M/1/N. In a later publication [24, 1972], Disney developed a solution for the steady-state probabilities for this system, although computations are cumbersome.

The need to understand complex computer systems, particularly those involving time-sharing, have initiated a considerable amount of work in specialized queuing networks. Typical of this work is a model of a time-sharing system by Adiri and Avi-Itzhak [1, 1969]. Multiple queues of varying priority are serviced by a single facility. Once a job (customer) receives service (a time slice), it either leaves the system or joins a lower priority queue for additional service. (This service discipline is designed to discourage long jobs.) Poisson arrivals and exponential service are assumed.

There are several other similar models. In a model by Chang, an incoming job was assigned a priority which it retained until finished

[15, 1968]. Schrage [64, 1967] and Takács [66, 1968] permitted general service time distributions. A system in which each job requires exactly N services was studied by Enns [27, 1969].

A multiprogrammed computer with a fixed number of tasks (like IBM's OS/MFT operating system) was modeled by Avi-Itzhak and Heyman [4, 1973]. Each task resides in core and requires, with some probability, service on various peripheral devices each time the task is active. These devices are modeled as homogeneous parallel servers with exponential service times. Customer cycle times and device utilizations were found. The sensitivity to the distribution assumptions was investigated and it was concluded that the distribution assumptions were critical when there were few jobs and many peripheral devices. There was little sensitivity to these assumptions if there were more than ten jobs in the system, however. Another interesting conclusion was that system performance was improved by having single-server peripherals with short mean service times rather than having multiple-server devices with proportionately slower service.

Work of the type reported in this section appears responsive to Saaty's enjoiner to study whole systems. Some elements of the underlying behavior of a system are evident in the last article by Avi-Itzhak and Heyman. Sufficient generality is not achieved, however, to confidently extend the results to other systems. Continued study of the interface between individual queues thus seems warranted; this interface will be examined in the following chapters.

CHAPTER III

SUPERPOSITION OF ERLANG PROCESSES

A process in which events occur through time and for which the times between events are identically distributed, mutually independent random variables is called a renewal process. Such processes are typically assumed as the input, or arrival, process to a queuing system.

In a network, however, the arrivals to a service facility may come from several different sources. These sources are merged together into a single stochastic process by the superposition of the arrival epochs of the source processes. This is illustrated for two source processes in Figure 1.

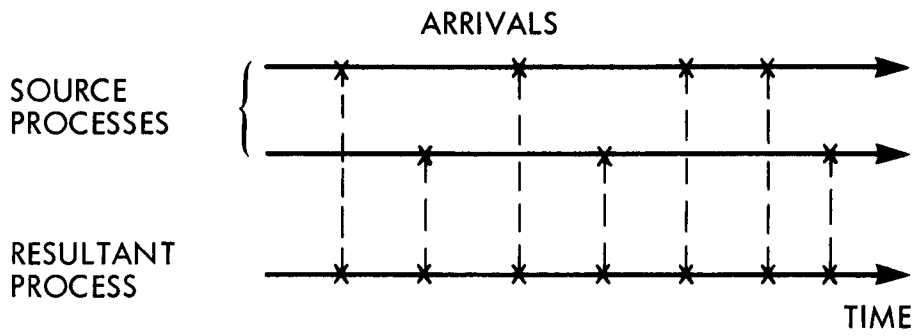


Figure 1. Superposition of stochastic processes.

It is well known that the superposition of Poisson source processes will yield a Poisson resultant process (see, for example, Cox [20, 1962, page 73]). An equivalent result does not hold for most other interarrival time distributions, however. Typically, the

resultant interarrival times will not have the same distribution as the sources, nor will these times be mutually independent.

This has severely limited the study of multiple inputs to queues. Anker and Gafarian considered a system with multiple Poisson arrival streams to a single server with exponential service times [2, 1961]. The important characteristic of the model was that each source imposed a different mean service rate on the server. Harris, also working with a single, exponential server, permitted single or doublet arrivals from multiple Poisson sources [35, 1965].

A departure from Poisson arrivals is seen in a model by Young in which there are two sources, one Poisson, the other Erlang [68, 1966]. The multiple servers have exponentially distributed service times. No queue is permitted. Young shows that the state probabilities of the system are a function of the arrival rate in the two channels and not dependent on the shape parameter of the Erlang source. The application is to an adaptive control model in which the state of the system influences the number of stages in the Erlangian arrival stream.

In this chapter the superposition of nonhomogeneous renewal processes with Erlang distributed interarrival times will be examined. Theory related to this development is found in Cox [20, 1962, Chapter 6] and Feller [29, 1966, Chapter XI]. The distribution and variance of the resultant interarrival times will be developed first. The dependency introduced by superposition will then be explored with emphasis on the effect that the dependency has on the behavior of a subsequent queue. Some observations about the convergence of superposed processes to a Poisson process will also be made and the robustness of queue

behavior to the shape of the source distributions will be briefly examined.

The Distribution of Superposed Erlang Processes

Consider M source streams, each of which is a renewal process with interarrival time distributions $F_j(t)$, $j=1,\ldots,M$. The M processes are mutually independent. Assume the system has been operating since time $-\infty$ and that a customer arrives from arbitrary source j. Denote the event that the arrival is due to this source as E_j. Further, assume that this event occurs at time zero. The probability that the next interarrival interval, I, due to an arrival from any source, is less than time t is given by

$$\Pr(I < t|E_j) = 1 - \prod_{i=1}^{M} (1 - \Pr(t_i < t)) \qquad (3.1)$$

where t_i is the time to an arrival from source stream i.

For the case of i=j,

$$\Pr(t_j < t) = F_j(t) \quad . \qquad (3.2)$$

For all other sources, $\Pr(t_i < t)$ is the residual waiting time distribution of $F_i(t)$, which, under steady-state conditions is given by

$$H_i(t) = \lambda_i \int_0^t (1 - F_i(y))\, dy \quad . \text{ (Feller [29, 1966, p. 354])} \qquad (3.3)$$

The λ_i in (3.3) is the mean arrival rate for source stream i. Expression (3.1) can now be written

$$\Pr\,(I < t|E_j) = 1 - (1 - F_j(t)) \prod_{\substack{i=1 \\ i\neq j}}^{M} (1 - H_i(t)) \quad . \tag{3.4}$$

For Erlang source streams, the distribution functior for interarrival times is

$$F_i(t) = 1 - e^{-\lambda_i m_i t} \sum_{r=0}^{m_i-1} \frac{(\lambda_i m_i t)^r}{r!} \tag{3.5}$$

with m_i being the shape parameter of the distribution.

The residual waiting time distribution for an Erlang process is developed in Appendix B; see expression (B.2), page 108.

$$H_i(t) = 1 - e^{-\lambda_i m_i t} \sum_{r=0}^{m_i-1} \left(\frac{m_i - r}{m_i}\right) \frac{(\lambda_i m_i t)^r}{r!} \quad . \tag{3.6}$$

Turning now to the unconditional distribution of the merged interarrival time, which will be denoted G(t), one may write

$$G(t) = \Pr\,(I < t) = \sum_{j=1}^{M} \Pr\,(E_j)\,\Pr\,(I < t|E_j) \quad . \tag{3.7}$$

The probability that the last event occurred in stream j is proportional to that stream's total contribution to the merged process:

$$w_j = \Pr\,(E_j) = \frac{\lambda_j}{\sum_{i=1}^{M} \lambda_i} \quad . \tag{3.8}$$

Finally, noting the similarity between expressions (3.5) and (3.6) permits G(t) to be specifically written as

$$G(t) = 1 - e^{-t\sum_{i=1}^{M}\lambda_i m_i} \sum_{j=1}^{M} w_j \prod_{i=1}^{M} \left\{ \sum_{r_i=0}^{m_i-1} A_{ir_ij} \frac{(\lambda_i m_i t)^{r_i}}{r_i!} \right\} \qquad (3.9)$$

where

$$A_{ir_ij} = \begin{cases} \left(\frac{m_i - r_i}{m_i}\right) , & i \neq j \\ 1 , & i = j \end{cases} .$$

The corresponding density function is also derived in Appendix C and appears as expression (C.5) on page 112. Plots of this function are shown in Figures 2, 3, and 4. Figure 2 illustrates the effect of increasing numbers of source streams, each of which has a shape parameter of 2. Figure 3 gives equivalent curves for a shape parameter of 15. Figure 4 shows the effect of increasing shape parameter in a system with two homogeneous sources. In all of these cases, the means of the source processes were chosen to yield a mean of 1 in the resultant process. It is apparent that this resultant mean is given by

$$E(t) = \frac{1}{\sum_{i=1}^{M} \lambda_i} . \qquad (3.10)$$

The variance of the resultant process is derived in Appendix C, pages 112 through 115. The results from which the variance is obtained are as follows:

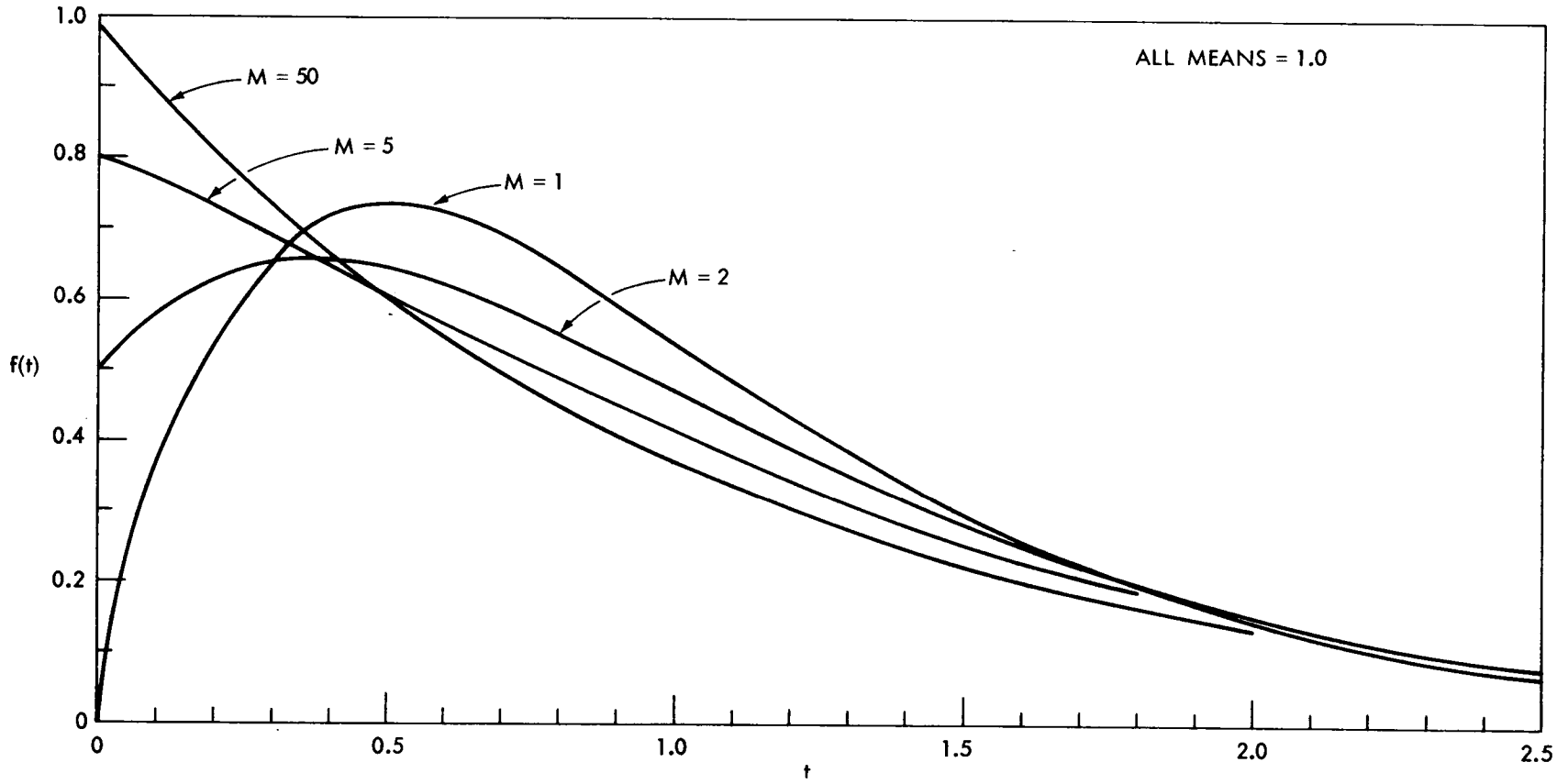


Figure 2. Density functions of superposed interarrival times for M sources, m = 2.

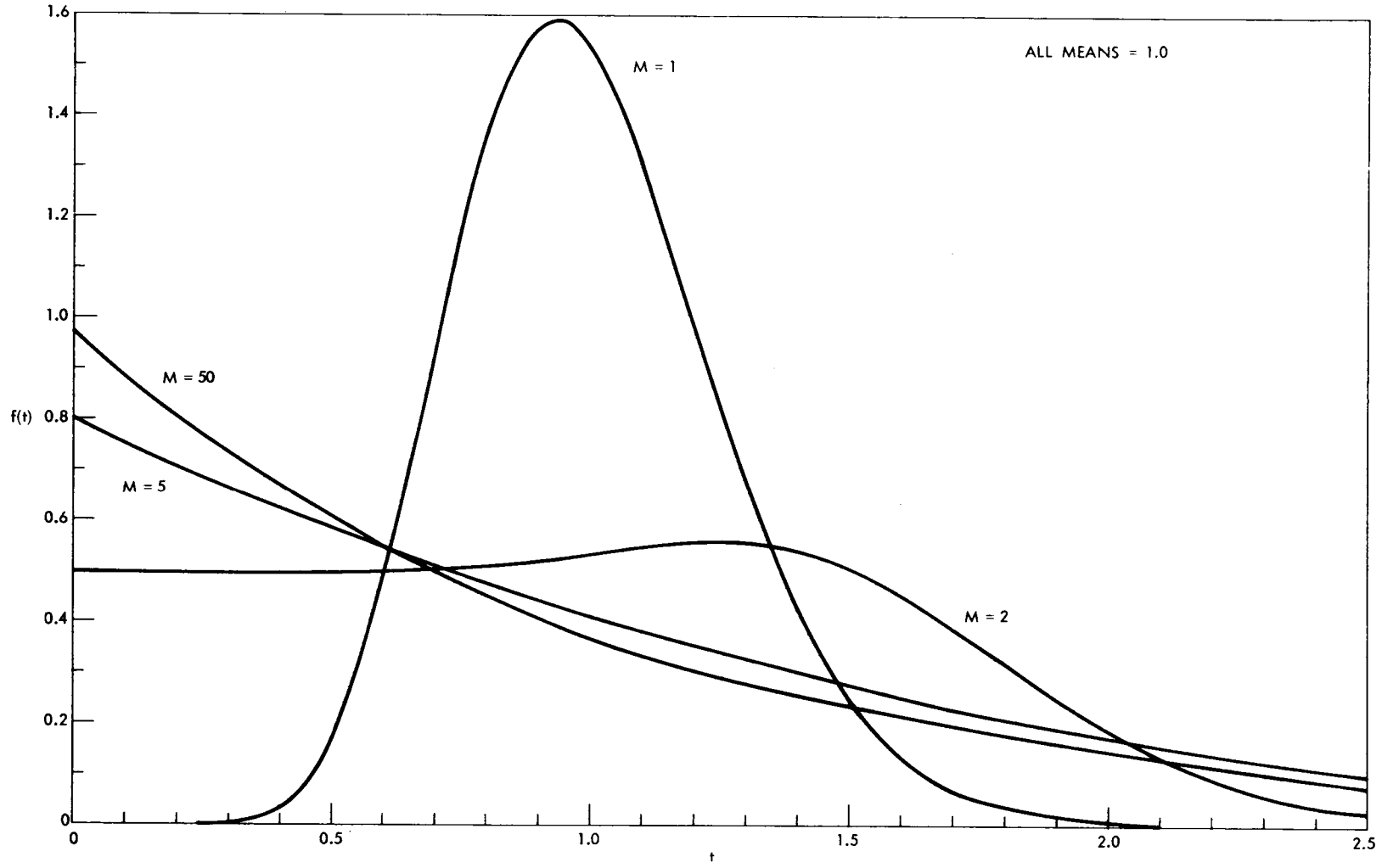


Figure 3. Density functions of superposed interarrival times for M sources, m = 15.

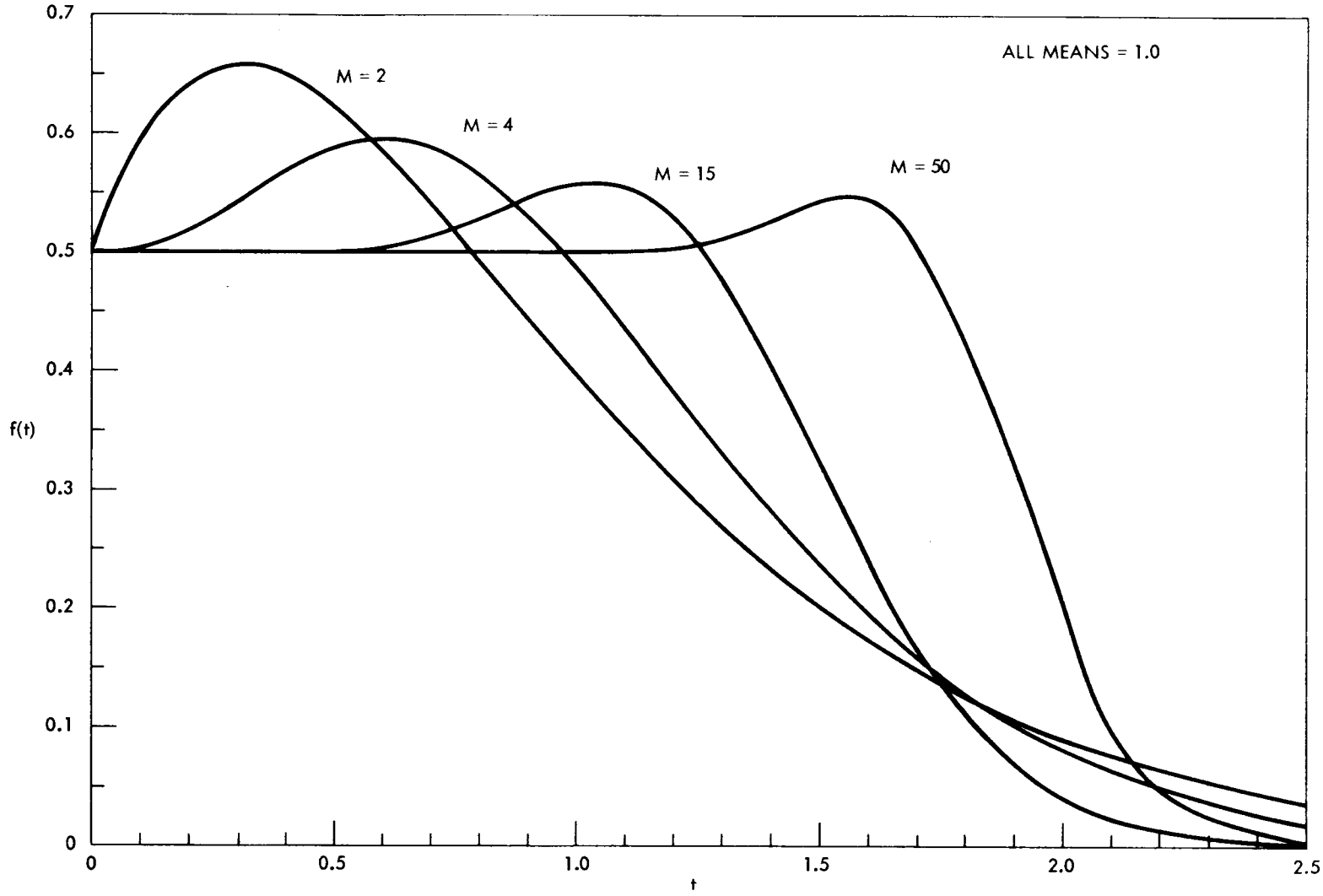


Figure 4. The effect of source shape parameter on two superposed processes.

Let:

$$C_{ir_ij} = \frac{A_{ir_ij}\,(\lambda_i m_i)^{r_i}}{r_i!} \qquad \text{for all } j \quad ,$$

and define coefficients, $B_{\ell j}$, as the sum of all products, $\prod_{i=1}^{M} C_{ir_ij}$, such that $\sum_{i=1}^{M} r_i = \ell$. Finally, define $N = \sum_{i=1}^{M} (m_i - 1)$. Then

$$E(t^2) = 2 \sum_{j=1}^{M} w_j \sum_{\ell=0}^{N} \frac{B_{\ell j}\,(\ell + 1)!}{\left(\sum_{i=1}^{M} \lambda_i m_i\right)^{\ell+2}} \tag{3.11}$$

Table 1 shows the effect that the input stream shape parameter and number of input streams has upon the variance of the resultant process. Homogeneous input processes were used.

Figures 2 and 3 on pages 24 and 25, respectively, and Table 1 all illustrate the tendency for superposed processes to converge to a Poisson process. In the figures, the resultant density of interarrival times for 50 superposed source streams is almost indistinguishable from an exponential density with unit mean. The table similarly shows the coefficient of variation approaching 100, the convergence being more rapid when the source processes themselves have a higher variance. This effect is moderated, insofar as the behavior of a subsequent queuing system is concerned, by the effect of the mutual dependency of interarrival times induced by superposition.

Table 1. Coefficient of variation of interarrival times for homogeneous source streams and resultant processes.

Number of Streams, M	Input Stream Shape Parameter, m	CV of Source Process	CV of Resultant Process
2	2	70.71	79.06
5	2	70.71	87.93
10	2	70.71	92.68
50	2	70.71	98.16
2	15	25.82	59.49
5	15	25.82	81.73
10	15	25.82	90.45
50	15	25.82	94.23
2	4	50.00	67.46
2	8	35.36	61.81

Dependency Introduced by Superposition

The fact that superposition of renewal processes typically introduces dependency in the resultant interarrival intervals is well documented (see Cox [20, 1962, pp. 75-76]). The basic mechanism that generates the dependency is that the residual waiting time is not generally independent of the elapsed time in the interval. The following example illustrates this point for a particular arrival pattern from two homogeneous Erlang processes. Figure 5 shows the arrival configuration.

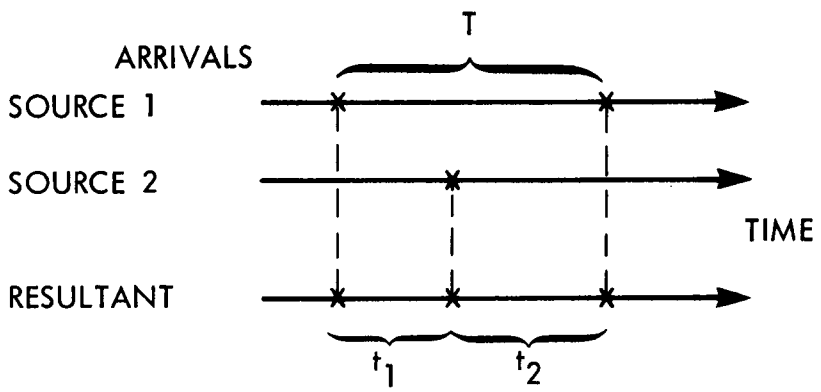


Figure 5. An arrival pattern for two source processes.

An arrival from source 2 occurs in time t_1 following an arrival from source 1. The dependency of time t_2 upon t_1 can be seen by writing the following:

$$F_{t_2}(t) = \Pr(t_2 < t \mid T > t_1) = \frac{\Pr(t_2 < t \text{ and } T > t_1)}{\Pr(T > t_1)} \qquad (3.12)$$

$$= \frac{\Pr(t_2 < t)}{\Pr(T > t_1)}$$

For homogeneous Erlang distributions, (3.14) is specifically,

$$F_{t_2}(t) = 1 - e^{-\lambda m t} \quad \frac{\sum_{r=0}^{m-1} \frac{(\lambda m)^r (t + t_1)^r}{r!}}{\sum_{r=0}^{m-1} \frac{(\lambda m)^r t_1^r}{r!}} \tag{3.13}$$

For the special case of exponential distributions, (3.13) readily reduces, as expected, to

$$F_{t_2}(t) = 1 - e^{-\lambda t} \quad .$$

Similar results can be demonstrated for arrival patterns other than the one shown in Figure 5.

The effect the dependency has upon queue performance has not been developed analytically. Therefore, to measure the dependency introduced, two queuing systems were simulated. Each system consisted of multiple, independent source processes with Erlang distributed interarrival times. These processes were merged into a single input stream to the service facility. Thus, the input stream contained the dependency introduced by superposition. In the results which follow, this will be referred to as the dependent case.

The second queuing system, the independent case, had for its input stream a renewal process with interarrival times drawn from the distribution described by expression (3.9). Parameters for the source streams were the same as for the first system. Thus, the two queuing systems were identical in all ways except for the mutual independence, or lack thereof, of the interarrival times.

Comparison of mean queue lengths between the dependent and independent cases was the primary measure of the effect of dependency. The validation of these simulation results is discussed in Appendix E. Generally, the mean queue lengths may be accepted as valid to within 7%, which is adequate for the purposes of this comparison.

Over 30 cases were simulated, most of which involved homogeneous input streams. Table 2 shows some of the more important results for variations of a single parameter. It may be noted that higher values of the shape parameter of the interarrival time distributions have an important effect on queue behavior. Similarly, there is a range of numbers of sources which have a marked effect. Lower values of traffic intensity exhibit a significant percentage change in mean queue lengths but the absolute differences are small. Service shape parameter and the number of servers appears to be relatively unimportant.

Results from variations of more than one parameter at a time are shown in Table 3. The first line is a base case from which variations are compared. The second case varies both the input and service shape parameters. When these parameters were varied individually, the ratios of independent to dependent mean queue lengths were 1.55 and 1.18, respectively. One may note that the cumulative effect of simultaneous variation is greater than either of these: 2.00. Similar results are shown on the third line for variations of input shape and traffic intensity. The final case represents a "worst" case, incorporating the values of each parameter which yields, individually, the largest ratio. Again, the cumulative effect is greater than any of the individual effects. It is not clear from these results, however, that

Table 2. Mean queue length of dependent and independent cases - single parameter variations.

Symbols:
- m: arrival shape parameter
- k: service shape parameter
- M: number of sources
- ρ: traffic intensity
- s: number of servers
- $\overline{Q}_i$: mean queue length, independent case
- $\overline{Q}_d$: mean queue length, dependent case
- R: $\overline{Q}_i/\overline{Q}d$

Variation with Arrival Shape Parameter

	m	$\overline{Q}_i$	$\overline{Q}_d$	R
M=2	1	5.8	5.8	1.00
ρ=.9	2	4.1	3.7	1.11
k=2	5	3.4	2.6	1.31
s=2	15	3.1	2.0	1.55

Variation with Service Shape Parameter

	k	$\overline{Q}_i$	$\overline{Q}_d$	R
M=2				
ρ=.2	2	4.1	3.7	1.11
m=2	4	3.3	2.8	1.18
s=2	8	2.8	2.5	1.12

Variation with Traffic Intensity

	ρ	$\overline{Q}_i$	$\overline{Q}_d$	R
M=2				
m=2	.7	.72	.58	1.24
k=2	.9	4.1	3.7	1.11
s=2	.98	25.8	24.0	1.08

Variation with Number of Servers

	s	$\overline{Q}_i$	$\overline{Q}_d$	R
M=2				
m=2	2	4.1	3.7	1.11
k=2	10	3.4	2.8	1.21
ρ=.9	30	2.2	1.9	1.13

Table 2 (Continued)

Variation with Number of Sources

	M	$\bar{Q}_i$	$\bar{Q}_d$	R
	1	3.7	3.7	1.00
m=2	2	4.1	3.7	1.11
k=2	5	4.7	3.8	1.24
s=2	10	5.1	3.9	1.31
ρ=.9	20	5.5	4.1	1.34
	100	5.8	4.4	1.31
	∞	5.8	5.8	1.00

Variation with Numbers of Sources, Low Input Process Variance

	M	$\bar{Q}_i$	$\bar{Q}_d$	R
m=15	1	1.9	1.9	1.00
k=2	2	3.1	2.0	1.55
s=2	8	4.9	2.3	2.13
ρ=.9	50	5.7	3.3	1.73
	∞	5.8	5.8	1.00

Table 3. Mean queue length of dependent and independent cases--variation of more than one parameter.

(No. Input Streams) M	(Input Shape) m	(Service Shape) k	(Traffic Intensity) ρ	(No. Servers) s	(Mean Queue Length, Independent) $\bar{Q}_i$	(Mean Queue Length, Dependent) $\bar{Q}_d$	$(\bar{Q}_i/\bar{Q}_d)$ R
2	2	2	.9	2	4.1	3.7	1.11
2	15	4	.9	2	2.2	1.1	2.00
2	15	2	.98	2	19.2	12.5	1.54
8	15	4	.9	10	3.3	.85	3.88

there is any definite predictive relationship between single-parameter and multiparameter variation effects.

Four additional sets of cases, shown in Table 4, illustrate that the composition of the source processes is important in determining queue behavior. The important feature in each of these comparisons is that the variance of the resultant distribution is equal for both of the cases being compared.

In the first two comparisons, unequal numbers of Erlang source processes with differing shape parameters were used to achieve similar resultant interarrival time distributions. Since the effect of dependency is sensitive to the number of source processes, the differing behavior exhibited within each comparison could be expected. In comparisons 3 and 4, however, the number of source processes were the same. Source processes with identical uniform distributions were chosen which, when superposed, yielded interarrival time distributions with means and variances equal to that obtained from the superposition of the indicated Erlang processes. The performances of these resultant processes in queuing systems were different, the difference being more pronounced for sources with high variance. It is thus apparent that knowing the mean and variance of the source processes is insufficient for predicting queue behavior; certainly radical differences in the distribution shapes can have significant effect.

Thus far, only the mean of the queue length has been discussed. The effect that superposition-induced dependence has upon the variance of queue length is also of interest. Table 5 gives three examples of the effect. One may notice that the standard deviation of queue length

Table 4. Mean queue length of dependent and independent cases--comparison of source processes with equal variance in the resultant process.

No. of Inputs (M)	Input Shape	Coef. of Var. of Resultant Process	Mean Queue Length Independent ($\overline{Q}_i$)	Mean Queue Length Dependent ($\overline{Q}_d$)	Ratio $\overline{Q}_i/\overline{Q}_d$
COMPARISON 1; ERLANG SOURCES					
2	$m_1 = 1$; $m_2 = 2$	0.778	4.8	4.8	1.00
8	All m = 15	0.778	4.8	2.3	2.09
COMPARISON 2; ERLANG SOURCES					
2	$m_1 = m_2 = 2$	0.625	4.1	3.7	1.11
4	$m_1 = 7$; $m_2 = m_3 = m_4 = 5$	0.625	4.2	2.5	1.72
COMPARISON 3, ERLANG VS. UNIFORM SOURCES					
2	$m_1 = m_2 = 2$	0.625	4.1	3.7	1.11
2	Identical Uniform	0.625	3.1	1.8	1.71
COMPARISON 4; ERLANG VS. UNIFORM SOURCES					
2	$m_1 = m_2 = 15$	0.354	3.1	2.0	1.55
2	Identical Uniform	0.354	3.0	1.7	1.74

Note: In all cases, service shape (k) = 2, traffic intensity (ρ) = 0.9, and number of servers (s) = 2.

Table 5. Effect of dependency on standard deviation of queue length.

No. of Inputs (M)	Shape of Input (m)	Traffic Intensity (ρ)	Independent s.d.	Independent c.v.	Dependent s.d.	Dependent c.v.
2	2	.90	5.02	123.7	4.87	126.5
2	2	.98	23.65	98.5	18.07	100.6
10	2	.90	7.31	126.3	4.98	122.5

In all cases, there are two servers with Erlang shape parameter 2.

is altered by about the same percentage as is the mean. The coefficient of variation, therefore, is almost equal for the dependent and independent cases.

The Limit Distribution of Merged Processes

The fact that the superposition of many renewal processes results in a process which is close to Poisson is discussed by both Cox [20, 1962, pp. 77-79] and Feller [29, 1966, pp. 355-356]. Cox shows that for homogeneous sources, the limiting distribution of interarrival times is exponential as the number of sources approaches infinity.

This can be demonstrated for the merged Erlang processes by writing expression (3.9) for homogeneous sources and normalizing by setting $\lambda = 1/M$. Then

$$G(t) = 1 - e^{-mt} \left(\sum_{r=0}^{m-1} \frac{(mt)^r}{M^r r!} \right) \left(\sum_{r=0}^{m-1} \left(\frac{m-r}{M}\right) \frac{(mt)^r}{M^r r!} \right)^{M-1}$$

$$= 1 - e^{-mt} \left(1 + O\left(\frac{1}{M}\right)\right) \left(1 + \frac{(m-1)\,t}{M} + O\left(\frac{1}{M^2}\right)\right)^{M-1} ,$$

and $\lim_{M \to \infty} G(t) = 1 - e^{-mt}\, e^{(m-1)t} = 1 - e^{-t}$.

In queuing applications, the rapidity of this convergence can be deceiving. Figures 2 and 3 on pages 24 and 25 illustrated that the distribution of interarrival times resulting from the superposition of 50 identical Erlang processes was almost indistinguishable from an exponential distribution. The coefficients of variation for these two cases also show evidence of the convergence; they are 98.2 for

sources with shape parameter of 2 and 94.2 when the parameter equals 15.

Figures 6 and 7 graphically portray the last two sets of data from Table 2, page 24. As expected, when independent samples are taken from the resultant distributions with large numbers of sources ($M > 20$), the queue behaves as if the input were Poisson. One may also note, however, that the dependency introduced by superposition still affects queue performance. Furthermore, it would appear that the effect of dependency on queue length decays quite gradially.

The Superposition of Interarrival Processes with Small Variances

Figure 4, page 26, leads to the conjecture that, as the variance of interarrival times of two homogeneous source processes becomes small, the distribution of the superposed interarrivals may be uniform. That this is so may be easily demonstrated. Writing (3.9) for M=2 gives, after rearranging and normalizing to obtain a mean of 1:

$$G(t) = 1 - e^{-mt}\left(\sum_{r=0}^{m-1} \frac{(\frac{mt}{2})^r}{r!}\right)\left(\sum_{r=0}^{m-1} (\frac{m-r}{m}) \frac{(\frac{mt}{2})^r}{r!}\right)$$

$$= 1 - e^{-mt}\left(\sum_{r=0}^{m-1} \frac{(\frac{mt}{2})^r}{r!}\right)\left(\sum_{r=0}^{m-1} \frac{(\frac{mt}{2})^r}{r!} - \frac{t}{2}\sum_{r=0}^{m-2} \frac{(\frac{mt}{2})^r}{r!}\right)$$

$$= 1 - [e^{-\frac{mt}{2}}(\sum_{r=0}^{m-1} \frac{(\frac{mt}{2})^2}{r!})][e^{-\frac{mt}{2}}(\sum_{r=0}^{m-1} \frac{(\frac{mt}{2})^r}{r!} - \frac{t}{2}\sum_{r=0}^{m-2} \frac{(\frac{mt}{2})^r}{r!})]$$

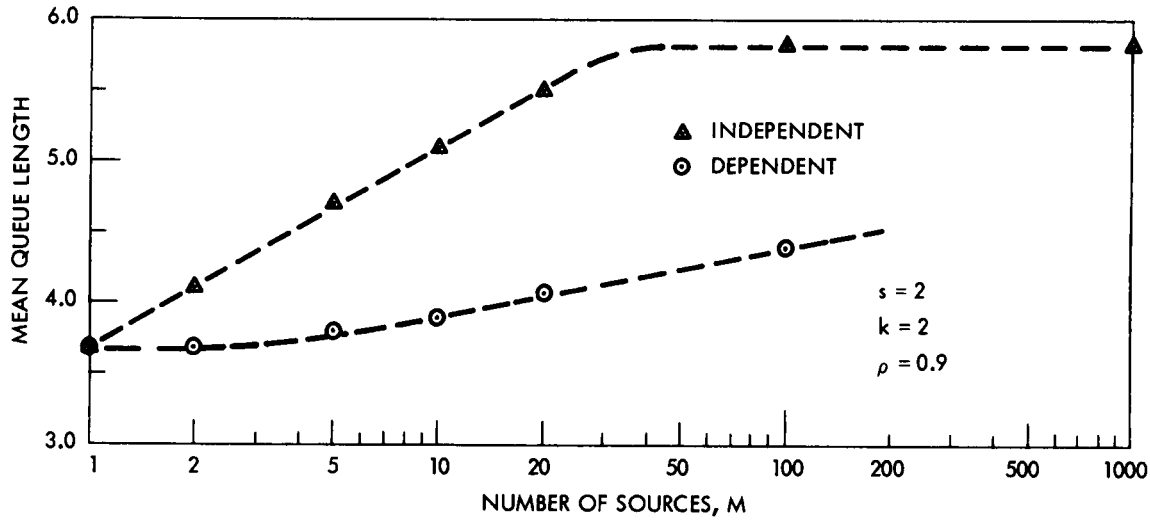


Figure 6. Effect of dependency introduced by superposition of M sources with shape parameter 2.

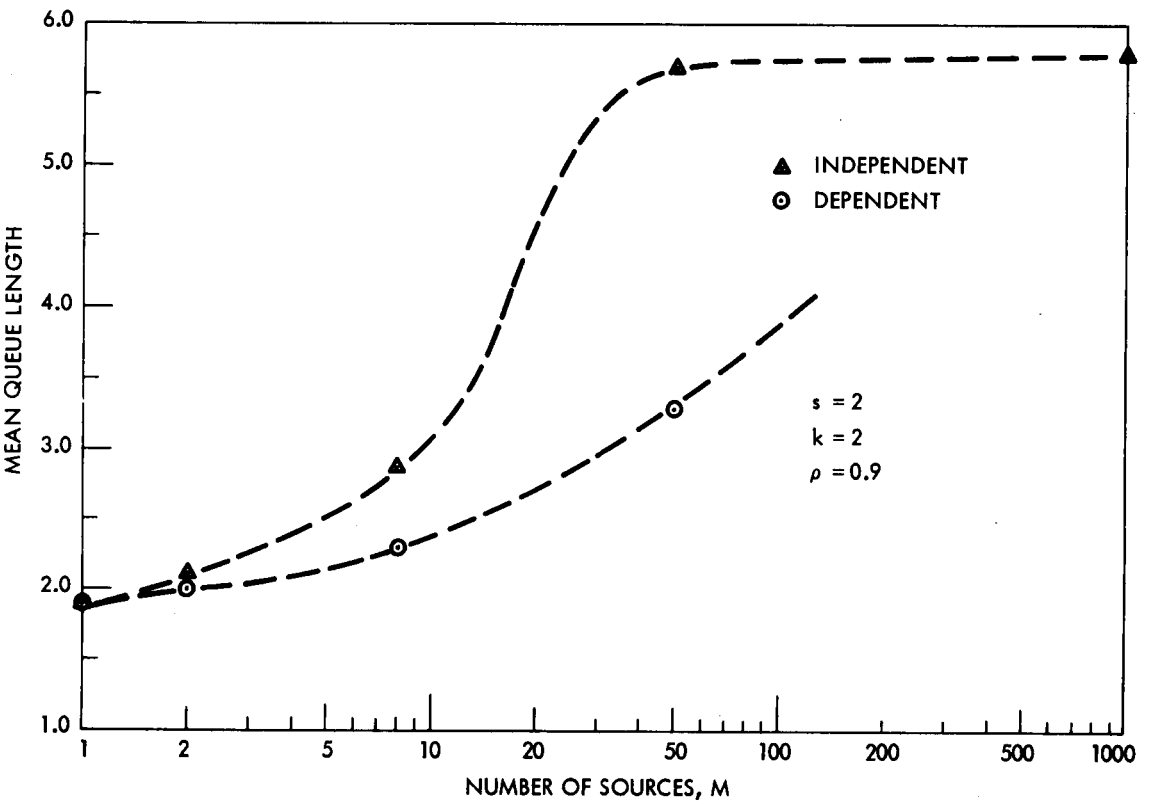


Figure 7. Effect of dependency introduced by superposition of M sources with shape parameter 15.

It is shown in Appendix C, page 115, that

$$\lim_{m\to\infty} e^{-\frac{mt}{2}} \sum_{r=0}^{m-n} \frac{(\frac{mt}{2})^r}{r!} = \begin{cases} 1, & t < 2 \\ 1/2 & = 2 \\ 0, & t > 2 \end{cases}$$

Thus

$$\lim_{m\to\infty} G(t) = \begin{cases} t/2, & 0 \leq t \leq 2 \\ 1, & t > 2 \end{cases}$$

Similarly, for M=3, the triangular distribution emerges as the limiting distribution as m increases:

$$\lim_{m\to\infty} G(t) = \frac{2t}{3} - \frac{t^2}{9}, \quad 0 \leq t \leq 3 \quad .$$

For M=4

$$\lim_{m\to\infty} G(t) = \frac{3t}{4} - \frac{3t^2}{16} + \frac{t^3}{64}, \quad 0 \leq t \leq 4 \quad .$$

The rapidity of these convergenses is shown in Figure 8 which plots the shape parameter of the resultant process against that of the homogeneous input processes for various numbers of inputs.

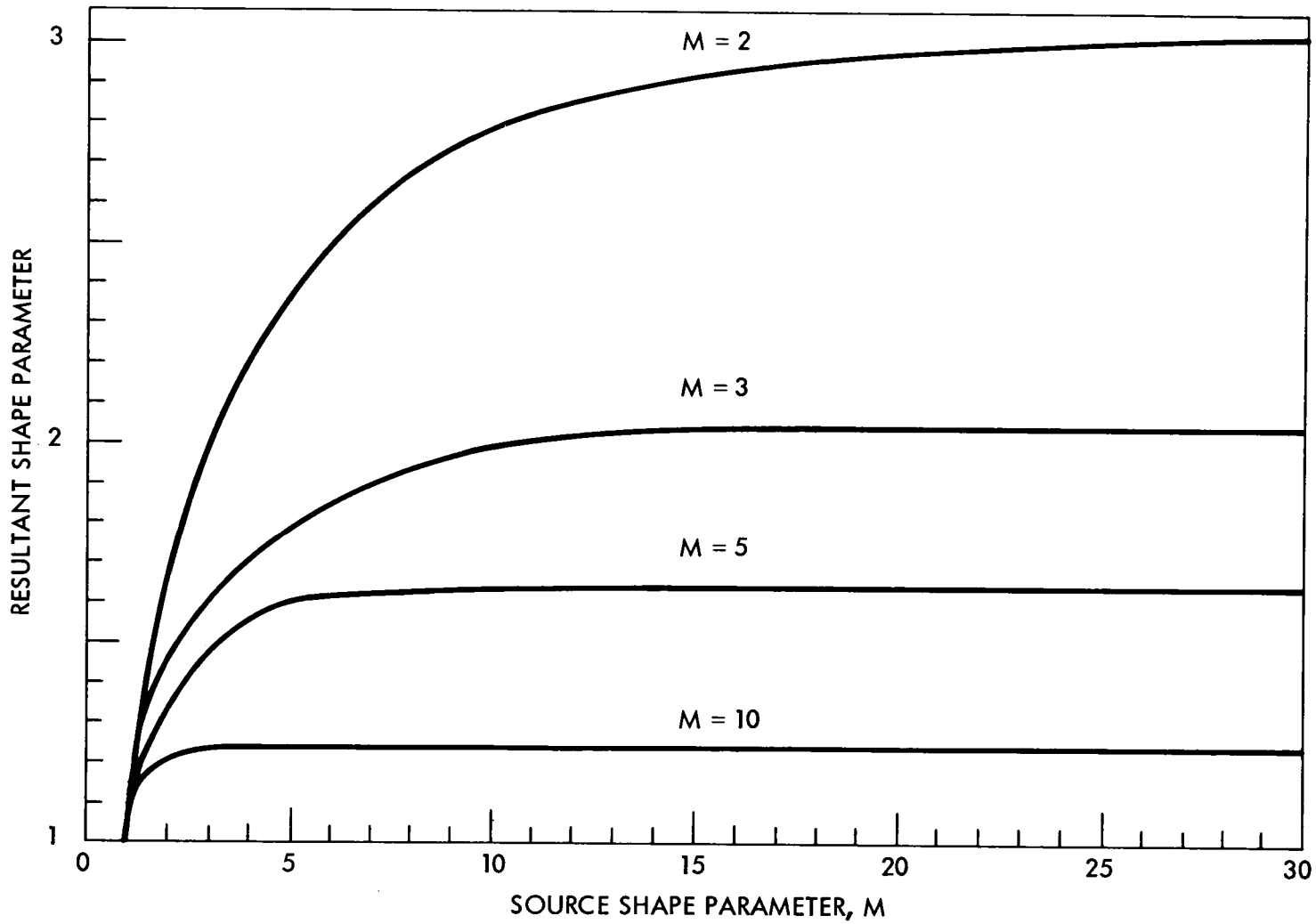


Figure 8. Shape parameter of resultant process versus shape parameter of M sources.

Estimation of Mean Queue Length for Queues with Superposed Input Processes

Figures 6 and 7 on page 40 immediately suggest a method of estimating queue behavior when the input consists of superposed Erlang renewal processes. For low numbers of homogeneous source processes, the system will behave as if the input were a single Erlang renewal process with the appropriate mean interarrival time. Mean queue length may be similarly estimated for nonhomogeneous sources. Consider a system with E_k/s service and M Erlang sources with parameters λ_i and m_i, $i=1,\ldots,M$. Let the total traffic intensity be ρ, $\rho < 1$. First, for each source, the mean queue length, q_i, is found assuming the total input is due to source i. Thus, q_i is the mean queue length of an $Em_i/E_k/s$ system with traffic intensity ρ. The estimate of system mean queue length, $\bar{Q}$, is the weighted average of the q_i, the weight being the proportion of input due to each source, w_i:

$$\bar{Q}(\text{estimate}) = \sum_{i=1}^{M} w_i \, q_i \quad .$$

The following example illustrates the method. Consider a queuing system with two servers, each of which has an Erlang distributed service time with shape parameter 2. The traffic intensity is 0.9 and there are two Erlang source processes with the following parameters:

	Shape Parameter (k)	Mean Interarrival Rate (λ)
Stream 1:	1	4/5
Stream 2:	16	1/5

The mean interarrival rate of the resultant process is 1. From Hillier and Lo, the mean queue lengths of queuing systems with the described service and input processes of shape parameters 1 and 16 are:

Input Shape Parameter	Mean Queue Length
1	5.773
16	1.9058

The estimated mean queue length for the system with merged inputs is the weighted average of the above:

$$\bar{Q}(\text{estimate}) = (\tfrac{4}{5})(5.773) + (\tfrac{1}{5})(1.9058) = 5.000 \quad .$$

The value obtained from simulation is 5.063. As Figures 6 and 7 on page 40 indicate, the actual mean queue length will be slightly higher than that obtained from the estimate.

It is apparent that, as the number of source processes increase, estimates of mean queue length will require correction. Additional experimental results would be required to provide the data necessary to accomplish such corrections over a wide range of queue configurations. Even without correction, however, the above estimating procedure can be used to establish a lower bound on mean queue length. Further, an upper bound is given by the queue's behavior with Poisson input. In practical problems, the difference in these two bounds would indicate whether or not additional information is required.

The Superposition of Processes with Gamma Distributed Interarrival Times

In analyzing networks, it will be necessary to superpose process with gamma distributed interarrival times. The expressions previously derived for integer valued shape parameters do not apply, requiring the following numerical procedure. For M source process, let the distribution of the ith source be $F_i(t)$ with parameters λ_i and m_i. From an arbitrary point in time, the time until the next arrival from source i is given by the residual time distribution, $H_i(t)$. Numerical integration is used:

$$H_i(t) = \lambda_i \int_0^t [1 - F_i(x)]\, dx \quad .$$

Since the sources are independent, the residual time distribution of the resultant process, $H_R(t)$, is

$$H_R(t) = 1 - \prod_{i=1}^{M} [1 - H_i(t)] \quad .$$

The interarrival time distribution of the resultant process, G(t), can be obtained:

$$G(t) = 1 - \frac{1}{\lambda_R} \frac{d}{dt} H_R(t)$$

where λ_R is the resultant arrival rate. Numerical differentiation is avoided by noting that

$$\frac{d}{dt} H_R(t) = - \frac{d}{dt} \prod_{i=1}^{M} [1 - H_i(t)] = - \sum_{i=1}^{M} \frac{Q}{1 - H_i(t)} \frac{d}{dt} [1 - H_i(t)]$$

where

$$Q = \prod_{i=1}^{M} [1 - H_i(t)] \quad .$$

Then

$$\frac{d}{dt} H_R(t) = \sum_{i=1}^{M} \frac{Q \lambda_i [1 - F_i(t)]}{1 - H_i(t)} \quad .$$

Further numerical procedures yield the mean and variance.

Accuracy of the process is observed by comparison with the Erlang results for integer values of the shape parameter. Also, the mean arrival rate obtained from this procedure is compared to the known resultant mean; less than 1% difference is typically achieved by dividing the range of significant values of the function into 1000 intervals. Solution time on an IBM 360/195 averaged about five seconds.

Summary

The distribution of interarrival times resulting from the superposition of heterogeneous Erlang or gamma renewal processes was developed. Further, the dependency among arrival intervals which is introduced by superposition was experimentally examined. It was found that this dependency often has a significant effect on the mean queue length of a subsequent service system. A method of estimating this queue length was illustrated.

CHAPTER IV

INTERDEPARTURE TIMES FOR ERLANG QUEUING SYSTEMS WITH ONE OR TWO SERVERS

The 1956 finding by Burke, cited earlier, that departures from an M/M/s system form a Poisson process provided some early encouragement to the study of networks of queues. It was soon recognized, however, that equivalent results would not be forthcoming for queuing systems with other time distributions. Reich's proof that the output of an $E_2/E_2/1$ system was not Erlangian came in 1957. He further showed that, if the input and output processes of a single channel queue are Poisson, the service time is exponentially distributed.

An additional complication is that, in general, the interdeparture times will not be mutually independent. Disney and others have recently shown that, for single server systems, mutual indpendence of output intervals is achieved only if the system is in steady state and one of the following holds: service time is zero; no queue is allowed; maximum queue length is one and service times are deterministic; or the system is M/M/1 with an infinite queue permitted [25, 1973]. The last of these possibilities is the most important and had been noted earlier (see, for example, Finch [30, 1959]).

In this chapter approximations to the interdeparture time distributions for Erlang systems with one and two servers will be developed. The correlation of successive interdeparture intervals will then be examined using a simulation approach. As with dependency

introduced by superposition, the emphasis will be on the affect of dependency on the behavior of a subsequent queuing system.

The Output of an $E_m/E_k/1$ System

The interdeparture time distribution of single channel queues was developed by Chang using complex-variable theory [14, 1963]. He applied the method to the M/M/1 queue, duplicating the well-known result that the output of such a system is Poisson. Application to more general systems becomes increasingly difficult, however. In this section, an approximation of the interdeparture time distribution for the $E_m/E_k/1$ system will be developed. It will be shown that this approximation yields good results over a wide range of parameters. Furthermore, the approach developed forms the basis for a similar approximation for the two-server system.

The following random variables will be required:

t_a , the interarrival times,

t_s , the service time,

t_{ar}, the residual time until an arrival,

t_d , the interdeparture time, and

y , $t_a - t_s$.

Additionally, define

$a(t_a)$, the Erlang density function of interarrival times,

λ , the arrival rate,

m , the shape parameter of the arrival distribution,

$b(t_s)$, the Erlang density function of service times,

μ , the mean service rate,

k , the shape parameter of the service distribution,

$a_r(t_{ar})$, the residual time distribution of the arrival process,

$a_y(y)$, the density function of y,

$c(t_d)$, the density function of interdeparture times,

P_0 , the probability of the system being empty at the moment following a service completion, and

P_{00} , the probability of the system being in state zero at a regeneration point.

For Poisson arrivals, the density function of interdeparture times is given by Cox [20, 1965, p. 299]:

$$c(t_d) = P_0 \int_0^{t_d} b(t_d - t)\, a(t)\, dt + (1 - P_0)\, b(t_d) \quad . \qquad (4.1)$$

For Erlang distributed interarrivals, the first term of the expression becomes dependent upon prior states of the system. A first order approximation can be written by replacing $a(t)$ in (4.1) with $a_r(t)$. The expression thus obtained was found to give good results for higher values of traffic intensity--above about 0.7. To correct for low values of traffic intensity, the first term of (4.1) was divided into two cases: one in which the previous regeneration state was zero, and one in which it was nonzero. In the former case, the remaining time to an arrival can be seen to be

$$t_a - t_s \mid t_a > t_s \quad .$$

The time to an arrival for the nonzero case was again approximated by the function $a_r(t)$. The corrected expression for interdeparture time thus becomes

$$c(t_d) = P_0\, p_{00} \int_0^{t_d} b(t_d - t)\, a_y(t)\, dt$$

$$+ P_0(1 - p_{00}) \int_0^{t_d} b(t_d - t)\, a_r(t)\, dt$$

$$+ (1 - P_0)\, b(t_d) \quad . \tag{4.2}$$

Now, from Appendix B, page 109,

$$a_r(t) = \lambda\, e^{-\lambda m t} \sum_{i=0}^{m-1} \frac{(\lambda m)^i}{i!}\, t^i \quad . \tag{4.3}$$

Further, $a_y(t)$ can be found by writing the joint density of t_a and t_s and making the transformation

$$y = t_a - t_s$$
$$z = t_a \quad .$$

Thus, in the region $y > 0$, the density of y is

$$a_y(y) = \int_y^{\infty} a(z)\, b(z - y)\, dz \quad . \tag{4.4}$$

Using the method of Appendix D, page 117, (4.4) reduces to a function of the form

$$a_y(y) = e^{-\lambda m y} \sum_{i=0}^{k+m-2} c_i\, y_i \quad . \tag{4.5}$$

The transition probability, p_{00}, is given by the probability that serivce time is less than an arrival interval. Thus

$$p_{00} = \int_{t_a=0}^{\infty} \int_{t_s=0}^{t_a} a(t_a)\, b(t_s)\, dt_s\, dt_a \quad . \tag{4.6}$$

The state probability, P_0, is now found by requiring the expected value of interdeparture times be equal to the expected value of interarrival times. Thus

$$P_0\, p_{00}[E(y|y > 0) + \tfrac{1}{\mu}] + P_0(1 - p_{00})[\tfrac{m+1}{2\lambda m} + \tfrac{1}{\mu}]$$

$$+ (1 - P_0)\, \tfrac{1}{\mu} = \tfrac{1}{\lambda} \quad . \tag{4.7}$$

The only remaining task is the evaluation of the two convolution integrals in (4.2). Noting the similarity in functional form of the densities in (4.3) and (4.5), the same development is applicable to both convolutions. Consider

$$\int_0^t b(t - u)\, a_r(u)\, du \quad .$$

The Laplace transforms of $a_r(t)$ and $b(t)$ are, respectively,

$$L_1(s) = \lambda \sum_{r=1}^{m} \frac{(\lambda m)^{r-1}}{(s + \lambda m)^r}$$

$$L_2(s) = \frac{(\mu k)^k}{(s + \mu k)^k} \quad .$$

The transform of the convolution is thus

$$L_T(s) = L_1(s)\, L_2(s) = \lambda(\mu k)^k \sum_{r=1}^{m} \frac{(\lambda m)^{r-1}}{(s + \lambda m)^r (s + \mu k)^k} \quad . \tag{4.8}$$

If $\mu k \neq \lambda m$, partial fraction expansion of (4.8) gives

$$L_T(s) = \lambda(\mu k)^k \sum_{r=1}^{m} (\lambda m)^{r-1} \left(\sum_{i=1}^{r} \frac{K_{1i}}{(s + \lambda m)^i} + \sum_{i=1}^{k} \frac{K_{2i}}{(s + \mu k)^i} \right), \tag{4.9}$$

where

$$K_{1i} = \frac{(-1)^{r-i} \binom{r + k - i - 1}{k - 1}}{(\mu k - \lambda m)^{r+k-i}} \quad , \text{ and}$$

$$K_{2i} = \frac{(-1)^{k-i} \binom{r + k - i - 1}{r - 1}}{(\lambda m - \mu k)^{r+k-i}} \quad .$$

Finally,

$$\int_0^t b(t-u)\, a_r(u)\, du =$$

$$= \lambda(\mu k)^k \sum_{r=1}^{m} (\lambda m)^{r-1} \left(\sum_{i=1}^{r} \frac{K_{1i}\, t^{i-1}\, e^{-\lambda m t}}{(i-1)!} + \sum_{i=1}^{k} \frac{K_{2i}\, t^{i-1}\, e^{-\mu k t}}{(i-1)!} \right). \tag{4.10}$$

For the case $\mu k = \lambda m$, (4.10) is

$$\lambda \sum_{r=1}^{m} \frac{(\lambda m)^{k+r-1}\, e^{-\lambda m t}}{(r+k-1)!} \quad . \tag{4.11}$$

With a similar evaluation of the first convolution integral, the interdeparture time density function for $E_m/E_k/1$ queues is completed.

Figures 9 and 10 illustrate the interdeparture time density function for a variety of parameters. As one would expect, the service shape parameter is most influential on the outcome when traffic intensity is high and the arrival shape is important for low congestion. This is further illustrated in Figure 11 which plots the coefficient of variation of interdeparture times against the service shape. A comparison of these functions to Erlang distributions is shown in Figure 12; cases were chosen for illustration which contrast the degree of resemblance.

Traffic intensity (ρ) has an important effect on the interdeparture time distribution for most choices of parameters. Figure 12 plots the coefficient of variation of the output distribution versus the traffic intensity for several cases. The variance in interdeparture time distribution is apparent for the extreme values of traffic

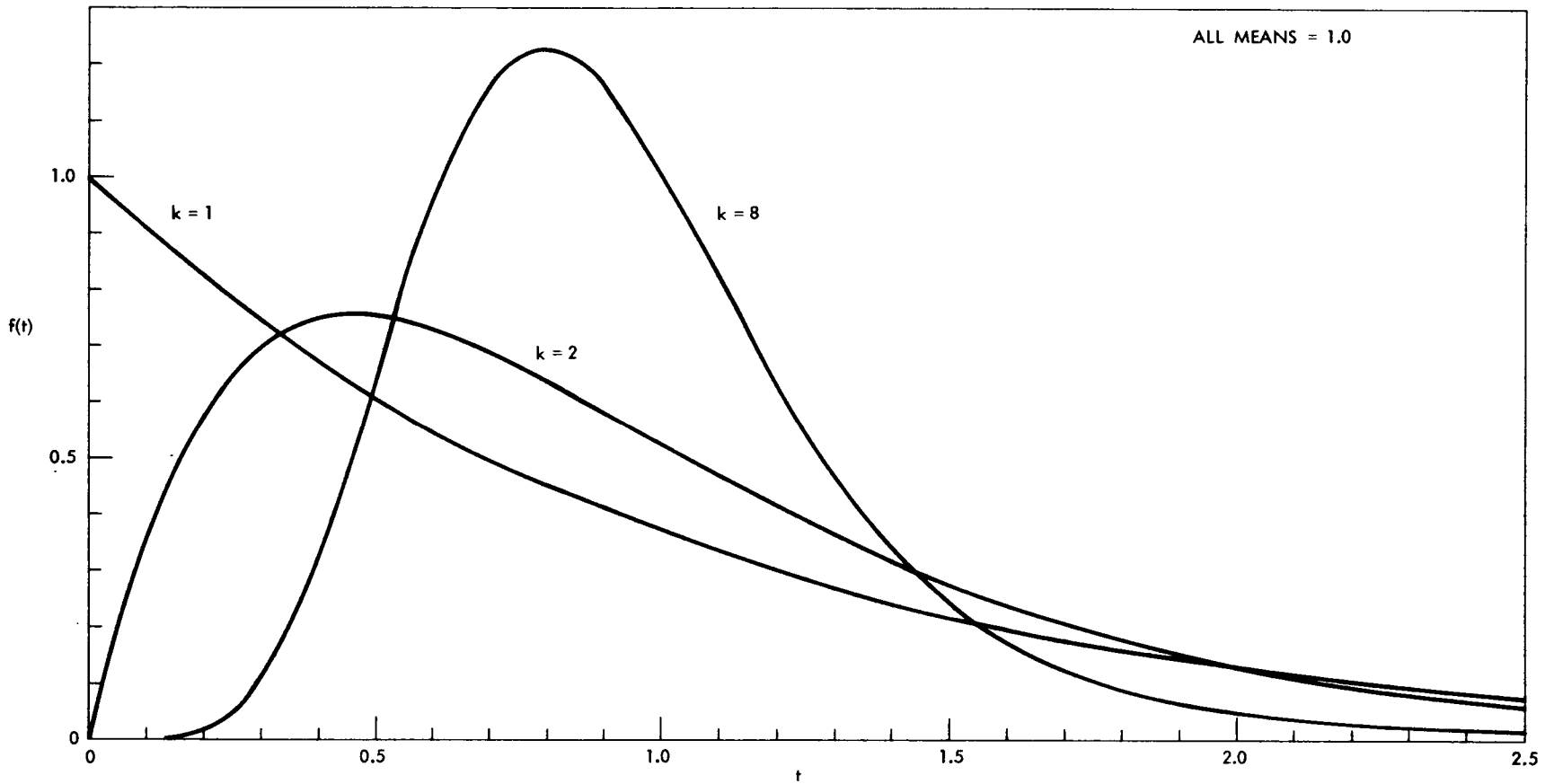


Figure 9. Density functions of interdeparture times of $M/E_k/1$ queues with traffic intensity of 0.9.

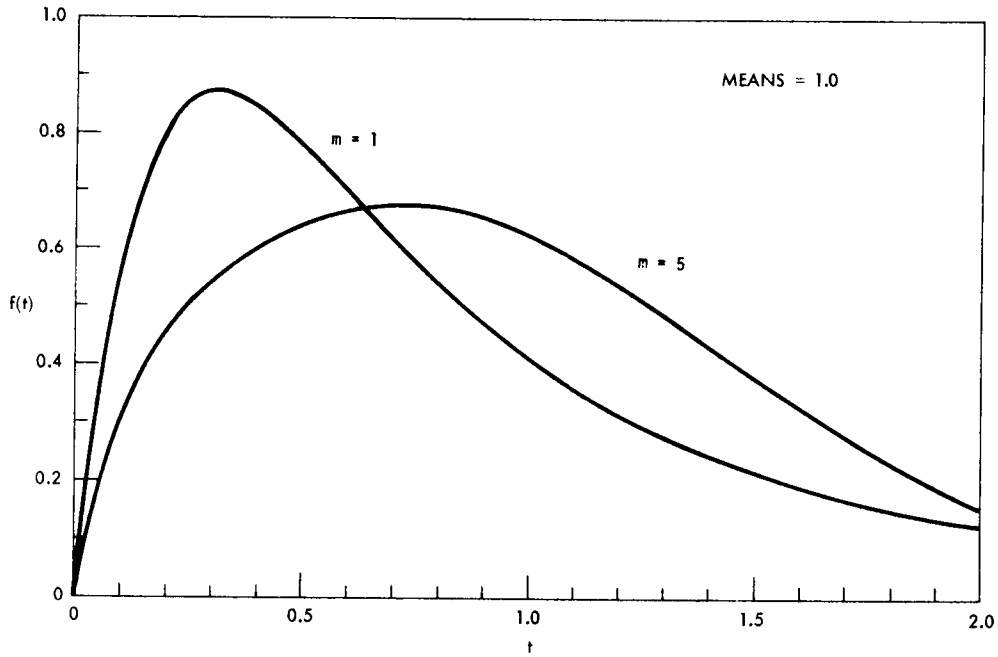


Figure 10. Density function of interdeparture times of $E_m/E_2/1$ systems with traffic intensity of 0.5.

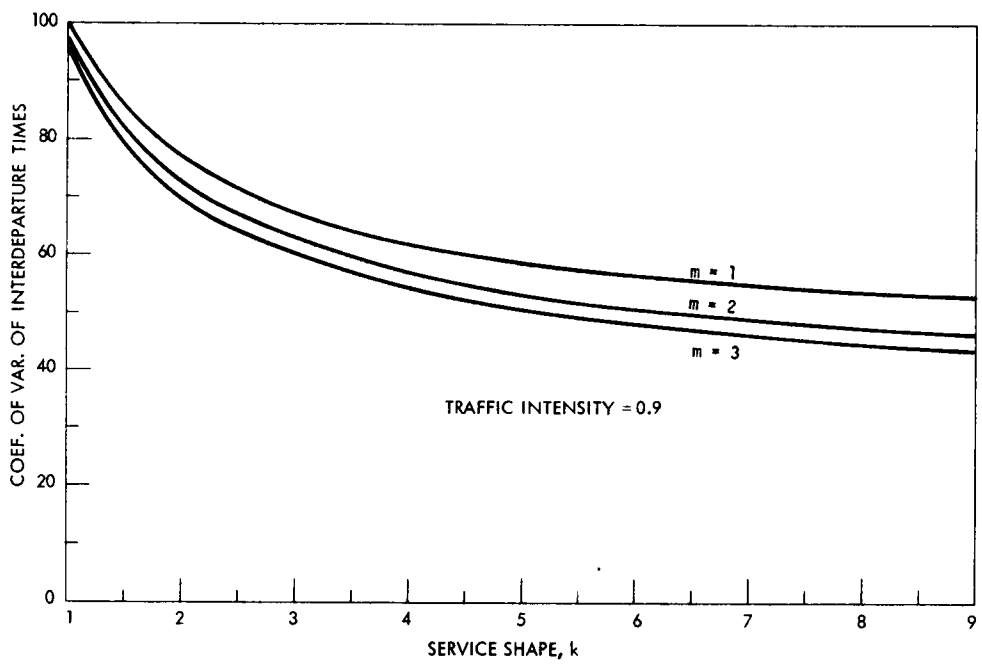


Figure 11. Coefficient of variation of interdeparture times of $E_m/E_k/1$ queues.

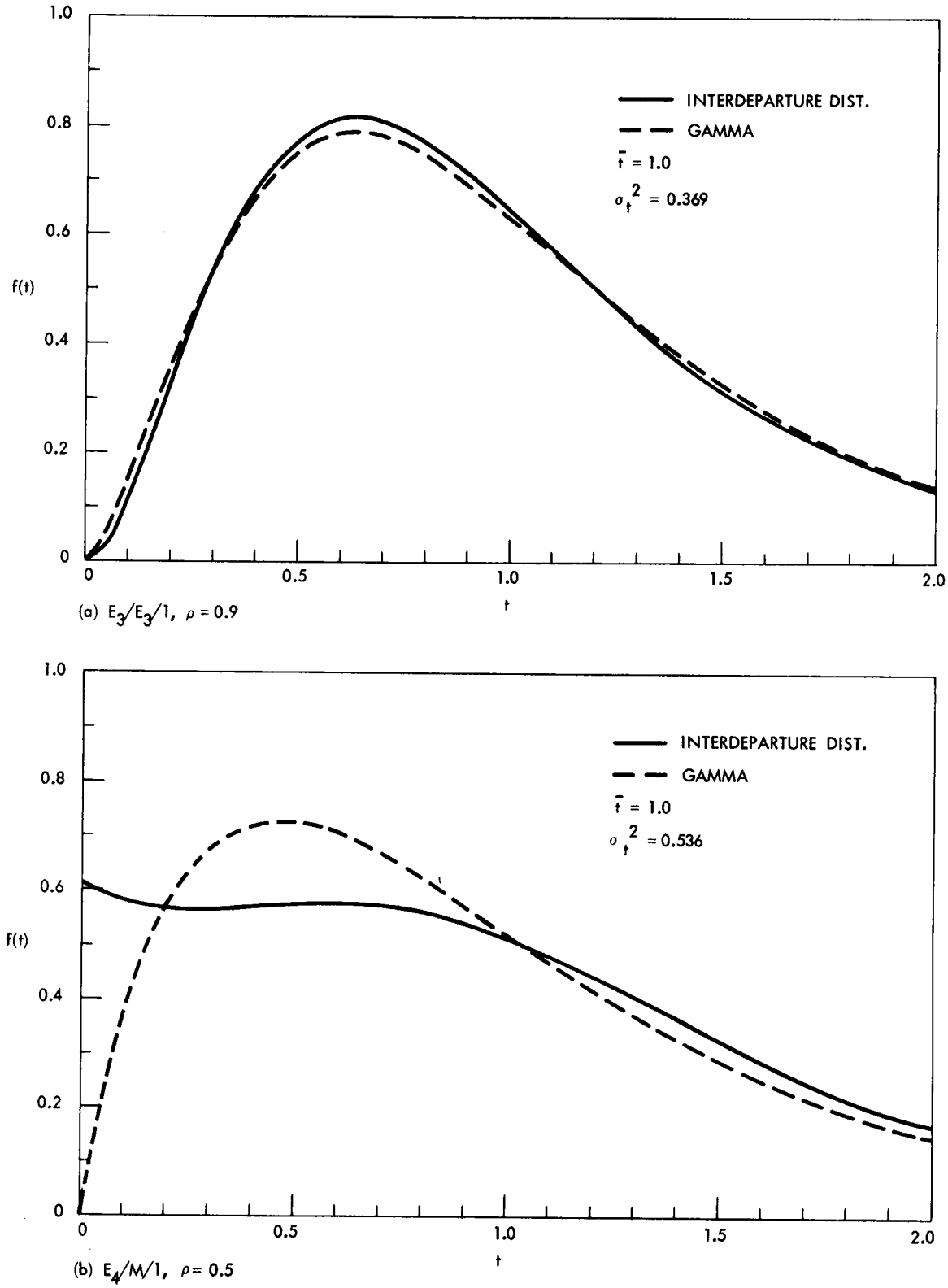


Figure 12. Comparison of density functions of interdeparture times to gamma densities of similar means and variances, single-server queues.

intensity. At $\rho = 0$, the output of the system will be identical to the input. At $\rho = 1$, equivalent to a system in which the server is never idle, the output will be identical to the service distribution. As Figure 13 illustrates, the function developed in this section gives results consistent with these boundary conditions. Further evidence of the validity of the function is indicated by comparison to simulations which were performed for various intermediate values of traffic intensity. These simulation results are also shown in Figure 13.

The Output of an $E_m/E_k/2$ System

The development of an approximation to the interdeparture time distribution for the two-channel queue requires notation in addition to that introduced in the preceding section. Define

- P_1 , the probability of being in state 1 at the regeneration point,
- P_2 , the probability of being in state 2 or greater at the regeneration point,
- t_{sr} , a random variable distributed by the residual time distribution of service time,
- t_1 , a random variable representing the time from a regeneration point to completion of service in one channel (arbitrarily designated Channel 1),
- t_2 , a similar random variable for the other channel, and
- t_{min_i}, the minimum of t_1 and t_2, given the system is in state i at the regeneration point.

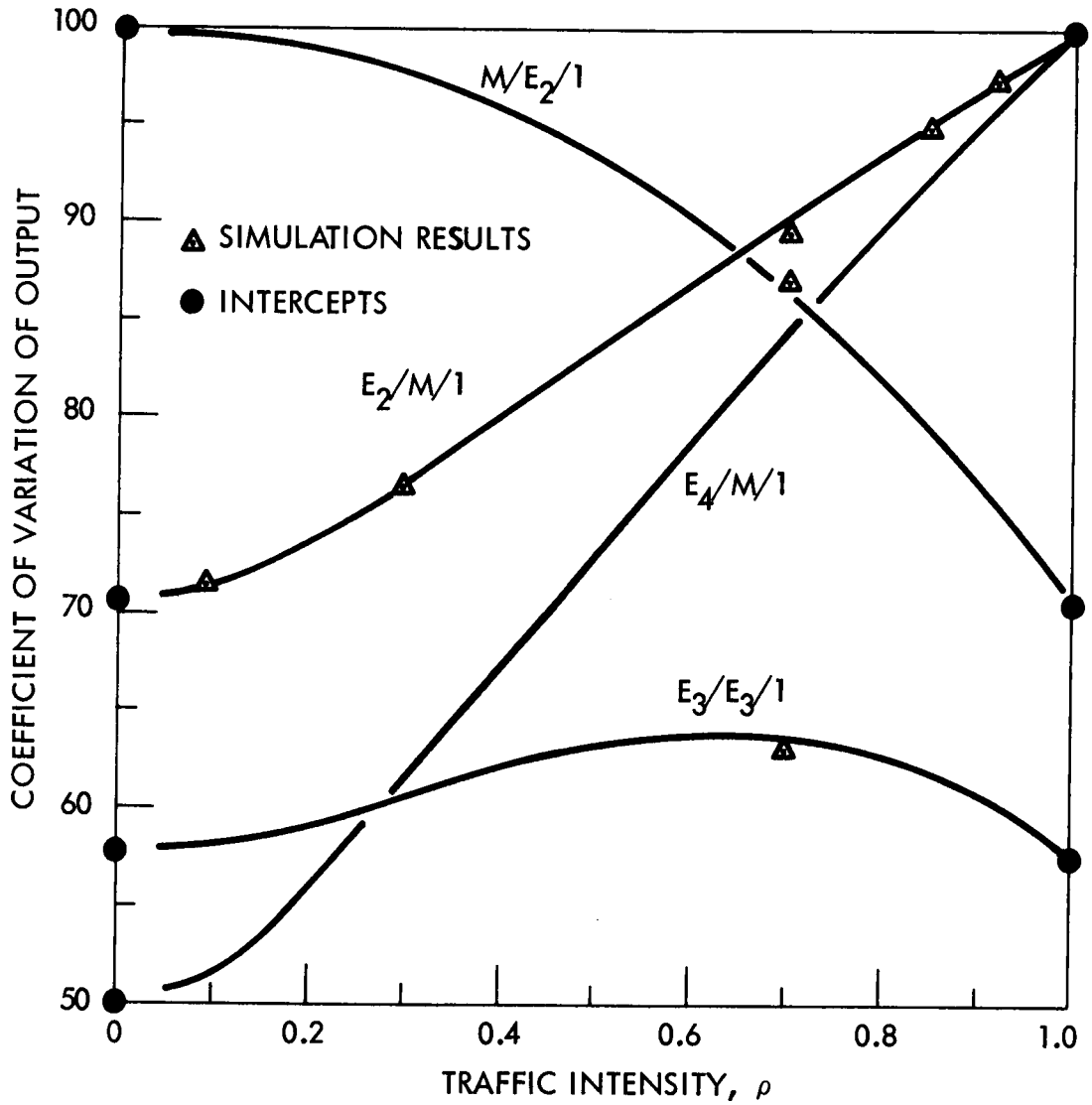


Figure 13. Coefficient of variation of interdeparture times versus traffic intensity for single-server queues.

As was the case with the single-server queue, a function based only upon the present state of the system at a regeneration point will provide a first estimate of the true interdeparture time distribution. This can be written as

$$C(t) = \sum_{i=0}^{2} P_i \Pr(t_{min_i} < t) \quad . \tag{4.12}$$

Tests with this function, as with the single-server system, show good results for high values of traffic intensity. The obvious reason is that prior states of the system do not affect the interdeparture time when a queue exists--and a queue is increasingly likely when the traffic intensity is high.

Incorporation of a correction for low traffic intensity again involves dividing the P_0 term in (4.12) into two cases, one for a prior state of zero and one for a nonzero prior state. Let $t_{min_{00}}$ and $t_{min_{10}}$ represent the minimum of t_1 and t_2 for these two cases. Then

$$C(t) = P_0 \, p_{00} \Pr(t_{min_{00}} < t) + P_0(1 - p_{00}) \Pr(t_{min_{10}} < t)$$

$$+ P_1 \Pr(t_{min_1} < t) + P_2 \Pr(t_{min_2} < t) \quad . \tag{4.13}$$

The state probabilities, P_i, are developed in Appendix D, pages 120 through 124. Again, they are chosen to yield a mean departure rate equal to the mean arrival rate. The additional equation required is based on the transition probabilities among states.

Determining C(t) now reduces to the evaluation of the conditional probabilities, $Pr(t_{min_i} < t)$. The two cases for i = 0 will now be examined.

Consider a system at the regeneration point with both channels empty and no customers in the queue. One customer will be partially through the arrival process. Let t_x be the time until this next arrival.

For the case of a prior system state of 1, t_x will be approximated by the residual time distribution of an Erlang process, earlier designated t_{ar}. The density function is given in expression (4.3), page 50. For a prior state of zero, t_x will be the random variable y of the preceding section with the density function of expression (4.5). Since the two density functions are similar in form, the developments of the densities of $t_{min_{00}}$ and $t_{min_{10}}$ proceed similarly. The case for $t_{min_{10}}$ follows:

This first arrival following regeneration will immediately enter service in Channel 1. Let t_{s_1} be the service time. Thus, the elapsed time to a departure from Channel 1, t_1, will be

$$t_1 = t_{ar} + t_{s_1} \quad .$$

Let t_a be the interarrival time between the first and second customers following the regneration point. The time until the second customer's completion of service is thus

$$t_2 = t_{ar} + t_a + t_{s_2} \quad .$$

The distribution of the random variable

$$t_{min_{10}} = \min(t_1, t_2)$$

cannot be found directly since t_1 and t_2 are not independent. However, one may write

$$Pr(t_{min_{10}} < \tau) = 1 - Pr(t_{min_{10}} > \tau) =$$

$$= 1 - [Pr(\tau < t_1 < t_2) + Pr(\tau < t_2 < t_1)] \quad (4.14)$$

Now define new random variables

$$y_1 = \tau - t_{ar}$$

$$y_2 = t_1 - t_{ar} = t_{s_1}$$

$$y_3 = t_2 - t_{ar} = t_a + t_{s_2} \quad .$$

Note that y_1, y_2, and y_3 are mutually independent. The probability density function of y_1 is obtained by a change of variable:

$$f_1(y_1) = \lambda e^{-\lambda m(\tau - y_1)} \sum_{r=0}^{m-1} \frac{(\lambda m)^r (\tau - y_1)^r}{r!} |-1| =$$

$$= \lambda e^{-\lambda a \tau} e^{\lambda a y_1} \sum_{r=0}^{m-1} \frac{(\lambda m)^r}{r!} \sum_{i=0}^{r} (-1)^i \binom{r}{i} y_1^i \tau^{r-i}$$

$$- \infty < y_1 < \tau \quad . \qquad (4.15)$$

The density function of y_2, $f_2(y_2)$, is obviously the same as for the service time. Finally, by use of Laplace transforms, the density function of y_3 is (for $\mu k \neq \lambda m$)

$$f_3(y_3) = (\mu k)^k (\lambda m)^m \left\{ e^{-\lambda m y_3} \sum_{i=0}^{m-1} \frac{(-1)^{m-1-i} \binom{m+k-i-2}{m-1-i} y_3^i}{i!(\mu k - \lambda a)^{m+k-i-1}} \right.$$

$$\left. + e^{-\mu k y_3} \sum_{i=0}^{k-1} \frac{(-1)^{k-1-i} \binom{m+k-i-2}{k-i-1} y_3^i}{i!(\lambda a - \mu k)^{m+k-i-1}} \right\} ,$$

$$0 < y_3 < \infty \quad . \tag{4.16}$$

Expression (4.14) can now be written

$$Pr(t_{min_{10}} < \tau) = 1 - [Pr(y_1 < y_2 < y_3) + Pr(y_1 < y_3 < y_2)] .$$

Now let $f_T(y_1, y_2, y_3)$ be the joint probability density function of y_1, y_2, and y_3. Then

$$f_T(y_1, y_2, y_3) = f_1(y_1)\, f_2(y_2)\, f_3(y_3) \quad .$$

Therefore

$$Pr(y_1 < y_2 < y_3) = \int_{y_1=0}^{\tau} \int_{y_2=y_1}^{\infty} \int_{y_3=y_2}^{\infty} f_T(y_1, y_2, y_3)\, dy_3\, dy_2\, dy_1$$

$$+ \int_{y_1=-\infty}^{0} \int_{y_2=0}^{\infty} \int_{y_3=y_2}^{\infty} f_T(y_1, y_2, y_3)\, dy_3\, dy_2\, dy_1 \quad . \tag{4.17}$$

A similar expression, with the order of integration on y_2 and y_3 interchanged, gives the $Pr(y_1 < y_3 < y_2)$. Evaluation of these expressions is obtained by the procedure described in Appendix D, pages 117 through 120. This completes the distribution of $t_{min_{10}}$.

The distributions of t_{min_1} and t_{min_2} can be obtained more directly. For t_{min_1}, define channel 1 as the channel occupied at the regeneration point. Thus

$$t_1 = t_{sr},$$

and

$$t_2 = t_{ar} + t_s \quad .$$

Independence of all service and interarrival times permits writing

$$Pr(t_{min_1} < \tau) = 1 - Pr(t_1 > \tau)\ Pr(t_2 > \tau) \quad . \tag{4.18}$$

The distribution of t_1 is the residual time distribution of the service process and the distribution of t_2 is given by expression (4.10), page 53.

Finally, the distribution of t_{min_2} is obtained from

$$Pr(t_{min_2} < \tau) = 1 - Pr(t_{sr} > \tau)\ Pr(t_s > \tau) \quad , \tag{4.19}$$

which completes the terms for the interdeparture time distribution in (4.13). Characteristics of the distribution are illustrated in the following figures.

Figure 14 gives plots of the density function at a traffic intensity of 0.9; the importance of the service time shape parameter when congestion is high is evident. Figure 15 shows the effect of the interarrival variance for low traffic intensity. A contrast in the output distributions for one- and two-server systems is shown in Figure 16, and in Figure 17 the output distribution is compared to an Erlang distribution with similar mean and variance.

Finally, Figure 18 plots the coefficient of variation of the output versus traffic intensity for various cases. As for the single-server queue, the adequacy of the approximation at extreme values of congestion is shown to be excellent. Simulation results for intermediate values of traffic intensity indicate some error in the output distribution. This is due to the fact that no correction for prior states was made in the distribution for the term involving P_1, the probability of a single customer in the system at the regeneration point. The correction to this term is more difficult than that for the P_0 term. Since P_1 remains relatively small over all ranges of traffic intensity, this amendment to the function was not undertaken.

The Effects of Dependency in Interdeparture Times

The results of Disney and others, showing that the interdeparture times from queuing systems are generally not independent, was cited earlier. In this section the effects of this dependency will be examined. As was the case for dependency introduced by superposition, the behavior of a subsequent queuing system was used as a measure of this effect.

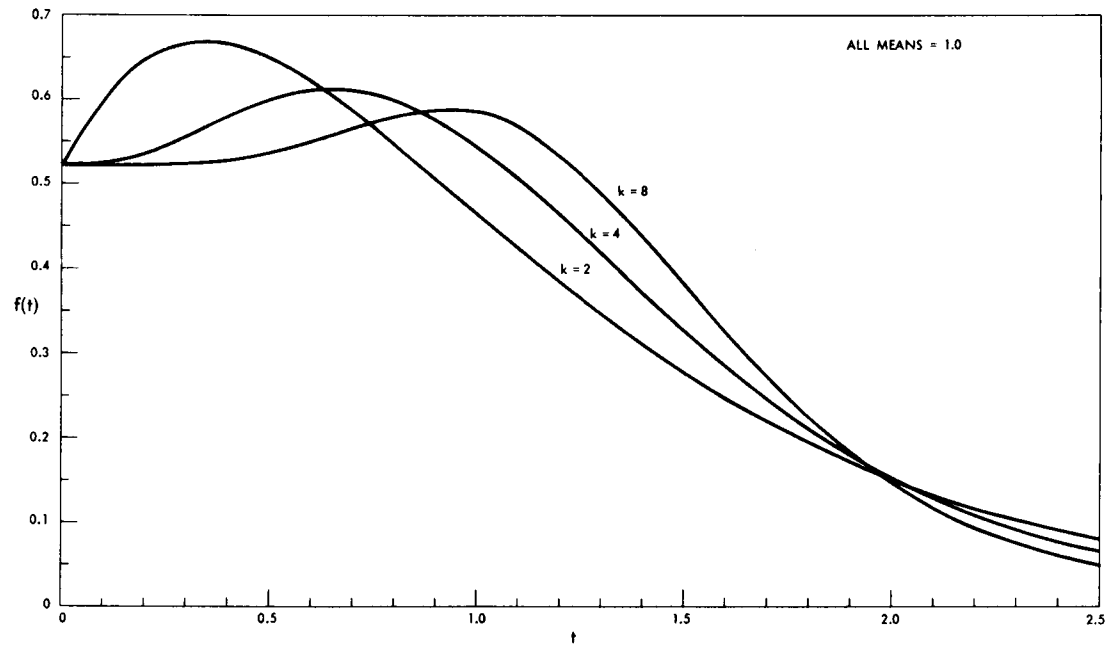


Figure 14. Density functions of interdeparture times from $E_2/E_k/2$ queues with traffic intensity 0.9.

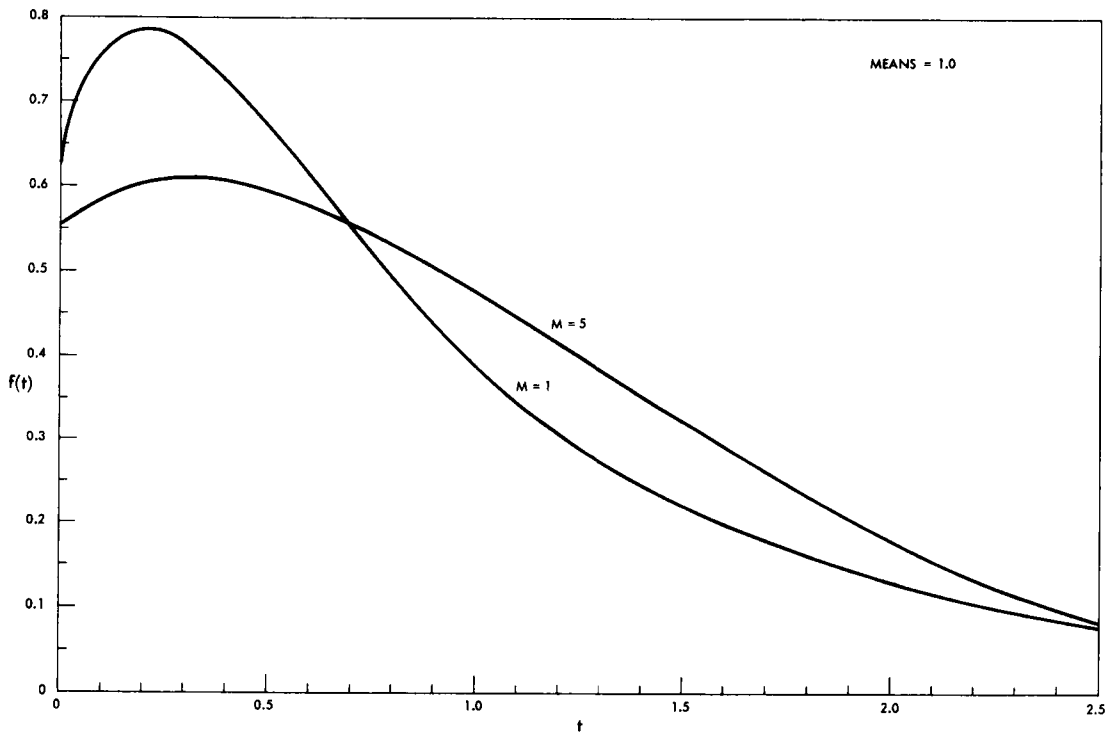
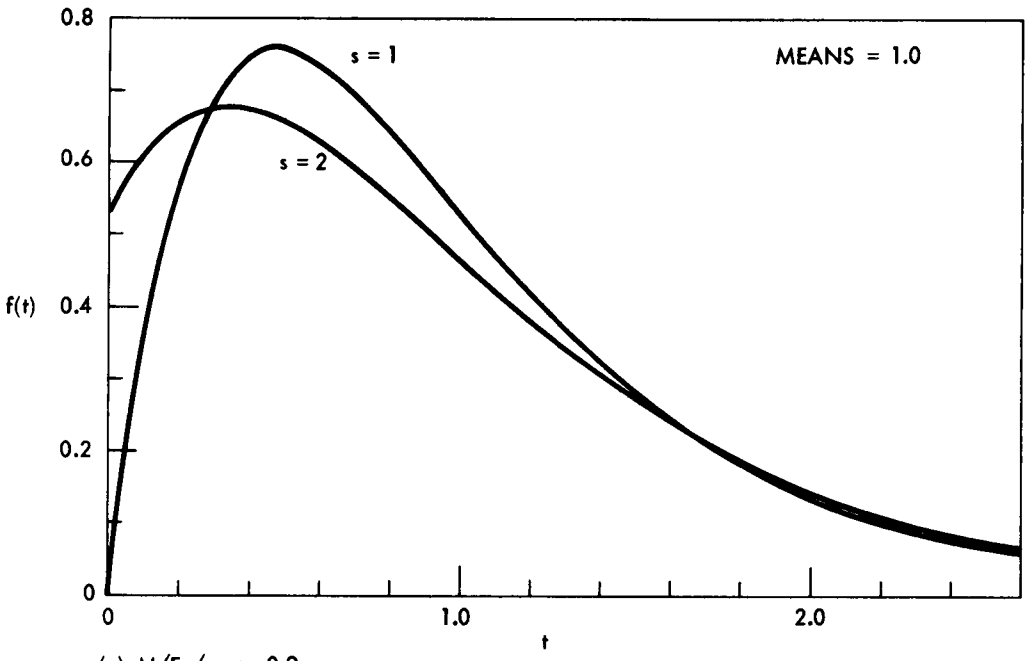
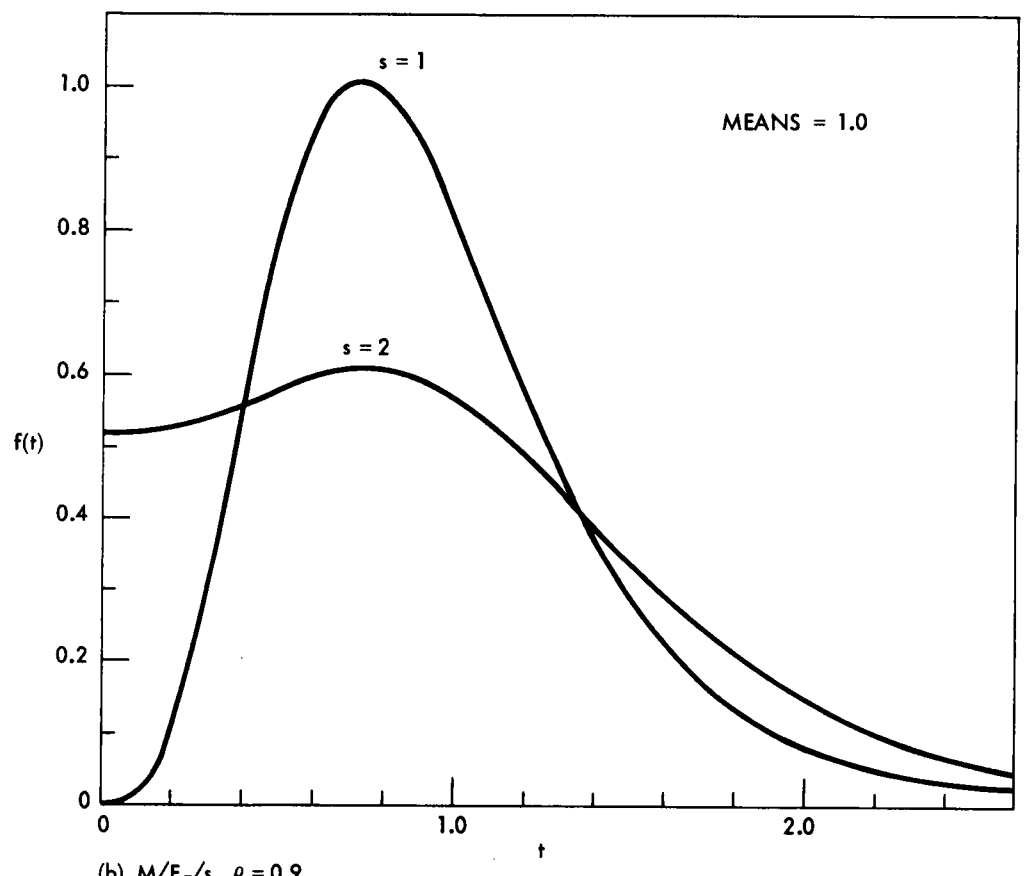


Figure 15. Density functions of interdeparture times from $E_m/E_2/2$ queues with traffic intensity 0.5.

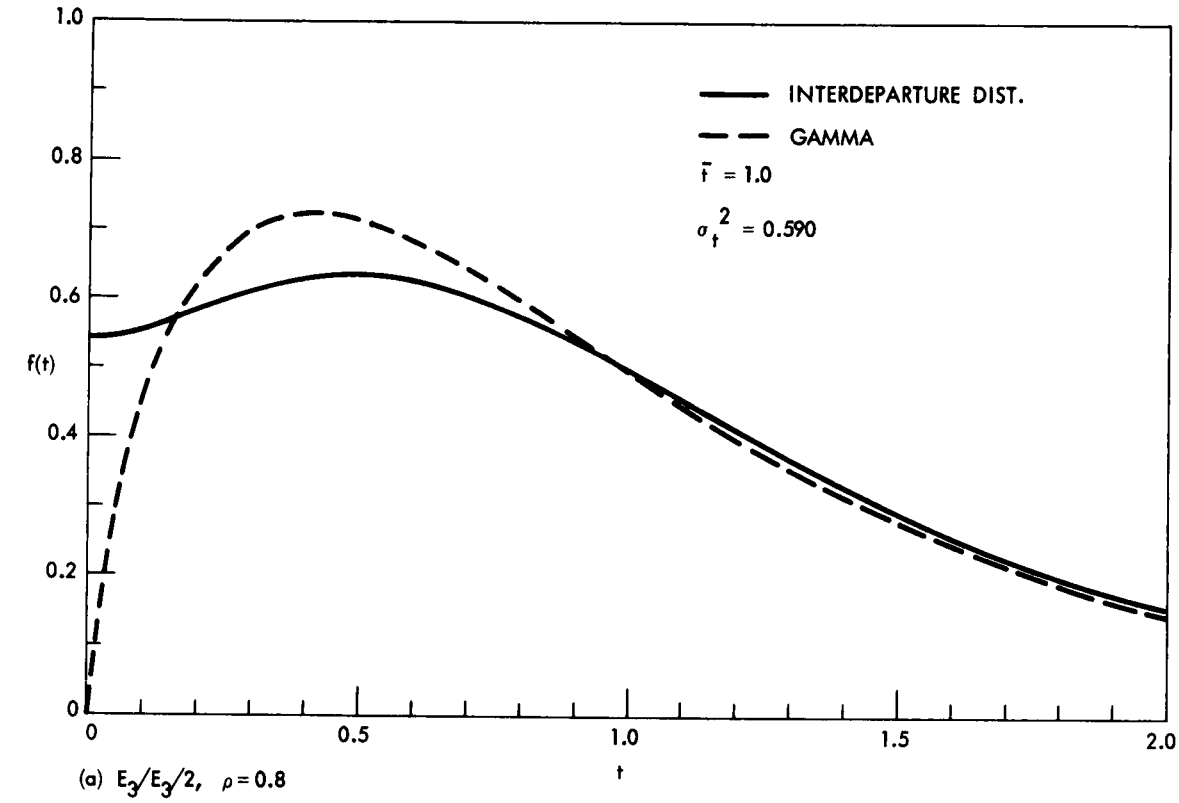


(a) $M/E_2/s$, $\rho = 0.9$

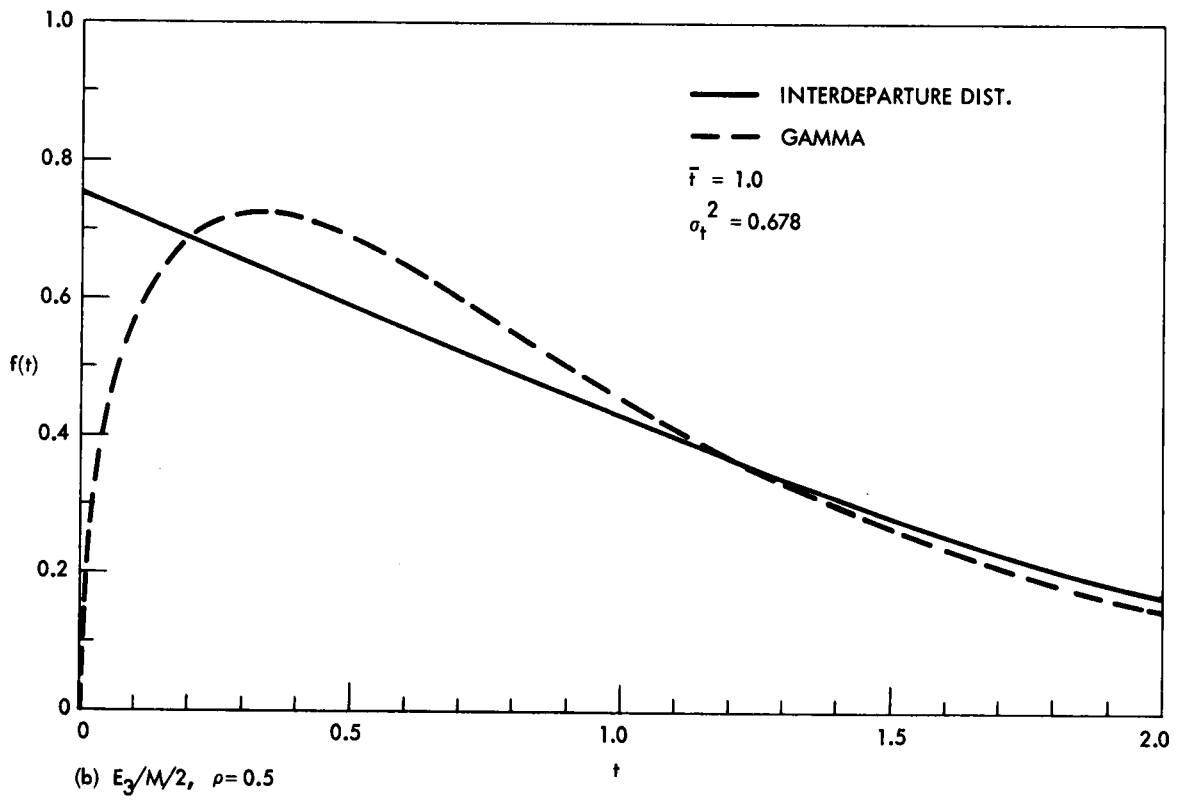


(b) $M/E_5/s$, $\rho = 0.9$

Figure 16. Comparison of interdeparture time density functions of one- and two-server queues.



(a) $E_3/E_3/2$, $\rho = 0.8$



(b) $E_3/M/2$, $\rho = 0.5$

Figure 17. Comparison of interdeparture time and gamma density functions with similar means and variances for two-server queues.

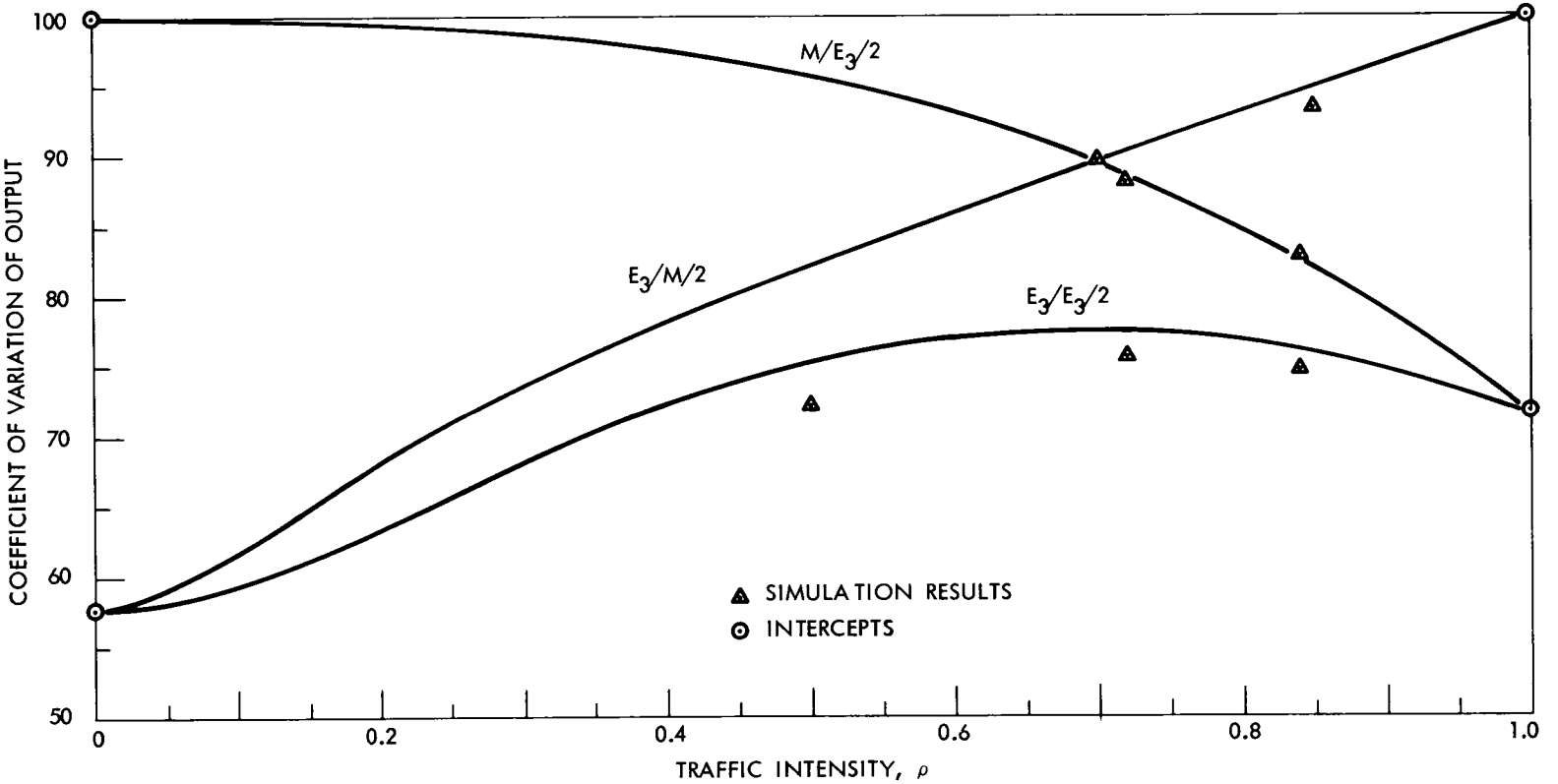


Figure 18. Coefficient of variation of interdeparture times versus traffic intensity for two server queues.

The two systems shown in Figure 19 were simulated. In the dependent case, the $E_m/E_k/s$ system whose output dependency was to be tested was first simulated. The output process of this system was then used as input to a "target" queue. The mean queue length of the target queue, $\bar{Q}_d$, and the standard deviation of queue length, s_{qd}, were determined.

These results were compared to those of a case using independent interarrivals as the input process to an identical target system. The interarrival times were drawn from the interdeparture time distribution of the test queue. The two cases thus differ only by the dependence or independence of the interarrival times presented to the target queue. The mean and standard deviation of queue length, $\bar{Q}_i$ and s_{qi}, were determined for the independent case.

Experiments were chosen to represent reasonably wide variations in test queue parameters: interarrival time variance, service time variance, and traffic intensity. The high cost of these simulations precluded more extensive tests. The experiments typically required about six minutes of CPU time on an IBM/360 model 195. Accuracy of the results thus achieved is discussed in Appendix E; mean queue lengths are valid to within about 7%.

The results of these tests are given in Table 6. It is apparent from the ratio of mean queue lengths for the two cases that the dependency in the output process has relatively little effect on the target queue behavior. It should be noted that the first experiment of Table 6 was repeated with a four-server target queue with service shape parameter 5, and the ratio of mean queue lengths was not substantially

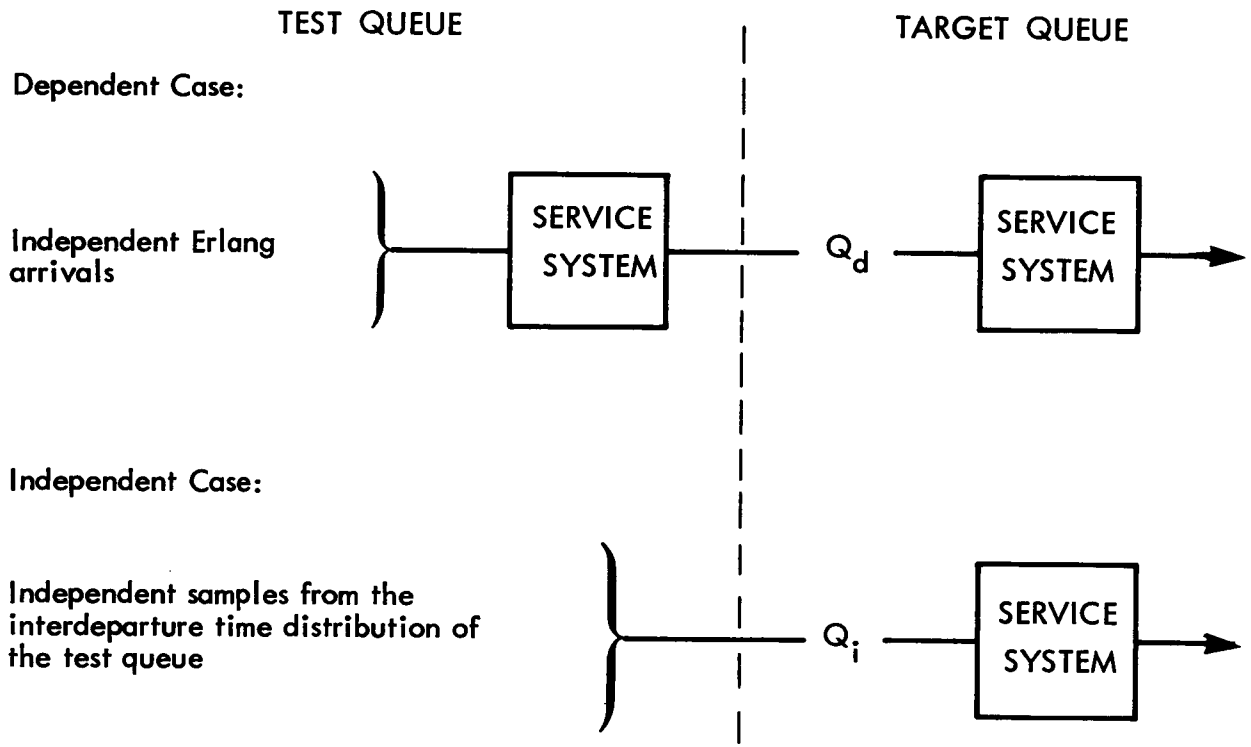


Figure 19. Diagram of simulations used to test dependency among interdeparture times.

Table 6. Effect of dependency in interdeparture times, simulation results.

Test Queue Parameters				Equivalent Output Shape Param.	Independent Case* Target Queue Statistics			Dependent Case* Target Queue Statistics			Ratio
m	k	s	ρ		$\bar{Q}_i$	s_{q_i}	CV_i	$\bar{Q}_d$	s_{q_d}	CV_d	$\bar{Q}_i/\bar{Q}_d$
1	2	2	.9	1.41	6.67	7.58	114	6.77	8.29	122	.99
4	2	2	.9	1.56	6.40	7.04	110	5.41	5.82	108	1.18
1	5	2	.9	1.82	6.05	7.42	116	6.05	7.26	120	1.00
1	2	5	.9	1.19	7.35	7.98	118	6.64	7.95	120	1.11
1	2	1	.9	1.68	6.35	7.05	115	6.85	8.16	119	.93
1	2	2	.98	1.56	6.40	8.23	118	6.62	7.79	118	.97
1	2	2	.7	1.19	7.35	9.01	120	7.98	9.71	122	.92
1	10	1	.98	7.38	4.06	4.83	119	4.66	5.69	122	.87

*For all experiments the target queue was single server with exponential service time $\rho = .9$.

different. Although not conclusive, this indicates that the results are not particularly sensitive to the composition of the target queue. One may also note that the variance of queue length is little affected by the dependency.

Cox has shown that large correlations in successive interdeparture times occur if the coefficient of variation of the service time is small and traffic intensity is near 1 [21, 1965]. The last experiment in Table 6 is such a case. Even for this case, subsequent queue behavior is not radically altered.

Summary

Functions which approximate the interdeparture time distribution of one- and two-server Erlang queuing systems were derived. The dependency among interdeparture intervals was explored and was generally found to be of limited importance to the performance of a subsequent queue.

CHAPTER V

SOME SECONDARY EFFECTS OF DEPENDENCY IN ARRIVAL AND DEPARTURE PROCESSES

In Chapter III the superposition of several renewal processes was examined with considerable attention focused on the effect of the dependency in the resultant interarrival times on the performance of a queuing system. In Chapter IV, the dependency inherent in the interdeparture times of a queuing system with an independent input process was similarly considered.

It is apparent, however, that in complex networks the above assumptions about the independence of the sources will not hold. In this chapter, the affect of relaxing these assumptions will be briefly examined. First, the superposition of two processes which do not have independent interarrival times will be considered. This will be accomplished by superposing the output of two queuing systems. A comparison to a similar case with independent source processes will be made.

The final section will concern the dependency in the output process of a queue when the interarrival times to that queue are not mutually independent.

Superposition of Sources with Dependent Interarrival Times

Figure 20 illustrates the system which was simulated to test the effect of nonindependent source streams on the interarrival time dependency of superposed processes. The output of identical queuing systems was used to provide processes with correlated interarrival

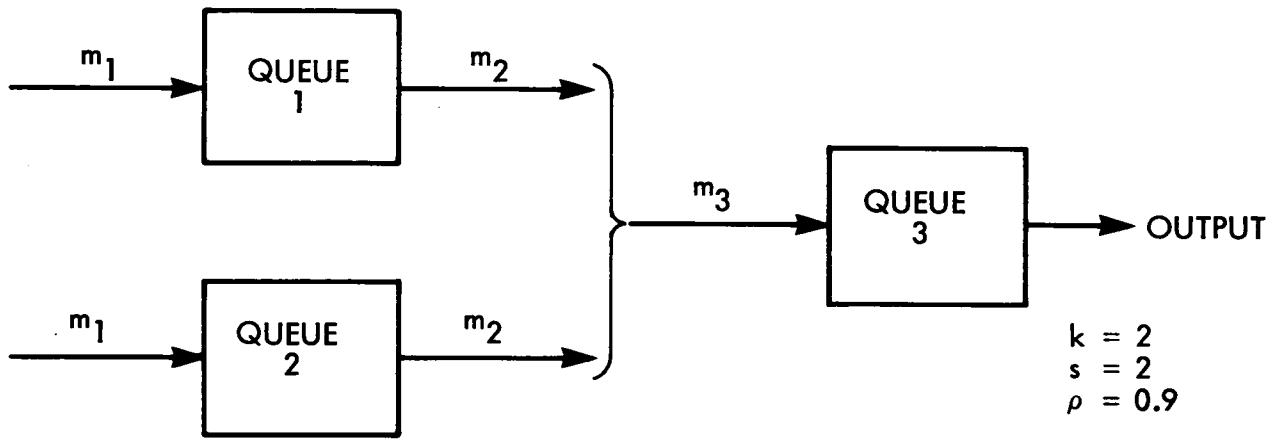


Figure 20. The model used to test the effect of interarrival dependency on the interdeparture dependency.

times. These processes were then merged and used as input to a queuing system (Queue 3 in the figure). This queue was a two-server system with service shape parameter equal to two; traffic intensity was 0.9.

Table 7 gives results of four experiments. The first column gives the interarrival shape parameter (m_1); the next three describe Queues 1 and 2 in terms of service shape parameter (k), number of servers (s), and traffic intensity (ρ). Column 5 gives the shape parameter of an Erlang process with the same mean and variance as the output from Queues 1 and 2 (m_2). The final three columns show the mean queue length of Queue 3. The first of these is the result obtained in Chapter III using merged renewal processes with Erlang distributions with shape parameter m_2. Next, the results for the merger of the output processes is shown. Finally, for reasons discussed below, the last column gives the results for cases in which the sources are mutually independent samples from the output distributions.

The first three cases shown were chosen to yield an output process from both Queues 1 and 2 which had a variance equal to that of an Erlang process with shape parameter 2. These three experiments can thus be compared to results in Chapter III. (See the fifth line of of Table 2, page 32.) Table 7 shows that the behavior of Queue 3 is significantly affected, with mean queue lengths varying from 2.9 to 4.7.

One major question remains: Is this differing behavior due to the varying correlations introduced by the three queuing systems, or is it due to the shape of the output distributions? In Figure 21, the density functions of these three interdeparture processes are plotted. Although the functions appear very different, all three have means and

Table 7. Effect of dependency in source processes on the dependency resulting from superposition.

Source Shape m_1	Queues 1 and 2 k	s	ρ	Equiv. Output Shape m_2	Mean Length of Queue 3 Ind. Erlang Sources (Note 1)	Dep. Output Sources (Note 2)	Ind. Output Sources (Note 3)
4	4	2	.90	2.00	3.71	2.90	3.84
1	10	2	.91	2.00	3.71	3.26	3.52
1	10	1	.75	2.00	3.71	4.72	3.51
1	2	2	.90	1.41	4.74	4.47	--

Notes: 1. Source streams were Erlang distributed renewal processes.

2. Sources were the output of Queues 1 and 2, including dependency introduced.

3. Sources were independent samples from the output distributions of Queues 1 and 2.

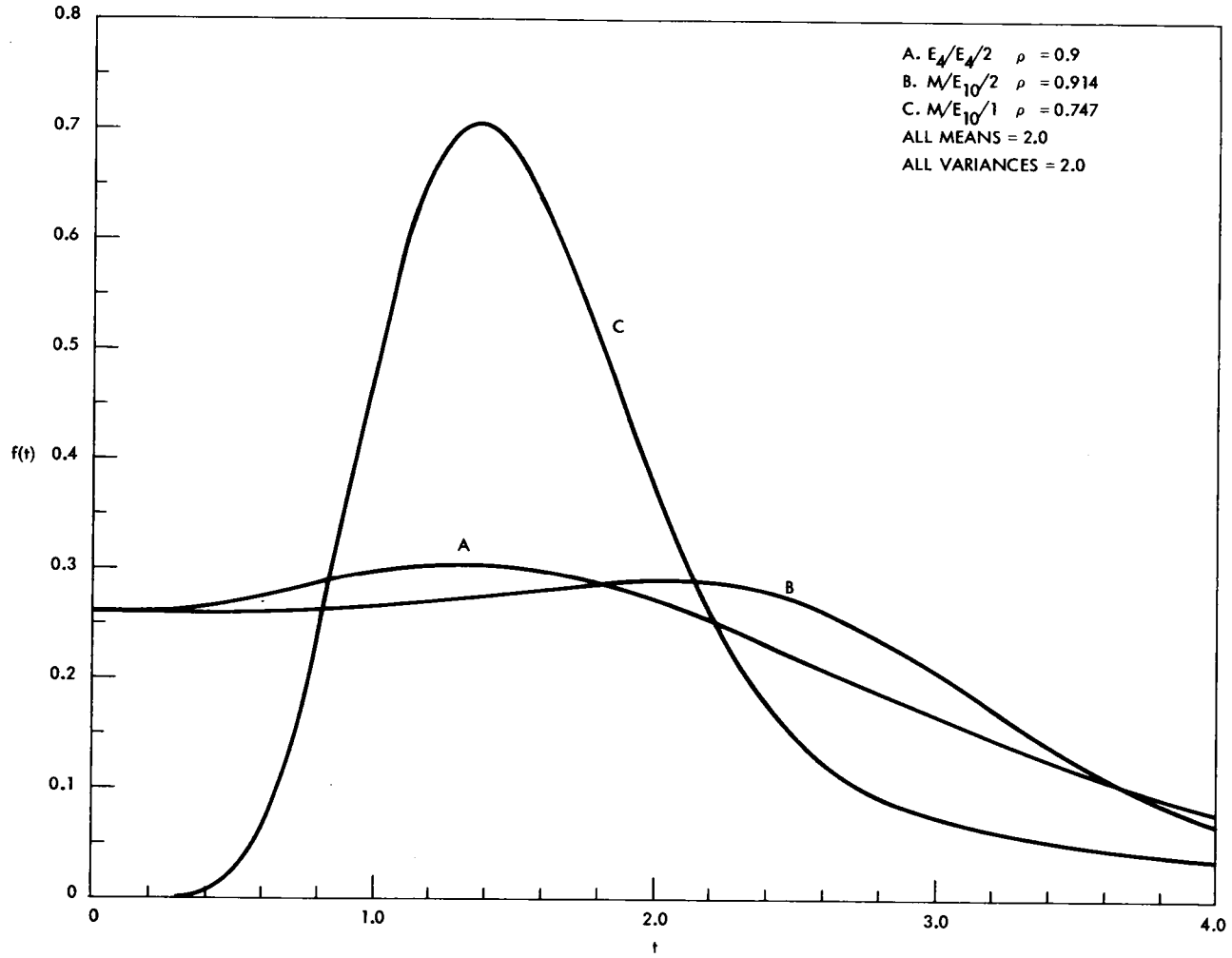


Figure 21. Density functions of interdeparture times from queues 1 and 2.

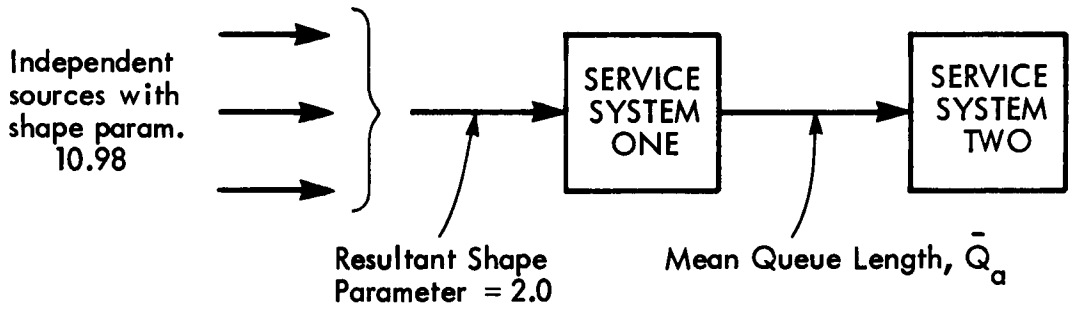
variances equal to 2.00. To determine if this shape variation is significant, interarrival processes were formed from independent samples from these distributions; these processes were then merged and used as input to Queue 3. The final column of Table 7 shows that the results varied little from those obtained with Erlang inputs. This indicates that the differing performance for the three cases is due primarily to the different dependency effects.

The results from these first three experiments run counter to the objective of developing a technique of queuing network analysis. These cases are, however, somewhat extreme. In the first case, the input process has a coefficient of variation of 50. Figures 13 and 18 (pages 58 and 68, respectively) show that, for many values of system parameters, such a low variation would not be encountered. Experiments 2 and 3 involved a low coefficient of variation in the service time--a condition which introduces high correlation in the output.

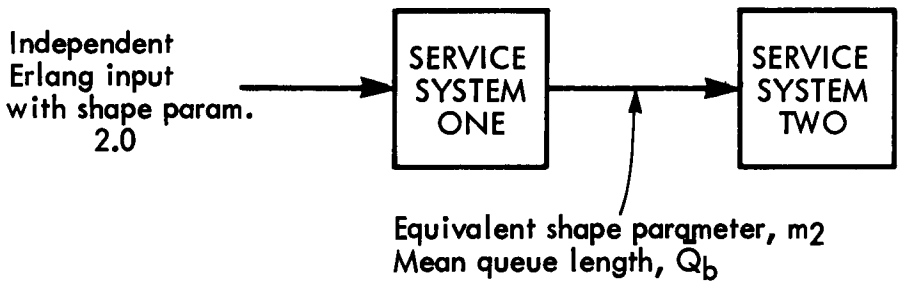
The final experiment in Table 7 illustrates the result for more moderate parameters. In this case, mean queue length is little affected by the dependency in the source streams.

The Effect of Interarrival Time Dependency on Interdeparture Time Dependency

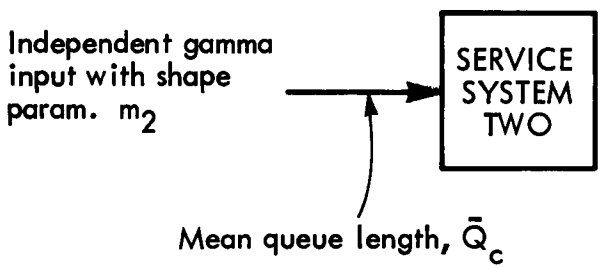
Figure 22 shows three systems used to observe the effect of dependency in the input process upon the dependency in the output process of a queue. The first system is a two-queue configuration in which the interarrival times to the first queue contain dependency introduced by the superposition of three similar processes. Each of



(a) System with dependent input to first queue.



(b) System with independent input to first queue.



(c) System with independent input to second queue.

Service System One: 2 server, E_2 distribution
Service System Two: 1 server, exponential

Figure 22. Systems used for comparing effect of input dependency on output dependency.

these sources is a renewal process with gamma-distributed interarrival times, the parameters of which are chosen to yield a resultant process with unit mean and a variance equal to that of an Erlang process with a shape parameter of 2. If it were not for the dependency introduced by superposition, this system would behave very similarly to one in which the input to the first queue was E_2 (Figure 22b). Comparisons of the mean length of the second queue for various traffic intensities of the first queue are made for these two cases. A third case, shown in Figure 22c, gives queue performance for an equivalent system with independent input to the second queue.

The results of Chapters III and IV indicate that correlation derived from superposition is more influential on queue behavior than is the correlation introduced by a prior queuing system. The current results, shown in Table 8, are consistent with these earlier findings. The greatest impact on the mean length of the second queue occurs when the traffic intensity of the first queue is low. In these cases, the dependency introduced by superposition is passed through the first queue to affect the second. At higher traffic intensities, however, the first queue buffers the second against the superposition-induced dependency.

Table 8. Interdeparture time dependency as affected by interarrival time dependency.

Traffic Intensity of 1st Queue	Mean Lengths of Second Queue			
	Dependent Input to 1st Queue ($\overline{Q}_a$)	Independent Input to 1st Queue ($\overline{Q}_b$)	Independent Input to 2nd Queue ($\overline{Q}_c$)	Equiv. Shape (m_2)
0.0	4.4	5.9	5.9	2.0
0.2	4.7	5.8	6.1	1.8
0.9	5.1	5.7	6.6	1.5
1.0	6.0	6.0	6.4	1.6

CHAPTER VI

ANALYSIS OF QUEUING NETWORKS

Two fundamental tools for analyzing networks of queues are now available. In Chapter III, the interarrival time variance of a process resulting from superposition of several processes was developed. And in Chapter IV, estimates for the variance of the interdeparture time of one- and two-server Erlang queues were discussed. Furthermore, the effect of the dependency introduced by these processes was explored for a variety of parameter values.

These findings are adequate to permit analysis of certain network configurations. First, it was found that when a small number of processes were superposed to provide input to a service system, the mean queue length could be estimated by taking a weighted average of queue lengths which would result from the individual source processes. Over the range of parameters tested, up to five source streams could be handled in this manner with less than 10% error in the estimate.

In Chapter IV it was found that dependency in the output process of queuing systems was of little importance in predicting the performance of a subsequent queue. In the worst case, less than a 20% change in mean queue length resulted from this correlation; less than 10% difference was noted for other sets of parameters. This conclusion cannot reliably be extrapolated to ranges of system parameters not considered in Chapter IV. In particular, low values of traffic intensity (less than 0.7) were not tested.

In Chapter V it was found that some dependency in source processes prior to superposition could be tolerated without major effect on system behavior. Further, as might be expected, the existence of waiting lines in the system tend to prevent dependencies in the stochastic process from accumulating in the system.

Two problems will now be analyzed using these guidelines. To avoid unexplored parameter ranges, distribution shape parameters were limited to a maximum of four and all traffic intensities were 0.7 or greater. The first problem is a series of four queues with two servers each. The more extreme dependencies introduced by superposition are thus not present in this example. The second problem is more complex, involving a converging network. The conceptual basis for this example is a computer system with multiple remote job entry stations.

Estimates of mean queue length were made for all queues in both systems using the above concepts. The systems were then simulated and the results compared to these estimates.

Series Queue Example

Four queues in tandem with two parallel servers each are analyzed first. Each queue has a traffic intensity of 0.9. Service time distributions are Erlang and the input to the first queue is Poisson. The mean length of the first queue is obtained directly from the tables by Hillier and Lo [41, 1971]. The methods of Chapter IV were then used to obtain an estimate of the variance of the inter-departure time distribution from the first queue. This variance was then used to obtain an equivalent shape parameter for the input to the second queue. Interpolation of Hillier and Lo's results provide the

estimate of the mean length of the second queue. Figure 23, which relates the shape parameter of the input to a queue to an equivalent shape parameter of the output, was developed from the methods of Chapter IV. This figure is now used to obtain the equivalent shape parameter for the input to the third queue. This process is then repeated for the fourth queue. Table 9 shows the results of these estimates contrasted to those obtained by a simulation of the system.

Queuing Network Example

Figure 24 represents a computer system with five remote batch entry points which are serviced by two satellite computers (Queues 1 and 2). These inputs, along with a local entry point, provide input to a multi-programmed computer (Queue 3). A job is completed when output is printed on one of four printers (Queue 4). Input rates and service characteristics are listed on the figure.

Analysis proceeds in a manner similar to the preceding example. Beginning with the superposed inputs, resultant equivalent shape parameters are found. From these, and utilizing Figure 23, equivalent shape parameters for the output processes of Queues 1 and 2 are estimated. Also, the mean queue lengths of Queues 1 and 2 are estimated using the weighted average of performance resulting from the individual inputs. This process continues until all queues are analyzed. Table 10 gives the results of the estimates and contrasts these findings to the outcome of a simulation of the system.

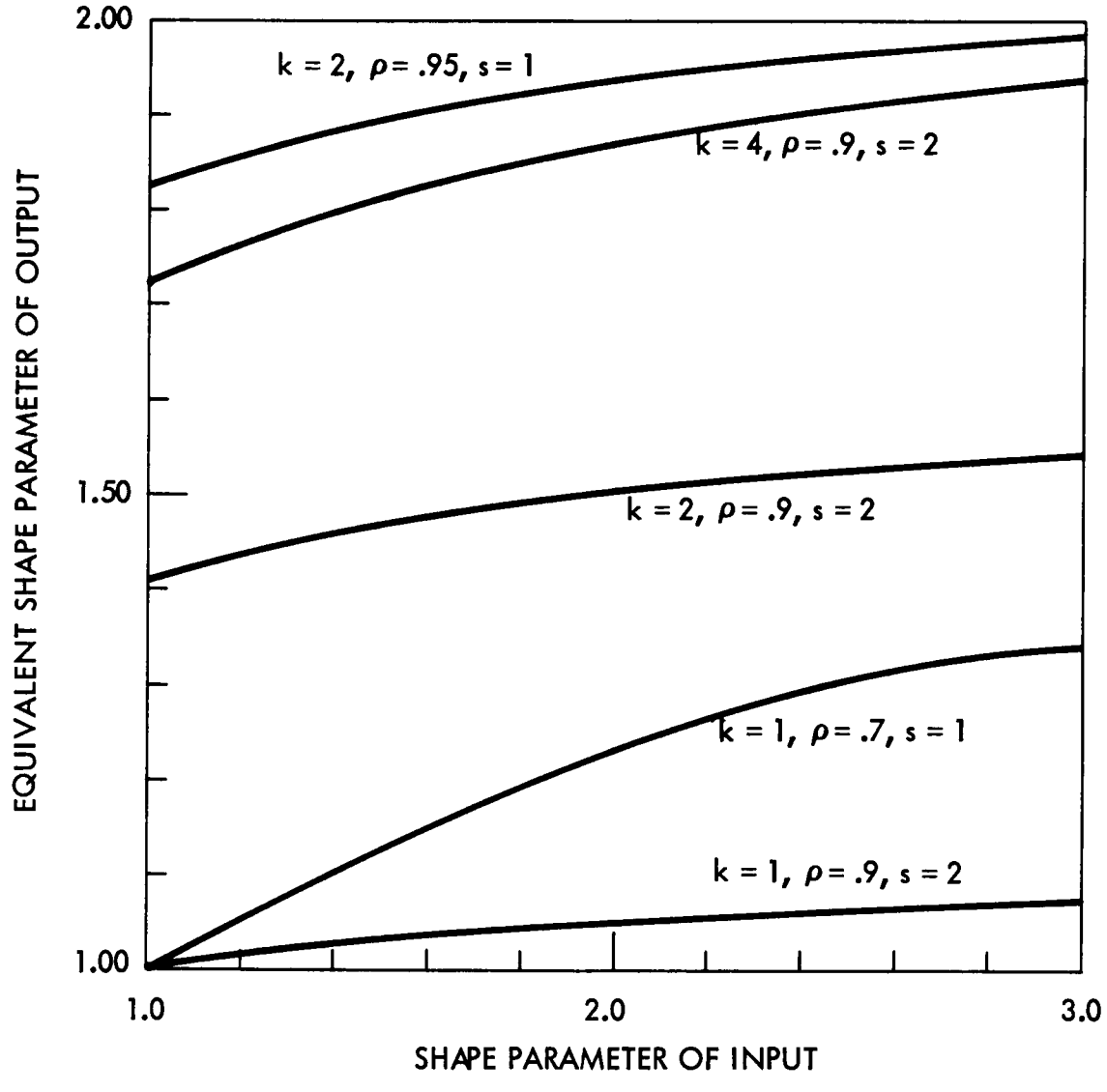
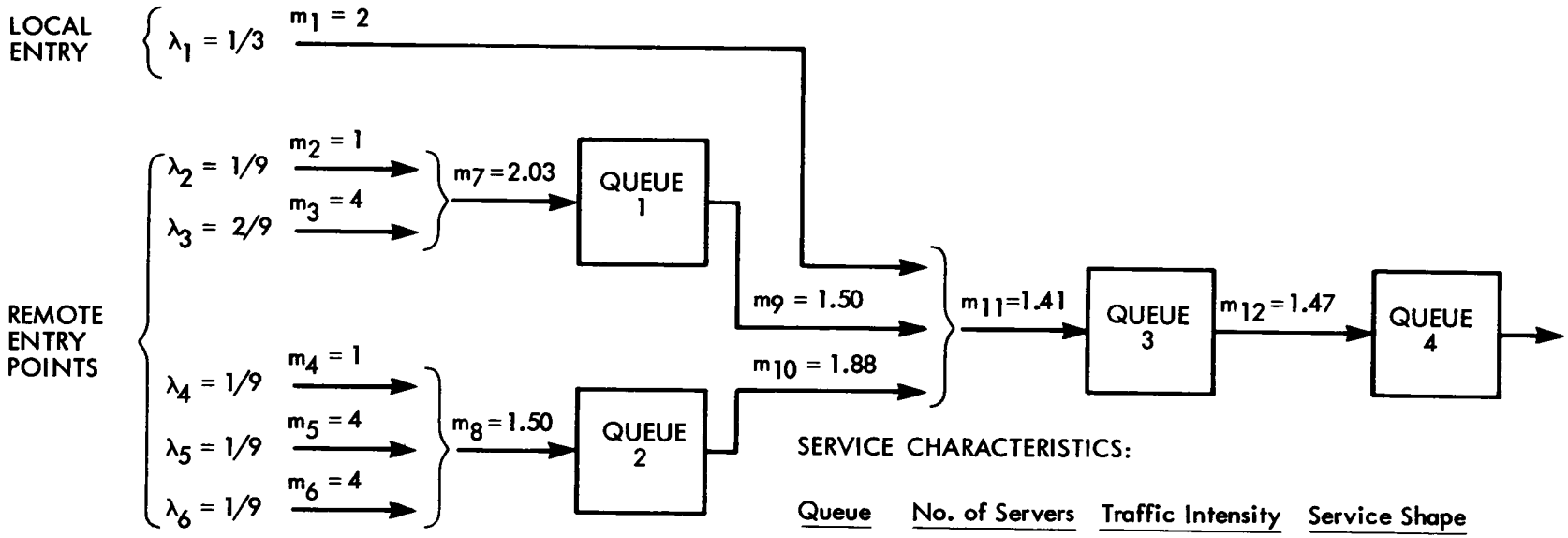


Figure 23. Equivalent shape parameter of output process as a function of input process shape.

Table 9. Series queue example results.

Queue Number	Service Shape Parameter	Equivalent Interdeparture Shape	Estimated Mean Queue Length	Mean Queue Length From Simulation	Ratio of Estimate/ Simulation
1	4	1.72	4.82	4.82	1.00
2	2	1.48	4.09	4.08	1.00
3	1	1.03	6.34	6.10	1.04
4	2	1.42	5.68	5.53	1.03



SERVICE CHARACTERISTICS:

Queue	No. of Servers	Traffic Intensity	Service Shape
1	1	0.7	1
2	1	0.95	2
3	2	0.9	2
4	4	0.9	1

Figure 24. Configuration of sample network.

Table 10. Queuing network example results.

Queue Number	Estimated Mean Queue Length	Mean Queue Length from Simulation	Ratio of Estimate/Simulation
1	1.12	1.14	0.98
2	8.89	9.00	0.99
3	3.96	4.08	0.97
4	5.80	5.40	1.07

CHAPTER VII

CONCLUSIONS

The outlook for developing a method of analyzing very general networks of queues is mixed. In the preceding chapter, two sample problems were treated quite successfully. These problems were, however, specially chosen to take advantage of the rather extensive data gathered in earlier chapters. Also, approximations were liberally used: the distribution of interdeparture times was an approximation, Hillier and Lo's results for queues with Erlang inputs were used for non-Erlang cases, interdeparture time dependency was ignored, and superposition-induced dependency was handled in an inexact manner. On the other hand, none of these approximations were casually employed. Each was separately supported by rather detailed comparisons and experiments. Furthermore, some cumulative effects had been explored.

This general approach, characterized by careful approximation and application of empirical results, seems to be one of the more likely avenues of future research in queuing networks. In the final section of this chapter some conclusions about the potential of such work will be advanced. First, however, the specific accomplishments of the current research will be reviewed with an emphasis on the utility of the findings. Extensions to this work will then be proposed.

Summary of Results

There are two products resulting from the work of Chapter III on superposition of input processes. First, the derivation of the distribution of the resultant process permits the calculation of an equivalent

input shape parameter. This is then used, along with the work on output distributions, to estimate the variance of the interdeparture times of a queue to which this input is presented.

It was also shown that the dependency introduced by superposition typically has a significant effect on queue behavior. The mean queue length could, however, often be closely approximated by using a weighted average technique. The restriction to this approach was that the number of source processes must be small (less than about five). For large numbers of sources, bounds on the mean queue length could be established.

A related result of Chapter III was the observation that superposition induced dependency continues to affect queue behavior in cases where there are a relatively large number of source processes. It was seen, for example, that the merger of 100 E_2 sources results in an interarrival time distribution which is almost indistinguishable from an exponential distribution. The behavior of a subsequent queue, however, was significantly different from that of a system with Poisson input, indicating that the dependency from superposition was still effective.

In Chapter IV it was found that, over a wide range of system parameters, the dependency introduced in the output process of a service facility is of little importance to the behavior of a subsequent queue. Also, good approximations to the output distribution were obtained for one- and two-server systems. The variance in interdeparture times thus obtained is used in evaluating the performance of a subsequent service system.

The way in which dependency effects may accumulate, or alternately be diluted, in networks was briefly explored in Chapter V. Most disturbing from the standpoint of developing a method for analyzing networks was the result of superposing processes with highly correlated interarrival times. Such processes could be generated in a queuing network by service systems with low variation in service times (coefficients of variation less than 50) or by low variances in both arrival times and service times. It was also found, however, that in systems with lower shape parameters, this type of correlation had little effect on queue performance.

It is reasonable to expect the existence of waiting lines in a network to dampen out dependencies in the stochastic processes. This conjecture was supported by additional experiments of Chapter V.

The simulations used to test dependency in superposed inputs provided, as a by-product, the opportunity to observe the distribution robustness of Erlang queuing systems. Superposition of Erlang processes results in a non-Erlang process. One set of experiments in Chapter III involved simulating a queuing system with independent interarrivals drawn from these non-Erlang distributions. In Table 11, the mean queue length resulting from these simulations are compared to values interpolated from Hillier and Lo's tables for Erlang inputs with the same variance. For most cases, agreement is within 5%--about the accuracy ascribed to the simulation results. (One exception is a case with high traffic intensity, a condition which increases the variance of queue length.) The strong indication of these comparisons is that the differences between the resultant distributions and Erlang

Table 11. Comparison of mean queue lengths for Erlang and non-Erlang input processes.

Input Shape Parameter (m)	Service Shape Parameter (k)	Number of Servers (s)	Traffic Intensity (ρ)	Erlang Mean Queue (Note 1)	Non-Erlang Mean Queue (Note 2)	Percent Difference
1.60	2	2	0.9	4.19	4.10	2.1
2.36	2	2	0.9	3.36	3.36	0.0
2.83	2	2	0.9	3.08	3.12	1.3
1.60	4	2	0.9	3.30	3.29	0.3
2.83	4	2	0.9	2.14	2.21	3.3
1.60	8	2	0.9	2.89	2.81	2.8
2.83	8	2	0.9	1.83	1.88	2.7
1.60	4	2	0.98	24.17	26.45	9.4
2.83	4	2	0.98	11.94	12.20	2.2
1.60	2	2	0.7	0.70	0.72	2.9
2.83	2	2	0.7	0.48	0.51	6.2
1.60	2	2	0.98	29.60	25.87	3.8
2.83	2	2	0.98	20.00	19.21	4.0
1.60	2	10	0.9	3.40	3.43	0.9

Table 11 (Continued)

Input Shape Parameter (m)	Service Shape Parameter (k)	Number of Servers (s)	Traffic Intensity (ρ)	Erlang Mean Queue (Note 1)	Non-Erlang Mean Queue (Note 2)	Percent Difference
1.29	2	2	0.9	4.79	4.74	1.0
1.16	2	2	0.9	5.17	5.11	1.2
1.10	2	2	0.9	5.40	5.52	2.2
1.00	2	2	0.9	5.77	5.83	1.0

Notes: 1. Interpolated from Hillier and Lo.

2. Simulation results using independent samples from superposed distribution.

distributions of similar mean and variance are unimportant in queuing systems.

Since approximations play a major role in the analysis approach used in this paper, biases in these estimates need to be considered. A qualitative examination of the various sources of error is encouraging. It appears that the weighted average method of analyzing systems with a few superposed input processes will consistently lead to an underestimate of the mean queue length. On the other hand, the approximation of the variance of interdeparture times seems to be consistently high, leading to an overestimate of subsequent queues. Neglecting the dependency introduced into the output process of a queue appears to give mixed results; similarly, using results for queues with Erlang inputs shows no consistent bias when used to approximate non-Erlang input systems. Although these arguments are not conclusive, they gain limited support from the accuracy achieved in analyzing the sample problems.

Extensions

The potential avenues of future research in queuing networks is practically unlimited; in this section, the more evident limitations of the current work will be addressed. First, systems with more than two servers per stage need consideration. A direct extension of the approaches used in Chapter IV appears to be cumbersome. In limited tests, a Monte Carlo scheme used in conjunction with earlier results has shown promise. As in Chapter IV, the distribution of interdeparture times is written

$$C(t) = \sum_{i=1}^{S} P_i \Pr\left(t_{min_i} < t\right) \quad ,$$

where s is the number of servers. Unlike the development in Chapter IV, however, the time from which departures are measured is not a regeneration point, but is an arbitrary instant. This makes the P_i equivalent to the long-range state probabilities of the system. These can be obtained from Hillier and Lo's results. Also, t_{min_i} is redefined as the residual time until a departure, given the system is in state i. Then define

t_{e_i} as the elapsed time to the earliest arrival from the (s - i) empty channels and

t_{f_i} as the time to the earliest arrival from the i full channels.

Then

$$Pr(t_{min_i} < t) = 1 - Pr(t_{e_i} > t)\ Pr(t_{f_i} > t) \quad .$$

For cases where i is two or less, an approach equivalent to that developed in Chapter IV provides $Pr(t_{e_i} > t)$. (An adjustment is required to accommodate the change from a regeneration point to an arbitrary point in time.) When i is greater than two, a Monte Carlo development of the distribution of t_{e_i} is used.

Computation time obviously will increase rapidly as the number of servers increases beyond 2. Some steps can be taken, however, to reduce this time. First, it is probably possible to eliminate the consideration of very unlikely events. For example, when there are many empty channels, the next departure will most likely take place in one of the first several service channels filled. Simulation of large numbers of channels may thus be avoided. Also supporting such a

strategy is the fact that, at higher traffic intensities, system states involving large numbers of empty channels are increasingly improbable.

Although simulation is still required in this approach, the efficiency of the computations will probably far exceed that of simulating the queue itself. The payoff seems, from very preliminary examinations, to encourage continuation of this approach.

A second major extension is the ability to handle decomposition of the event streams in the system. Some work by Disney and others was cited earlier on this subject. Much remains to be done, however, to discover the effects that decomposition will have on queue performance.

Also required for general network systems will be an ability to handle feedback within the system. This will most likely involve an iterative procedure, beginning with estimates of the characteristics of the feedback streams.

Studies of the effect of additional combinations of system parameters are required before a complete methodology can be claimed to exist. Low traffic intensities and low variances in arrival and service times are examples. Also, very high variances (coefficients of variation greater than 100) may be of interest.

Finally, confidence in the methodology must be established. This must be largely an empirical process involving the successful analysis of test problems. Additional consideration of the errors and biases of the individual approximations may also play a role in this effort.

Outlook for the Future

Any predictions about the eventual success of an analysis tool for queuing networks must certainly be provisional. There would appear

to be no reason why continued work would not be fruitful. It is apparent, though, that gains will be costly. Before additional work is launched, it is advisable that the worth of the final product be evaluated. The objectivity of such an evaluation, coming at the close of this effort, would certainly be open to question. Admittedly, however, even with the proposed extensions, the network which would yield to analysis would remain limited.

A major restriction is the exclusion of transient behavior. It is well recognized that many systems rarely reach a steady-state condition. Also, queue disciplines, nonstationary processes, reneging, and limitations on queue length are important topics which have not been addressed.

The value of the research which has been conducted, and that which is proposed, must be assessed in terms of these limitations. Even with these limitations, there would seem to be an increasing potential for application of queuing network theory. The complexity of computing systems--and further, the development of networks of computers--offers possibilities. Communications networks are obvious candidates and with somewhat more imagination, one might conceive of applications in economic systems.

Important systems do exist which can be represented as networks of queues, whether or not they will admit an approach of the above type. If they can be so analyzed, the advantage of a compact methodology over simulation should be sufficient to encourage continued research.

REFERENCES

REFERENCES

1. Adiri, I., and B. Avi-Itzhak, "A Time-Sharing Model with Many Queues," Operations Research, Vol. 17, No. 6 (November-December, 1969), pp. 1077-1089.

2. Ancker, C. J., Jr., and A. V. Gafarian, "Queuing with Multiple Poisson Inputs and Exponential Service Times," Operations Research, Vol. 9, No. 3 (May-June, 1961), pp. 321-327.

3. Avi-Itzhak, B., "A Sequence of Service Stations with Arbitrary Input and Regular Service Times," Management Science, Vol. 11, No. 5 (March, 1965), pp. 565-571.

4. Avi-Itzhak, B., and D. P. Heyman, "Approximate Queuing Models for Multiprogrammed Computer Systems," Operations Research, Vol. 21, No. 6 (November-December, 1973), pp. 1212-1230.

5. Avi-Itzhak, B., and M. Yadin, "A Sequence of Two Servers with No Intermediate Queues," Management Science, Vol. 11, No. 5 (March, 1965), pp. 553-564.

6. Bharucha-Reid, A. T., Elements of the Theory of Markov Processes and Their Applications, McGraw-Hill Book Company, Inc., New York, 1960.

7. Bhat, U. N., "Sixty Years of Queuing Theory," Management Science, Vol. 15, No. 6 (February, 1969), pp. B-280 - B-294.

8. Boling, R. W., "How to Analyze Sequential Crew Systems," Industrial Engineering, Vol. 2, No. 2 (February, 1970), pp. 12-18.

9. Brumelle, S. L., "Bounds on the Wait in a GI/M/k Queue," Management Science, Vol. 19, No. 7 (March, 1973), pp. 773-777.

10. Burke, P. J., "The Output of a Queuing System," Operations Research, Vol. 4, No. 6 (December, 1956), pp. 699-704.

11. Burke, P. J., "The Dependence of Delays in Tandem Queues," Annals of Mathematical Statistics, Vol. 35 (1964), pp. 874-875.

12. Burke, P. J., "The Output Process of a Stationary M/M/s Queuing System," Annals of Mathematical Statistics, Vol. 39 (1968), pp. 1144-1152.

13. Burke, P. J., "The Dependence of Sojurn Times in Tandem M/M/s Queues," Operations Research, Vol. 17 (1969), pp. 754-755.

14. Chang, W., "Output Distribution of a Single-Channel Queue," Operations Research, Vol. 11, No. 4 (July-August, 1963), pp. 620-623.

15. Chang, W., "Queues with Feedback for Time-Sharing Computer System Analysis," Operations Research, Vol. 16, No. 3 (May-June, 1968), pp. 613-627.

16. Cinlar, E., and R. L. Disney, "Streams of Overflow from a Finite Queue," Operations Research, Vol. 15 (1967), p. 131.

17. Conway, R. W., and W. L. Maxwell, "Network Dispatching by the Shortest Operation Discipline," Operations Research, Vol. 10, No. 1 (January-February, 1962), pp. 51-73.

18. Conway, R. W., W. L. Maxwell, and L. W. Miller, Theory of Scheduling, Addison-Wesley Publishing Company, Reading, Massachusetts, 1967.

19. Cooper, R. B., Introduction to Queuing Theory, The Macmillan Company, New York, 1972.

20. Cox, D. R., Renewal Theory, Methuen and Company, Ltd., London, 1962.

21. Cox, D. R., "Some Problems of Statistical Analysis Connected with Congestion," Proceedings of the Symposium on Congestion Theory, The University of North Carolina Press, Chapel Hill, 1965, pp. 289-316.

22. Daley, D. J., "Monte Carlo Estimation of the Mean Queue Size in a Stationary GI/M/1 Queue," Operations Research, Vol. 16, No. 5 (September-October, 1968), pp. 1002-1005.

23. Disney, R. L., "Analytic Studies of Stochastic Networks Using Methods of Network Decomposition," Journal of Industrial Engineering, Vol. 18, No. 1 (January, 1967), pp. 140-145.

24. Disney, R. L., "A Matrix Solution for the Two Server Queue with Overflow," Management Science, Vol. 19 (1972), pp. 254-265.

25. Disney, R. L., R. L. Farrell, and P. R. de Morais, "A Characterization of M/G/1 Queues with Renewal Departure Processes," Management Science, Vol. 19, No. 11 (July, 1973), pp. 1222-1228.

26. Dunn, P. F., C. D. Flagle, and P. A. Hicks, "The QUEUIAC: An Electromechanical Analog for the Simulation of Waiting-Line Problems," Operations Research, Vol. 4, No. 4 (August, 1956), pp. 648-662.

27. Enns, E. G., "Some Waiting-Time Distributions for Queues with Multiple Feedback and Priorities," Operations Research, Vol. 17, No. 3 (May-June, 1969), pp. 519-553.

28. Evans, R. V., "Queuing When Jobs Require Several Services Which Need Not be Sequenced," Management Science, Vol. 10, No. 2 (January, 1964), pp. 298-315.

29. Feller, W., An Introduction to Probability Theory and Its Applications, Volume II, John Wiley and Sons, Inc., New York, 1966.

30. Finch, P. D., "The Output Process of the Queuing System M/G/1," Journal of the Royal Statistical Society, Series B, Vol. 21, No. 2 (1959), pp. 375-380.

31. Friedman, H. D., "Reduction Methods for Tandem Queuing Systems," Operations Research, Vol. 13, No. 1 (January-February, 1965), pp. 121-131.

32. Ghosal, A., "Queues in Series," Journal of the Royal Statistical Society, Vol. 24, No. 2 (1962), pp. 359-363.

33. Gordon, W. J., and G. F. Newell, "Closed Queuing Systems with Exponential Servers," Operations Research, Vol. 15, No. 2 (March-April, 1967), pp. 254-265.

34. Gordon, W. J., and G. F. Newell, "Cyclic Queuing Systems with Restricted Length Queues," Operations Research, Vol. 15, No. 2 (March-April, 1967), pp. 266-277.

35. Harris, C. M., "Queues With Multiple Poisson Input," Management Science, Vol. 12, No. 4 (December, 1965), p. C-55.

36. Haydon, B. J., "System States for a Series of Finite Queues," Operations Research, Vol. 20 (1972), pp. 1137-1141.

37. Heffer, J. C., "Steady-State Solution of the $M/E_k/c$ (∞, FIFO) Queuing System," Canadian Operations Research Society Journal, Vol. 7 (1969), pp. 16-30.

38. Hildebrand, D. K., "On the Capacity of Tandem Server, Finite Queue, Service Systems," Operations Research, Vol. 16, No. 1 (January-February, 1968), pp. 72-82.

39. Hiller, F. S., and R. W. Boling, "The effect of Some Design Factors on the Efficiency of Production Lines with Variable Operation Times," Journal of Industrial Engineering, Vol. 17, No. 12 (December, 1966), pp. 651-658.

40. Hillier, F. S., and R. W. Boling, "Finite Queues in Series with Exponential or Erlang Service Times--A Numerical Approach," Operations Research, Vol. 15, No. 2 (March-April, 1967), pp. 297-303.

41. Hillier, F. S., and F. D. Lo, "Tables for Multiple-Server Queuing Systems Involving Erlang Distributions," Technical Report No. 31 (Office of Naval Research Contract N 00014-67-A-0112-0058), Department of Operations Research, Stanford University, December 28, 1971.

42. Hunt, G. C., "Sequential Arrays of Waiting Lines," Operations Research, Vol. 4 (1956), pp. 674-683.

43. Jackson, J. R., "Networks of Waiting Lines," Operations Research, Vol. 5, No. 4 (August, 1957), pp. 518-521.

44. Jackson, J. R., "Jobshop-like Queuing Systems," Management Science, Vol. 10, No. 1 (October, 1963), pp. 131-142.

45. Jackson, R. R. P., "Random Queuing Processes with Phase-Type Service," Journal of the Royal Statistical Society, Series B, Vol. 18, No. 1 (1956), pp. 129-132.

46. Kendall, D. G., "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of the Imbedded Markov Chain," The Annals of Mathematical Statistics, Vol. 24 (1953), pp. 338-354.

47. Kiefer, J., and J. Wolfowitz, "On the Theory of Queues with Many Servers," American Mathematical Society Transactions, Vol. 78 (1955), pp. 1-18.

48. Kingman, J. F. C., On the Algebra of Queues, Methuea & Co., Ltd., London, 1966.

49. Koenigsberg, E., "Finite Queues and Cyclic Queues," Operations Research, Vol. 8, No. 2 (March-April, 1960), pp. 246-253.

50. Love, R. F., "Steady-State Solution of the Queuing System $E_W/M/s$ with Batch Service," Operations Research, Vol. 18, No. 1 (January-February, 1970), pp. 160-171.

51. Maaloe, E., "Approximation Formulae for Estimation of Waiting Time In Multiple-Channel Queuing Systems," Management Science, Vol. 19, No. 6 (February, 1973), pp. 703-710.

52. Maxwell, W. L., "On the Generality of the Equation $L = \lambda W$," Operations Research, Vol. 18, No. 1 (January-February, 1970), pp. 172-174.

53. Mayhugh, J. O., "A Note on the Queue $M/E_k/r$," Management Science, Vol. 16 (1970), pp. 512-513.

54. Mayhugh, J. O., and R. E. McCormick, "Steady State Solution of the Queue $M/E_k/r$," Management Science, Vol. 14 (1968), pp. 692-712.

55. Nelson, R. T., "Waiting-Time Distributions for Application to a Series of Service Centers," Operations Research, Vol. 6, No. 6 (November-December, 1958), pp. 856-862.

56. Reich, E., "Waiting Times when Queues are in Tandem," The Annals of Mathematical Statistics, Vol. 28, No. 3 (September, 1957), pp. 768-773.

57. Reich, E., "Note on Queues in Tandem," Annals of Mathematical Statistics, Vol. 34 (1963), pp. 338-341).

58. Reich, E., "Departure Processes," Proceedings of the Symposium on Congestion Theory, The University of North Carolina Press, Chapel Hill, 1965, pp. 439-457.

59. Reitman, J., Computer Simulation Applications, John Wiley and Sons, Inc., New York, 1971.

60. Restrepo, R. A., "A Queue with Simultaneous Arrivals and Erlang Service Distribution," Operations Research, Vol. 13, No. 3 (May-June, 1965), pp. 375-381.

61. Saaty, T. L., "Resumé of Useful Formulas in Queuing Theory," Operations Research, Vol. 5, No. 2 (April, 1957), pp. 161-200.

62. Saaty, T. L., Elements of Queuing Theory with Applications, McGraw-Hill Book Company, New York, 1961.

63. Saaty, T. L., "Stochastic Network Flows: Advances in Networks of Queues," Proceedings of the Symposium on Congestion Theory, The University of North Carolina Press, Chapel Hill, 1965, pp. 86-107.

64. Schrage, L. E., "The Queue M/G/1 With Feedback to Lower Priority Queues," Management Science, Vol. 13, No. 7 (March, 1967), pp. 466-474.

65. Shapiro, S., "The M-Server Queue with Poisson Input and Gamma-Distributed Service of Order Two," Operations Research, Vol. 14, No. 4 (July-August, 1966), pp. 685-694.

66. Takács, L., "Two Queues Attended by a Single Server," Operations Research, Vol. 16, No. 3 (May-June, 1968), pp. 639-650.

67. Wishart, D. M. G., "A Queuing System with Chi-Squared Service Times Distribution," Annals of Mathematical Statistics, Vol. 27 (1956), pp. 768-779.

68. Young, J. P., "Administrative Control of Multiple-Channel Queuing Systems with Parallel Input Streams," Operations Research, Vol. 14, No. 1 (January-February, 1966), pp. 145-156.

69. Yu, O. S., "On the Steady-State Solution of an $E_m/E_k/r$ Queue with Heterogeneous Servers," Technical Report No. 38 (Office of Naval Research Contract N 00014-67-A-0112-0058), Department of Operations Research, Stanford University, June 23, 1972.

APPENDICES

APPENDIX A

ERLANG AND GAMMA DISTRIBUTIONS

The form of the Erlang distribution used in this paper conforms to generally accepted expressions used in management science literature. The density function is

$$f(t) = \frac{(\mu k)(\mu k t)^{k-1} e^{-\mu k t}}{(k-1)!}, \quad t > 0$$

where k is a positive integer. The expectation and variance are

$$E(t) = \frac{1}{\mu}$$

$$Var(t) = \frac{1}{\mu^2 k} .$$

The gamma distribution, of which the Erlang is a special case, is generally expressed in statistics literature in a form which leads to a different interpretation of the parameter μ. To avoid confusion within this paper, a somewhat nonstandard expression for the gamma distribution is used. The density function is

$$f(t) = \frac{(\mu k)(\mu k t)^{k-1} e^{-\mu k t}}{\Gamma(k)}, \quad t > 0$$

where k is any positive number. The expectation and variance are thus the same as for the above Erlang distribution.

APPENDIX B

THE RESIDUAL TIME DISTRIBUTION OF AN ERLANG DISTRIBUTED RENEWAL PROCESS

Consider random variables of the form

$$s_n = t_1 + t_2 + \ldots + t_n$$

where the t_i are identically distributed, mutually independent random variables with distribution function $F(t)$. The sequence of random variables, s_n, forms a renewal process. When such a process is used to represent the input to a queuing system, the t_i are the interarrival times and the points s_n on the time line are the instants of arrivals, or arrival epochs. Let the process be observed at an arbitrary point in time, s, occurring between epochs s_{n-1} and s_n. The residual waiting time, t_r, is given by

$$t_r = s_n - s \quad .$$

Feller [29, 1966, pp. 352-353] shows that the t_r are distributed by

$$H(t_r) = \lambda \int_0^{t_r} [1 - F(t)]dt \qquad (B.1)$$

where λ is the mean interarrival rate (arrivals/unit time).

Now let $F(t)$ be an Erlang distribution with density

$$f(t) = \frac{(\lambda m)^m \, t^{m-1} \, e^{-\lambda m t}}{(m-1)!} \quad .$$

The residual time distribution can be most easily derived by noting that, at time s, there is an equal probability of being in each of the m stages of the interarrival process. Thus

$$H(t_r) = \frac{1}{m} \sum_{r=1}^{m} \int_0^{t_r} \frac{(\lambda m)^r \, t^{r-1} \, e^{-\lambda m t}}{(r-1)!} \, dt =$$

$$= \frac{1}{m} \sum_{r=1}^{m} \left(1 - e^{-\lambda m t_r} \sum_{j=0}^{r-1} \frac{t_r^j \, (\lambda m)^j}{j!}\right) =$$

$$= 1 - \frac{1}{m} e^{-m t_r} \sum_{r=1}^{m} \sum_{j=0}^{r-1} \frac{t_r^j \, (\lambda m)^j}{j!} \quad .$$

Finally,

$$H(t_r) = 1 - e^{-\lambda m t_r} \sum_{r=0}^{m-1} \left(\frac{m-r}{m}\right) \frac{t_r^r \, (\lambda m)^r}{r!} \quad . \qquad \text{(B.2)}$$

The density function, $h(t_r)$, of the residual time is most easily obtained by differentiating expression (B.1):

$$H(t_r) = \frac{d}{dt_r} \lambda \int_0^{t_r} [1 - F(t)] dt =$$

$$= \lambda[1 - F(t_r)] =$$

$$= \lambda\left[1 - \left(1 - e^{-\lambda m t_r} \sum_{r=0}^{m-1} \frac{(\lambda m)^r \, t_r^r}{r!}\right)\right] \quad .$$

Thus

$$h(tr) = \lambda e^{-\lambda m t_r} \sum_{r=0}^{m-1} \frac{(\lambda m)^r \, t_r^r}{r!} \quad . \tag{B.3}$$

The Laplace transform of $h(t_r)$ is

$$L\{h(t_r)\} = h_L(s) = \lambda \sum_{r=0}^{m-1} \frac{(\lambda m)^r}{(s + \lambda m)^{r+1}} \quad . \tag{B.4}$$

The mean of the residual waiting time is given by

$$E(t_r) = - \frac{d}{ds} [h_L(s)]_{s=0}$$

$$= - \lambda \sum_{r=0}^{m-1} (\lambda m)^r \, (- r-1)(s + \lambda m)^{-(r+2)} \Big|_{s=0}$$

$$= \frac{\lambda}{(\lambda m)^2} \sum_{r=0}^{m-1} (r+1) = \frac{1}{\lambda m^2} \sum_{r=1}^{m} r \quad .$$

Finally,

$$E(t_r) = \frac{m+1}{2\lambda m} \quad . \tag{B.5}$$

APPENDIX C

THE DENSITY FUNCTION AND VARIANCE OF INTERARRIVAL TIMES RESULTING FROM THE SUPERPOSITION OF ERLANG RENEWAL PROCESSES

In this appendix the density function and variance of the interarrival time of the process resulting from the superposition of Erlang renewal processes are derived.

The Density Function

The following variables will simplify the expression of the density function of M superposed interarrival processes.

$$p_j(t) = 1 - F_j(t) = e^{-\lambda_j m_j t} \sum_{r=0}^{m_j-1} \frac{(\lambda_j m_j t)^r}{r!}$$

$$q_j(t) = 1 - H_j(t) = e^{-\lambda_j m_j t} \sum_{r=0}^{m_j-1} \left[\frac{m_j - r}{m_j}\right] \frac{(\lambda_j m_j t)^r}{r!}$$

$$U(t) = \prod_{i=1}^{M} q_i(t) = e^{-t \sum_{i=1}^{M} \lambda_i m_i} \prod_{i=1}^{M} \left\{ \sum_{r=0}^{m_j-1} \left[\frac{m_i - r}{m_i}\right] \frac{(\lambda_i m_i t)^r}{r!} \right\}$$

$$w_j = \frac{\lambda_j}{\sum_{i=1}^{M} \lambda_i}$$

The interarrival time distribution given in expression (3.9) on page 23 can now be written as

$$G(t) = 1 - \sum_{j=1}^{M} \frac{w_j\, p_j(t)\, U(t)}{q_j(t)} \quad . \tag{C.1}$$

The density function is

$$g(t) = - \sum_{j=1}^{M} w_j \left[\frac{p_j(t)}{q_j(t)} \left(\sum_{i=1}^{M} \frac{q_i'(t)\, U(t)}{q_i(t)} \right.\right.$$

$$\left.\left. + U(t)\, \frac{q_j(t) p_j'(t) - p_j(t)\, q_j'(t)}{(q_j(t))^2} \right]\right.$$

$$= - U(t) \sum_{j=1}^{M} w_j \left[\frac{p_j(t)}{q_j(t)} \sum_{i=1}^{M} \frac{q_i'(t)}{q_i(t)} \right.$$

$$\left. + \frac{q_j(t)\, p_j'(t) - p_j(t)\, q_j(t)}{(q_j(t))^2}\right] \quad . \tag{C.2}$$

It can be shown that

$$p_j'(t) = \frac{-e^{-\lambda_j m_j t}\, t^{m_j - 1}\, (\lambda_j m_j)^{m_j}}{(m_j - 1)!} \tag{C.3}$$

and from expression (B.3), page 109,

$$q_j'(t) = - \lambda_j\, e^{-\lambda_j m_j t} \sum_{r=0}^{m_j - 1} \frac{(\lambda_j m_j t)^r}{r!} \quad . \tag{C.4}$$

Noting that the exponential, $e^{-\lambda_j m_j t}$, is common to the numerator and denominator of the terms in (C.2), define

$$E_j(t) = e^{-\lambda_j m_j t}$$

and let

$$P_j(t) = p_j(t)/E_j(t) = \sum_{r=0}^{m_j-1} \frac{(\lambda_j m_j t)^r}{r!}$$

$$R_j(t) = -p_j'(t)/E_j(t) = \frac{t^{m_j-1}(\lambda_j m_j)^{m_j}}{(m_j-1)!}$$

$$Q_j(t) = q_j(t)/E_j(t) = \sum_{r=0}^{m_j-1} \left(\frac{m_j-1}{m_j}\right)\frac{(\lambda_j m_j t)^r}{r!}$$

$$S_j(t) = -q_j'(t)/E_j(t) = \lambda_j \sum_{r=0}^{m_j-1} \frac{(\lambda_j m_j t)^r}{r!} \quad .$$

Then, the density function simplifies to

$$g(t) = U(t) \sum_{j=1}^{M} \frac{w_j}{Q_j(t)} \left[P_j(t) \left(\sum_{i=1}^{M} \frac{S_i(t)}{Q_i(t)} - \frac{S_j(t)}{Q_j(t)} \right) + R_j(t) \right]. \quad (C.5)$$

The Variance of Interarrival Time

Due to the complexity of the density function, the second central moment is obtained as follows. Integration by parts gives

$$\int t^2\, g(t)\, dt = t^2\, G(t) - 2 \int t\, G(t)\, dt \quad . \tag{C.6}$$

The second term will be considered first. Let

$$W(t) = \int t\, G(t)\, dt \quad . \tag{C.7}$$

Obtaining G(t) from expression (3.9), page 23, and letting $a = \sum_{i=1}^{M} \lambda_i m_i$,

$$G(t) = 1 - \sum_{j=1}^{M} w_j\, e^{-at} \prod_{i=1}^{M} \left(\sum_{r_i=0}^{m_i-1} A_{ir_ij}\, \frac{(\lambda_i m_i t)^{r_i}}{r_i!} \right) \tag{C.8}$$

The product in (C.8) is next rewritten. Let

$$\Pi_j(t) = \prod_{i=1}^{M} \left(\sum_{r_i=0}^{m_i-1} A_{ir_ij}\, \frac{(\lambda_i m_i t)^2}{r!} \right)$$

and let

$$C_{ir_ij} = \frac{A_{ir_ij}\, (\lambda_i m_i)^{r_i}}{r_i!} \quad .$$

Then

$$\Pi_j(t) = \prod_{i=1}^{M} \sum_{r_i=0}^{m_i-1} C_{ir_ij} t^r =$$

$$= (C_{10j}\, t^0 + C_{11j}\, t + \ldots + C_{1,m_1-1,j}\, t^{m_1-1})$$

$$(C_{20j}\, t + C_{21j}\, t + \ldots C_{2,m_2-1,j}\, t^{m_2-1}) \ldots$$

$$\ldots (C_{M0j}\, t^0 + C_{m1j}\, t + \ldots + C_{M,m_M-1,j}\, t^{m_M-1}) =$$

$$= B_{0j}\, t^0 + B_{1j}\, t + \ldots + B_{Nj}\, t^N = \sum_{\ell=0}^{N} B_{\ell_j}\, t^{\ell}$$

where $N = \sum_{i=1}^{M} (m_i - 1)$,

and where each coefficient, $B_{\ell j}$, is the sum of all products, $\prod_{i=1}^{M} C_{ir_ij}$, such that $\sum_{i=1}^{M} r_i = \ell$.

Expression (C.7) can now be written

$$W(t) = \int t \left(1 - \sum_{j=1}^{M} w_j\, e^{-at} \sum_{\ell=0}^{N} B_{\ell j}\, t^{\ell}\right) dt =$$

$$= \frac{t^2}{2} - \sum_{j=1}^{M} w_j \sum_{\ell=0}^{N} B_{\ell j} \int e^{-at}\, t^{\ell+1}\, dt \quad . \qquad (C.9)$$

The integral in (C.6) can now be evaluated:

$$\int t^2 g(t)\, dt = t^2 G(t) - t^2 + 2 \sum_{j=1}^{M} w_j \sum_{\ell=0}^{N} B_{\ell j} \int e^{-at} t^{\ell+1}\, dt \quad . \quad (C.10)$$

Applying ℓ'Hospital's rule readily shows that

$$\lim_{t\to\infty} t^2[G(t) - 1] = 0 \quad .$$

The second central moment is, therefore

$$E(t)^2 = \int_0^{\infty} t^2 g(t)\, dt = 2 \sum_{j=1}^{M} w_j \sum_{\ell=0}^{N} B_{\ell j} \int_0^{\infty} e^{-at} t^{\ell+1}\, dt =$$

$$= 2 \sum_{j=1}^{M} w_j \sum_{\ell=0}^{N} \frac{B_{\ell j}(\ell+1)!}{a^{\ell+2}} \quad . \quad (C.11)$$

The variance of interarrival time is thus

$$\sigma^2(t) = 2 \sum_{j=1}^{M} w_j \sum_{\ell=0}^{N} \frac{B_{\ell j}(\ell+1)!}{\left(\sum_{i=1}^{M} \lambda_i m_i\right)^{\ell+2}} - \frac{1}{\left(\sum_{i=1}^{M} \lambda_i\right)^2} \quad . \quad (C.12)$$

<u>A Limit Relative to Processes with Small Variances</u>

To be shown is

$$\lim_{m\to\infty} e^{-\frac{mt}{2}} \sum_{r=0}^{m-n} \frac{\left(\frac{mt}{2}\right)^r}{r!} = \begin{cases} 0, & t/2 > 1 \\ 1/2, & t/2 = 1 \\ 1, & t/2 < 1 \end{cases} \quad (C.13)$$

with $0 < t < \infty$,

and n, a finite integer.

Expression (C.13) is a Poisson distribution function with parameter $(\frac{mt}{2})$. The limiting distribution as m approaches infinity of the random variable

$$y = \frac{r - \frac{mt}{2}}{\sqrt{\frac{mt}{2}}}$$

is normal with mean zero and variance one. Thus

$$\lim_{m\to\infty} \sum_{r=0}^{m-n} \frac{e^{-\frac{mt}{2}} (\frac{mt}{2})^r}{r!} = \lim_{m\to\infty} Pr(r < m - n+1) =$$

$$= \lim_{m\to\infty} Pr(y < \frac{\sqrt{m}\ (1 - t/2)}{\sqrt{t/2}} - \frac{n-1}{\sqrt{mt/2}}) \quad ,$$

which gives the result in (C.13).

APPENDIX D

DEVELOPMENT RELATED TO INTERDEPARTURE TIME DISTRIBUTIONS

Some Solution Procedures Used in the Interdeparture Interval Expressions

A recurring problem in the development of interdeparture time distributions is the manipulation of expressions involving terms of the form

$$\phi(t) = e^{-at} \sum_{i=0}^{n} C_i \, t^i \quad . \tag{D.1}$$

The following problem, although simpler than most, is illustrative:

$$\text{Evaluate} \qquad \int_{t_1=0}^{\infty} \int_{t_2=t_1}^{\infty} t_1 \, f_1(t_1) \, f_2(t_2) \, dt_2 \, dt_1$$

where

$$f_1(t_1) = e^{-at_1} \sum_{i=0}^{n_1} C_i \, t_1^i \quad ,$$

$$f_2(t_2) = e^{-bt_2} \sum_{i=0}^{n_2} D_i \, t_2^i \quad .$$

Solution of such problems is simplified by the fact that the functional form of the terms, $\phi(t)$, is preserved by both multiplication and integration. Thus a few subroutines can be written to manipulate the constants of the expressions during these operations. For the various calculations encountered, the following evaluations are required:

Summation simplifications

$$\text{Type 1: } \sum_{i=0}^{n} C_i \sum_{r=0}^{i+k} D_r\, t^r = \sum_{i=0}^{n+k} E_i\, t^i$$

$$\text{where} \quad E_i = D_i \sum_{j=i-k}^{n} C_j$$

$$\text{and} \quad E_i = D_i \sum_{j=0}^{k-1} C_j \quad \text{if} \quad k > 0 \quad .$$

$$\text{Type 2: } \sum_{i=0}^{n} C_i \sum_{r=0}^{i} D_{ir}\, t^r = \sum_{i=0}^{n} E_i\, t^i$$

$$\text{where} \quad E_i = \sum_{r=i}^{n} C_r\, D_{ri} \quad .$$

Multiplication

$$(e^{-at} \sum_{i=0}^{n_1} C_i\, t^i)(e^{-bt} \sum_{j=0}^{n_2} D_j\, t^j) = e^{-(a+b)t} \sum_{k=0}^{n_1+n_2} E_k\, t^k$$

$$\text{where} \quad E_k = \sum_{\substack{\text{all} \\ i\&j}} C_i\, D_j \text{ such that } i+j=k.$$

Integration, upper limit a variable

$$\int_0^t e^{-ay} \sum_{i=0}^{n} C_i\, y^{i+k}\, dy = K - e^{-at} \sum_{i=0}^{n+k} D_i\, t^i$$

where $K = \sum_{i=0}^{n} \frac{C_i(k+1)!}{a^{k+i+1}}$

and D_i is obtained from the Type 1 summation simplification of

$$\sum_{i=0}^{n} \frac{C_i(k+i)!}{a^{k+i+1}} \sum_{r=0}^{k+i} \frac{a^r}{r!} \quad .$$

Integration, lower limit a variable

$$\int_t^{\infty} e^{-ay} \sum_{i=0}^{n} C_i\, y^{i+k}\, dy = e^{-at} \sum_{i=0}^{n+k} D_i\, t^i$$

where D_i is the same as the preceding case.

Additionally, subroutines providing evaluation of integrals over constant limits are written for

$$\int_{\alpha}^{\beta} e^{-at} \sum_{i=0}^{n} C_i\, t^{i+p}\, dt$$

with limits of

α	β
0	τ
$-\infty$	τ
0	∞

where τ is a positive number.

The sample problem can now be evaluated:

$$\int_0^\infty \int_{t_1}^\infty t_1 \, f_1(t_1) \, f_2(t_2) \cdot dt_2 \, dt_1 =$$

$$= \int_0^\infty t_1 \, f_1(t_1) \int_{t_1}^\infty e^{-b t_2} \sum_{i=0}^{n_2} D_i \, t_2^i \, dt_2 \, dt_1$$

$$= \int_0^\infty t_1 (e^{-at_1} \sum_{i=0}^{n_1} C_i \, t_1^i)(e^{-bt_1} \sum_{i=0}^{n_2} E_i \, t_1^i) dt \quad ,$$

where E_i results from the integration subroutine with variable lower limit. The integral then reduces, by application of the multiplication subroutine, to

$$\int_0^\infty e^{-(a+b)t_1} \sum_{i=0}^{n_1+n_2} F_i \, t_1^{i+1} \, dt_1 \quad ,$$

which is then evaluated by the integration routines with appropriate limits.

System State Probabilities at Regeneration Points in $E_m/E_k/2$ Queues

The state probabilities at regeneration points do not, in general, correspond to the long run state probabilities of the system. The interdeparture time distribution requires, for two-server systems, the probabilities for states zero and 1. Define

P_i : the probability of being in state i, i=0,1,2 at the regeneration point; i=2 is interpreted to include all states greater than 1.

E_i : the mean time until the next departure, given the system is in state i at the regeneration point,

P_{i0} : the transition probability from state i at one regeneration point to state zero at the next regeneration point,

λ, μ : mean arrival and service rates.

Three equations can be written in the P_i, the first based on conditional expectations of interdeparture times, the second from the transition probabilities, and the sum of the probabilities provides the third:

$$P_0 E_0 + P_1 E_1 + P_2 E_2 = 1/\lambda$$

$$P_0 P_{00} + P_1 P_{10} + P_2 P_{20} = P_0$$

$$P_0 + P_1 + P_2 = 1 \quad . \tag{D.2}$$

Evaluation of the E_i and p_{ij} will require the additional random variables:

t_a : the interarrival time,

t_{ar} : the residual waiting time until an arrival,

t_s : the service time,

t_{sr} : the residual service time.

The expected interdeparture time, given system state zero, will now be found. Obviously, no departure can occur until there is an arrival to the system. That arrival immediately enters service in either of the empty channels; the service channel entered will be designated Channel 1. The elapsed time from arrival to departure of this first customer, t_1, is obviously the service time:

$$t_1 = t_s$$

A second customer may also arrive and be served in Channel 2. The time from the first customer's arrival to the second customer's service completion, t_2, is

$$t_2 = t_a + t_s$$

The density functions of t_1 and t_2 are

$$f_1(t_1) = \frac{(\mu k)^k \, t_1^{k-1} \, e^{-\mu k t_1}}{(k-1)!} \qquad 0 < t_1 < \infty \quad ,$$

$$f_2(t_2) = (\mu k)^k \, (\lambda m)^m \left[e^{-\mu k t_2} \sum_{i=0}^{k-1} \frac{K_{1(i+1)}}{i!} t_2^i \right.$$

$$\left. + e^{-\lambda m t_2} \sum_{i=0}^{m-1} \frac{K_{2(i+1)}}{i!} t_2^i \right] \qquad 0 < t_2 < \infty \quad ,$$

where

$$K_{1i} = \frac{(-1)^{k-i}\binom{m+k-i-1}{m-1}}{(\lambda m - \mu k)^{m+k-i}} \quad ,$$

$$K_{2i} = \frac{(-1)^{m-i}\binom{m+k-i-1}{k-1}}{(\mu k - \lambda m)^{m+k-i}} \qquad \mu k \neq \lambda m \quad .$$

The expression for E_0 is

$$E_0 = Pr(t_1 < t_2)\, E(t_1 \mid t_1 < t_2) + Pr(t_2 < t_1)\, E(t_2 \mid t_2 < t_1)$$
$$+ E(\text{time to the first arrival}) \quad . \tag{D.3}$$

The random variables t_1 and t_2 are independent. Thus

$$Pr(t_1 < t_2)\, E(t_1 \mid t_1 < t_2) =$$

$$\int_{t_1=0}^{\infty} t_1\, f_1(t_1) \int_{t_2 = t_1}^{\infty} f_2(t_2)\, dt_2\, dt_1 \quad . \tag{D.4}$$

Expression (D.4) is easily evaluated using the method discussed in the preceding section. A similar expression for the second term of (D.3) can be written.

The last term in (D.3) is divided into two cases: a prior system state of zero and a prior nonzero state. The term, which must be consistent with the approximation of the interdeparture time distribution, is

E(time to the first arrival) =

$$p_{00}\, E(t_a - t_s \mid t_a > t_s) + (1 - p_{00})\, E(t_{ar}) =$$

$$= E(t_a - t_s) + (1 - p_{00})\, E(t_{ar}) \quad .$$

Now, $E(t_a - t_s)$ can be obtained from expression (4.5), page 51 and $E(t_{ar})$ is given in expression (B.5) page 109.

The value of E_1 is found in a similar manner. In this case, one customer will have partially completed service at the regeneration point. The time until completion of service, t_1, is distributed by the residual service time distribution. The time until a customer can complete service in the empty channel is the sum of the residual arrival time plus a complete service. Thus,

$$t_2 = t_{ar} + t_s \quad .$$

The distribution of t_2 is given in expression (4.10), page 53.
The expression for E_1 is, then

$$E_1 = Pr(t_1<t_2)\, E(t_1|t_1<t_2) + Pr(t_2<t_1)\, E(t_2|t_2<t_1) \quad . \qquad (D.5)$$

The terms are evaluated as in (D.4).

Finally, E_2 is found by noting that, at the regeneration point, one customer will be partially through service and another will be entering service. Thus

$$t_1 = t_{sr} \quad \text{and}$$

$$t_2 = t_s \quad .$$

The transition probabilities, p_{i0}, will now be considered. First, if the system is in state zero at a regeneration point, no new departure can occur until the next arrival. When that arrival occurs, it immediately enters service. Simultaneously, the next customer enters the arrival process. It can be seen that the only way for the system to again be in state zero at the next regeneration point is for the first customer to complete service before the second customer arrives. Thus

$$p_{00} = Pr(t_s < t_a) = \int_{t_a=0}^{\infty} \int_{t_s=0}^{t_a} b(t_s)\, a(t_a)\, dt_s\, dt_a \qquad (D.6)$$

If the system is in state 1 at the regeneration point, one customer will have partially completed service and another will be partially through the arrival process. To be in state zero at the next regeneration point,

$$p_{10} = Pr(t_{sr} < t_{ar}) \quad . \qquad (D.7)$$

Finally, it is apparent that, if the system is in state 2 (or greater) at a regeneration point, it cannot be in state zero at the next. Thus

$$p_{20} = 0 \quad . \qquad (D.8)$$

All coefficients in the set of equations (D.2) are now specified.

APPENDIX E

VALIDATION OF SIMULATION EXPERIMENTS

The simulation experiments reported in this paper were performed on an IBM model 360/195 computer using the General Purpose Simulation System language (GPSS Version 2). From initial experiments it was determined that simulation of 150,000 arrivals provided a reasonable balance between accuracy and cost. Over a wide variety of queue configurations, the 95% confidence limits of the mean queue length were generally about 7% of the mean. Computation time (central processing unit time) was typically about one minute per simulated service system. The first analysis of accuracy illustrated below involves a simple Erlang queuing system. Additional examples are given for the experiments of Chapters III and IV.

Simple Queue Simulation

In this example, an $M/E_3/2$ system with traffic intensity of 0.9 was simulated. Thirty independent samples of 5000 arrivals each were taken. For each of the samples, the queue was randomly initialized from the distribution of system states given by Hillier and Lo.

In addition to the mean queue length, the realization of traffic intensity for a sample is a random variable. Results for the 30 samples are given in Table 12.

A linear regression analysis, using traffic intensity as the independent variable, was then used to determine the variance of the mean queue length observations. The results were as follows:

Table 12. Results of simple queue simulation.

Sample No. (i)	Traffic Intensity (P_i)	Mean Queue ($\bar{Q}_i$)	Sample No. (i)	Traffic Intensity (P_i)	Mean Queue ($\bar{Q}_i$)
1	.891	5.254	16	.899	4.123
2	.888	4.742	17	.892	7.168
3	.896	5.335	18	.890	3.654
4	.891	5.585	19	.893	6.481
5	.888	3.779	20	.898	5.397
6	.888	4.195	21	.893	4.575
7	.891	4.197	22	.892	4.598
8	.896	4.782	23	.894	4.210
9	.890	4.143	24	.894	3.682
10	.892	5.318	25	.894	4.278
11	.893	3.825	26	.895	4.834
12	.891	4.344	27	.887	5.693
13	.890	4.370	28	.894	6.818
14	.886	4.942	29	.893	4.762
15	.896	4.869	30	.897	4.427

Grand mean queue length: 4.813

Grand mean traffic intensity: 0.892

Variance due to regression: 0.081

Experimental variance: 0.800

For the grand mean, 95% confidence limits are now obtained from the t distribution with 29 degrees of freedom:

$$\text{limit} = 2.04 \left(\frac{.800}{30}\right)^{\frac{1}{2}} = 0.333 \quad .$$

With reasonable confidence, then, the true mean lies within about 7.96% of the simulation grand mean. (The theoretical figure for mean queue length interpolated from Hillier and Lo's tables is about 4.6.)

Two Tandem Queues

A similar analysis was made using two queues in series--the type of experiment performed in Chapter IV. The first queue was $M/E_2/2$; the second had two servers with E_3 service time distribution. Both traffic intensities were 0.9. This, and most other, experiments differed from the one discussed above in that 15 samples of 10,000 arrivals each were simulated. Results are given in Table 13.

Superposition Experiments

Several experiments involving the superposition of input processes will now be examined. The cases shown were chosen to include the more extreme parameter variations. Throughout these experiments, 15 samples of 10,000 arrivals each were used. Table 14 gives the results. As may be expected, accuracy decreases as traffic intensity,

Table 13. Analysis of two series queue experiment.

	Queue 1	Queue 2
Grand Mean Queue Length	5.699	4.011
Variance Due to Regression	5.462	5.357
Experimental Variance	0.205	0.148
95% Confidence Interval	0.250	0.213
Confidence Interval as a Percent of Grand Mean	4.39	5.30
Theoretical Mean	5.80	N/A

Table 14. Analysis of superposition results.

	Case (See Note)					
	1	2	3	4	5	6
Grand Mean Queue Length	3.879	18.046	0.585	2.062	2.504	4.101
Grand Mean Traffic Intensity	0.901	0.978	0.698	0.900	0.901	0.898
Variance Due to Regression	0.300	27.215	0.006	0.985	0.367	3.733
Experimental Variance	0.243	33.438	0.001	0.043	0.038	0.263
95% Confidence Interval	0.273	3.195	0.016	0.114	0.107	0.283
Conf. Int. as Percent of Mean	7.03	17.7	2.76	5.54	4.28	6.91

Note: Case 1--base case; two E_2 input streams to an $E_2/2$ service system with traffic intensity 0.9.

Case 2--high traffic intensity (0.98).

Case 3--low traffic intensity (0.70).

Case 4--high input shape parameter (15).

Case 5--high service shape parameter (8).

Case 6--many input streams (20).

and hence the variance of queue length, increases. Daley [22, 1968], using GI/M/1 systems, determined that the number of arrivals required to obtain a given accuracy of mean queue length was proportional to $(1 - \rho)^{-2}$. This would suggest that 25 times as long a simulation run would be required for $\rho = .98$ as for $\rho = .90$ for comparable accuracy.

VITA

Robert Allen McLaren was born in Muleshoe, Texas on September 16, 1938. He attended elementary and high schools in Southern California, graduating in 1956. The Vanderbilt University School of Engineering was then entered, leading to a Bachelor of Engineering degree in Electrical Engineering in June, 1960. The following fall, he entered the Graduate School of Industrial Administration at the Carnegie Institute of Technology (now Carnegie-Mellon University). The two-year program led to a Master of Science degree in 1962.

Mr. McLaren was commissioned in the United States Air Force in November, 1962, and served as an engineer in a guided missile support organization for three years. He left active duty in 1965 and has been employed since then by the Nuclear Division of Union Carbide Corporation at Oak Ridge, Tennessee. He is currently head of an Operations Research Section.

While employed at Oak Ridge, he has been pursuing further graduate study at the University of Tennessee, completing the Doctor of Philosophy degree with a major in Management Science in December, 1974.

Mr. McLaren is married to the former Kathleen Stem of Nashville, Tennessee. They have a daughter, Joyce Anne, age two.