(FINAL REPORT)

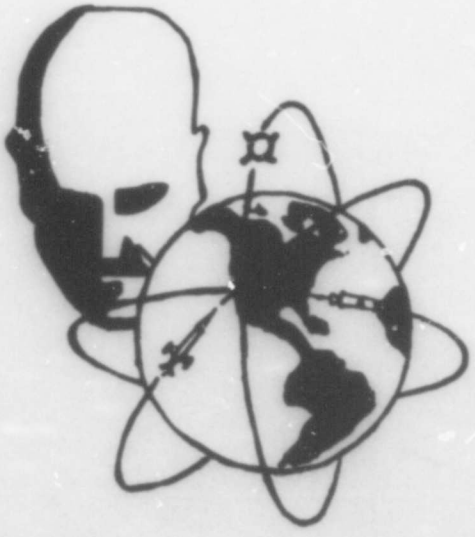
APPLICATION OF QUEUING THEORY TO INFORMATION SYSTEMS DESIGN

TECHNICAL DOCUMENTARY REPORT NO. ESD-TDR-64-428

JUNE 1964

David Nee

DIRECTORATE OF COMPUTERS
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts



(Prepared under Contract No. AF 19 (628)-2901 by the Stanford Research Institute, Menlo Park, California.)

APPLICATION OF QUEUING THEORY TO INFORMATION SYSTEMS DESIGN

Abstract

This research project was undertaken to investigate the application of queuing theory, in particular the priority queuing theory, to structuring of information systems.

A scheme of classifying queuing models and queuing literature was developed, and was used to classify most of the known priority-queuing literature. The waiting-time distribution for a multi-server, head-of-the-line, priority-queuing model was developed. Priority-queuing formulas were catalogued and listed in accordance with the classification scheme. A simple single-stage priority-queuing model was constructed to demonstrate the feasibility of using such a model to augment the analytical modeling techniques in investigating more complex priority-queuing systems. Finally, procedures for the application of the queuing model to structuring of information systems were outlined.

REVIEW AND APPROVAL

This technical documentary report has been reviewed and is approved.

JOHN B. CURTIS
2d Lt., USAF
Task Scientist

KEY WORD LIST

1. INFORMATION SYSTEMS
2. MODELS
3. SIMULATION
4. DESIGN
5. STATISTICAL ANALYSIS
6. QUEUING THEORY
7. OPERATIONS RESEARCH
8. PROBABILITY

CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

FOREWORD

This final report on Application of Queuing Theory to Information Systems Design was prepared by Stanford Research Institute, Menlo Park, California on Air Force Contract AF 19(628)-2901 as SRI Project 4513. The principal investigator is D. Nee.

The author wishes to acknowledge the contribution to this project by R. Davis and H. Aggarwal for their mathematical analysis of waiting-time distributions, and to V. Sagherian for his programming of the simulation model.

I. INTRODUCTION

Many difficult problem areas appear frequently in the process of analyzing or synthesizing information systems. One area which appears with regularity deals with queuing and congestion processes. Many idealized models of these processes have been developed, analyzed, and reported in the literature. Thus far, this literature has not been organized in such a way as to make it useful to the frequent searcher.

Often the queuing problem in an information system differs significantly from the idealized models treated. It is desirable, where possible, to develop an analytical model to the problem at hand, but the approach to this development is frequently quite difficult, if not impossible. In many cases the solution to the analytical model is in such a form that it is not readily usable. However, it is usually possible to develop an approximate analytical model, whose solution is simple enough for application.

In many information systems the service discipline is that of the priority type. Since 1954 a number of investigators have contributed significantly to development of priority queuing theory. The results of their finding were published in various mathematical-statistics journals and books. Most of the contributions are for single-server service systems. The multi-server priority queuing problem is essentially unsolved. Most of the practical priority-queuing problems are of this multi-server type.

The principal objectives of the research effort described in this report are

(1) To develop a means of classifying the queuing models and literature,

(2) To organize the results of the priority queuing studies in accordance to the classification schemes,

(3) To extend the priority queuing analysis to include that of multi-server service systems, and

(4) To establish guides and procedures for the application of the priority queuing theory to structuring of information systems.

The research effort covers the period between April 15, 1963 to April 14, 1964.

In Section II, a classification scheme is developed. This is a generalized scheme for classifying the queuing models in accordance to their input, queue, and service characteristics. The queuing literature is classified in accordance with the models investigated and the specific performance measures studied. A simplified version of this classification scheme is used to classify all the known priority queuing literature. The analytical development of priority queuing theory is briefly reviewed in Section III. This theory was extended in this project to include the analysis of the waiting time distribution of a single-server, head-of-the-line, priority queue system (which is a corrected and modified version given by Kestern and Runnenburg) and of a multi-server, head-of-the-line, priority queue system. Detailed derivations of these two waiting time distributions are given in Appendices A and B. Some of the useful results of priority queuing analysis are presented in Section III-C for reference. The digital priority-queuing simulator is briefly described in Section IV. Some examples illustrate the use of a small simulator as an experimental tool to augment the analytical queuing analysis for obtaining approximate solutions to some specific queuing problems. The procedures for applying the priority queuing models to structuring of information systems are discussed in Section V.

II THE CLASSIFICATION OF QUEUING MODELS AND QUEUING LITERATURE

A. INTRODUCTION

The development of queuing theory may be traced back to A. K. Erlang who, in 1909, published a paper called "The Theory of Probabilities and Telephone Conversation."[1]* Since then, hundreds of papers and several books on or related to queuing theory have been published in many languages. The writer knows of no previous systematic scheme for classifying and indexing this literature. It is felt that such a scheme is needed to assist investigators or users of queuing theory in determining what aspect of queuing has (or has not) been investigated, and where to find the information that is needed in solving particular queuing problems.

In 1953, Kendall[2] introduced a basic scheme of classifying queuing systems, together with some symbols of representation. In this scheme, a queuing system is classified by:

(1) Input--the inter-arrival times distribution

(2) Queue discipline--the rules that govern the order in which waiting customers are served

(3) Service mechanism--number of servers and the service-time distribution.

Four types of distribution, F(t), were used in the classification:

---

*References are listed at the end of the report.

(1) The regular, or constant, distribution, D--The time interval between successive events of this distribution is the same. Analytically, this distribution is expressed as

$$F(t) = 0 \qquad \text{for } t < a$$
$$F(t) = 1 \qquad \text{for } t \geq a \quad . \tag{1}$$

(2) The Poissonian, or negative exponential, distribution, M--The time interval between successive arrivals and the time interval required to serve successive customers are "at random." The analytical expression is

$$F(t) = 1 - e^{-at} \quad . \tag{2}$$

(3) The Erlangian distribution, E_k--This is an intermediate distribution between the regular and the Poissonian distribution, and its analytical expression is

$$dF(t) = \frac{(ka)^k}{\Gamma(k)} e^{-kat} t^{k-1} dt \quad . \tag{3}$$

Note that Eq. (3) becomes Eq. (2) when k = 1, and approaches Eq. (1) when k tends to infinity.

(4) The general distribution, G--This is to identify an arbitrary distribution whose analytical expression is given by an arbitrary function, F(t). When successive time periods, either the inter-arrival or the service-time periods, are statistically independent, it is termed a general independent distribution, GI.

Kendall in his study of queuing systems was mainly concerned with the ordered, the "first come, first served", queue discipline. Thus, his symbolic representation of a queuing model omits the queue discipline. The format of the symbolic representation is

X/Y/Z

where X stands for the arrival distributions

Y stands for the service distributions

Z stands for the number of servers.

For example, consider a queuing model with Poissonian arrival distribution, constant service-time distributions, and three servers; this is denoted by M/D/3.

The above scheme advanced by Kendall is a very convenient shorthand notation for describing a specific class of queuing models and has been generally accepted. However, this scheme cannot describe complex queuing models. For example, it cannot describe the characteristics of the sources of input beyond the distribution of inter-arrival times, and it cannot describe the various queue-disciplines.

Saaty[3] presented a rather comprehensive description of queuing systems. He divided a queuing operation into four parts: the input, the waiting line, the service facility, and the output. With each of these is associated a set of alternative assumptions concerning the queuing process. For example, the waiting line is associated with the queue-discipline assumption and with the characteristics of the queues and of the customers in the queues. The queue may have a fixed length, and the customers may become

impatient and leave the queue. Because some of the alternative assumptions mentioned in Saaty's work are not likely to occur in normal information systems, they have been omitted here, and only likely assumptions are considered.

Saaty[4] also introduced the concept of the queue graph, which is somewhat analogous to the schematic representation of an electrical network. If this concept can be fully developed it might provide a convenient graphical representation for various classes of queuing models. Much development seems to be required before it can be of practical use. Consequently, the idea of using a queue graph to symbolically represent a queuing system is not pursued in this paper.

In this paper Kendall's classification scheme is modified and expanded to adapt it to classifying basic queuing models that are commonly found in information systems. One such basic queuing model is that of the single-stage-queue shown in Fig. II-1. Here all inputs enter a "queue," are served by a "service" organization, and finally exit from the system.

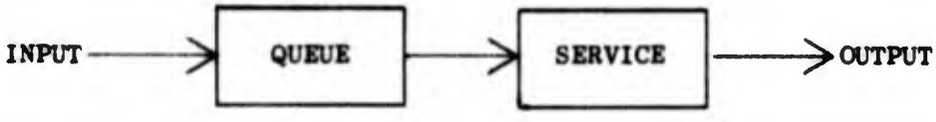


Fig. II-1

SINGLE-STAGE-QUEUE QUEUING MODEL

The model of a queuing system would be a combination of the basic queuing models. A basic queuing model can be classified according to its input, its queue, and its service characteristics. In Part C, these three primary

characterizations are explained and secondary classification is introduced within each of the primary classifications, together with a set of symbolic representations.

An investigator of a queuing model may study any one of the many measures of effectiveness for the model. Some examples of the more commonly studied measures are:

(1) The probability that a customer finds the service system free upon his arrival

(2) The probability of a customer's having to wait longer than a specified time

(3) The average (mean) waiting time of a customer

(4) The probability that the queue exceeds a certain length

(5) The average size of queue.

An investigator may study other measures, unrelated to the effectiveness of the model operation. One such measure is the interdeparture intervals of a queuing model. This measure is of special significance in the study of the tandem-queues model, where the output of one queue is the input to the next queue. The term "operational" measure is used in this section to cover both the "effectiveness" and "other" measures. Some of the more commonly used measures will be defined in more detail in Part D.

Queuing literature can usually be classified according to the queuing model or models considered, and according to the specific operational measure or measures investigated. Part E contains such a scheme of classification, which is applied to priority queuing literature.

B. BASIC QUEUING MODELS

A basic queuing model is defined as a queuing model that contains a single queuing stage (Fig. II-1). A queuing system model is a combination of basic queuing models. A basic queuing model can be classified according to its input, its queue, and its service characteristics. The symbols used to represent the characteristics are, in most cases arbitrary; however, the symbols for distributions are those used by Kendall. In this paper a unit of input is called a customer and a service channel is called a server.

1. Input

The input to a basic queuing model can be described in terms of the source of the customers, in terms of the form customers take, and in terms of the policy customers follow for queuing.

Customers could come from a population source that is limited, in which case the maximum number of customers from this source in queue at any one time is the size of the source. When the number of customers in queue from this source reaches the limit, the source cannot generate any further customers until the number in queue drops back below the limit. The population source could be unlimited, in which case the source will continue to generate customers regardless of the number of customers in the queue. The population source can also be characterized by the homogeneity of the population. A homogeneous population is one whose customers are not distinguishable from each other except by their order of arrival at the queue. A heterogeneous population is one whose customers are distinguishable into classes. Customers of a class may be distinguished from customers of another class by their input characteristics, by the queue characteristics,

by the service requirements, or by any combination of these. The input characteristics include the arrival pattern, the arrival distribution, and the arrival policy. The queue characteristics include the queue discipline and the queue policy, which will be explained in Part B-2. The service requirement includes the type of server required, the service-time distribution, and the service pattern, which will be explained in Part B-3.

Customers may arrive singly or in batches (groups). Their inter-arrival intervals may be distributed according to the four distributions defined by Kendall; the regular, the Poissonian, the Erlangian, and the general or general independent.

A customer may decide to join the queue regardless of the condition of the queue. In telephone traffic language this type of customer is called delayed-customer. A customer may decide upon arrival that the condition of the queue is such that he may not (or cannot) join it immediately. If the customer decides to try to join the queue at a later time, it is called repeat customer. Should the customer decide not to try, it is called lost customer. The many other possible types of policy are grouped under others. Table I is a listing of the input classifications together with their assigned symbols.

2. Queue

Arriving customers who decide to wait in line are said to have joined the queue. Physically the queue may take on the form of a single file or a multiple parallel file. The single-file form is sometimes called single-queue and the multi-file form is called multi-queue. The buffer storage system in some data handling systems is a multi-queue system which

Table I

INPUT CLASSIFICATION

Input Classification Parameters		Symbols
Population Size	Limited	L
	Unlimited	U
Population Characteristics	Homogeneous	I
	Heterogeneous	S
Pattern	Singly	I
	Batches	S
Distribution	Poissonian	M
	Regular	D
	Erlangian	E
	General/General Independent	G,GI
Policy	Delayed	D
	Repeat	R
	Lost	L
	Others	O

may contain several individual buffers (queues), one for each input device. Customers (information) generated by the input devices are filed into their respective queues. The size of queue may be limited by the physical facility to accommodate the customers. For example, in an electronic

message switching center, the queue size is limited by the size of magnetic core or drum or the disk or tape unit(s) employed. When the queue is reasonably larger than the traffic load expected, for all practical purposes it may be considered to be unlimited.

The rules that govern the order in which a queue is selected for service, in the case of multi-queues, and the order in which customers in a queue are selected for service is called queue discipline. Queue discipline can take many forms. In this paper only those forms that are most likely to occur in information systems are considered. The various queue disciplines will be discussed in the order of increasing complexity. The order of complexity is directly associated with the homogeneity of the population source, the number of queues, and the homogeneity of the servers. The simplest queue discipline occurs when the population source is homogeneous, and there is only one queue and one class of servers. The three most commonly discussed disciplines are:

(1) Ordered queue--Customers are given service according to their order of arrival in the queue. Another term for the ordered queue is first-come-first-served.

(2) Inverse-ordered queue--In this case the customer at the end of the queue is given service when a server becomes available. A common name for this discipline is "last come, first served."

(3) Random order queue--When a server becomes available, the next customer to be served is selected at random from the queue.

The next degree of complexity occurs when the population source is heterogeneous, and there is a single queue and a single server. The customer in queue can now be identified by class. One of the most common way of classifying customers is by their urgency level. Thus the population can be classified into priority classes, according to their urgency level. The urgency level of a class of population may remain static regardless of the changing condition of the service system, which is called static priority or just priority. The urgency level may change as a function of the condition of the service system. For example, the urgency of a customer may be given a higher level as waiting-time increases; this is termed dynamic priority. The rule that governs the order by which a priority customer is placed in service and the manner in which a displaced customer resumes service is called priority queue discipline. The three principal variations of priority queue discipline are:

(1) Head-of-the-line (non-preemptive or precedence)--The next customer to be served when a server becomes available is the highest priority customer in the queue who has the earliest arrival time.

(2) Interrupt (preemptive-resume)--The arrival of a customer whose priority class is higher than that of the customer currently in the server will interrupt the service. The interrupted customer

is placed at the head of the queue, and when the service does resume it will resume from the point of interruption.

(3) Pre-emptive (pre-emptive-repeat)--This differs from the interrupt case in that the service on the interrupted customer resumes from the beginning rather than from the point of interruption. Most of the "torn-tape" message centers employ this queue discipline.

The interrupt and pre-emptive priority can take on still a further variation. This is the so called conditional pre-emptive. Under this variation the higher-priority customer will interrupt only under certain conditions. If the length of time required to finish servicing the customer to be interrupted is comparatively short, the interrupting customer will wait for the service to end before entering the server.

A further complication can occur when the population source is heterogeneous, there is a single queue, and there are several homogeneous servers. Under this situation, the rule that governs the selection among the customers who are currently in servers can take on a number of variations. Two of the variations are listed in the order of the likelihood of being adopted in the information system:

(1) Latest time--The interrupted customer is the customer with the lowest priority and the latest arrival time.

(2) Random time--The interrupted customer is selected randomly among the lowest-priority customers, if there is more than one lowest priority customer.

Obviously there can be many other variations but they are deemed to be not practical from the point of view of designing information-system equipment.

A queue discipline that is common in communication system is the so called allocation queue discipline. Under this discipline, customers of some selected classes are given service by an allocated group of servers.

Finally when the queue is of the multi-queues type, there are rules that govern the order of selecting a particular queue and the manner in which the customers in the selected queue are served. Again there can be a wide variety of order of selecting a queue, however, only a few are likely to be found in the information system and they are:

(1) Cyclic order--Queues are numbered from 1 to n. If at any instant of time the last selected queue is the rth queue and a server becomes available, the next queue to be selected is the r+1th queue. When r is the nth queue, the next queue is the first queue. Most of the "stepping switch" service systems employ this order of selection.

(2) Lowest order--This differs from the cyclic order in that the next queue to be selected is always the queue that is the lowest queue with

waiting customers. In some "stepping swtich" service systems this order of selection is used and it is definitely biased toward the lower-numbered queue.

(3) Random order--As the name implies, the next queue to be selected is selected randomly among those queues that have waiting customers.

The queue discipline of the customers in the selected queue is the same as the queue disciplines of the single queue. In addition there is a rule governing the number of customers to be served from a selected queue before the next queue is selected. Three alternative rules can be adopted:

(1) Single--Serve one customer only

(2) Static empty--Empty the entire queue, but not including those customers that arrive after the time the queue is selected.

(3) Continuously empty--Empty the entire queue including those customers that arrive during the emptying process.

Customers waiting in the queue may decide to leave the queue for some reason. This type of queue-policy is called defection. If the defected customer should decide to re-enter the queue at a later time, the policy is defection with retrial. An example of a defection-with-no-retrial policy can be found in one type of input or output buffer operation (queue) of certain information processing systems. The buffer is loaded with messages (customers) starting from the first position of the buffer

to the last position of the buffer. When the last position in the buffer is loaded, new messages are loaded starting from the first position. The new message will replace the message that was in that position. Unloading follows the same order. For any reason, the unloading rate fails to keep up with the loading rate, it is possible to displace some messages before they are unloaded. The displaced messages are termed defected messages with no retrial.

The queue classification is listed in Tables II(a) and II(b). Table II(a) is the over-all queue-classification table and Table II(b) is the queue-discipline section of Table II(a). The order of classifying queue discipline is as follows:

(1) Queue number

(2) Queue size

(3) Queue discipline: (a) Selection order

(b) Service mode

(c) Service order

(d) Priority modifiers

(4) Queue policy.

An example of queue representation is then

$$2UCE_{s}P(t)_{pcl}{}^{N}$$

which stands for a two-queue queue model. The queue size is unlimited and defection is not allowed. The queue-discipline symbol stands for cyclic selection of the queues; once a queue is selected, the content of the queue, as of the time of selection, is emptied (served) according to a dynamic-priority discipline. This priority discipline is the conditional pre-emptive type, and

the pre-empted customer is the lowest-priority customer with the latest arrival time.

Table II

(a) QUEUE CLASSIFICATION

Queue-Classifications Parameters			Symbol
Number	Single		1
	Multiple		S
Size	Limited		L
	Unlimited		U
Queue-Discipline	See Table II(a)		
Policy	Non-defection		N
	Defection	No retrial	D
		Retrial	D_R

Table II (Cont.)

(b) QUEUE-DISCIPLINE CLASSIFICATION

	Queue-Discipline Parameters		Symbols
Service Order	Ordered (FCFS)		0
	Inverse Ordered (LCFS)		I
	Random		M
	Priority		P
	Allocation		A
Priority Modifiers	1st level	Static	
		Dynamic	(t)
	2nd level	Head-of-the-line	h
		Interrupt	i
		Pre-emption	p
	3rd level	Conditional	c
		Unconditional	u
	4th level	Latest time	l
		Random time	m
Multi-Queues	Selection Order	Cyclic	C
		Lowest	L
		Random	M
	Service Mode	Single	1
		Static Empty	E_s
		Dynamic Empty	E_d

3. Service

Customers in queues are serviced by a service organization. This service organization may contain one or more servers. Where there is more than one server, the servers may be arranged in serial, parallel, or serial-parallel fashion. It is assumed that no queue can exist within the service organization. If these servers have identical service characteristics and can serve any customer in the queue, they are called homogeneous servers. The most commonly-assumed service-time distributions are the negative-exponential, the constant, and the Erlangian. In some analytical analyses, the distribution is assumed to be a general function, F(t). The server may serve the customers singly or in batches. The service classification is listed in Table III.

Table III

SERVICE CLASSIFICATION

Service Classification Parameters		Symbol
Organization	Single	1
	Multiple-Serial	S_S
	-Parallel	S_P
	-Serial-parallel	S_M
Server Characteristics	Homogeneous	1
	Heterogeneous	S
Service Time Distribution	Negative Exponential	M
	Constant	D
	Erlang	E_k
	General	G
Service Pattern	Singly	1
	In Batches	S

C. OPERATIONAL MEASURES

The quality of a queueing-system performance can be expressed in terms of its operational measures. Some of the most commonly used measures are:

(1) Queue length--The queue length can be defined in two ways:

(a) The number of customers in the queue

(b) The number of customers in the queue and in the server.

(2) Waiting time--The time that a customer spends waiting in the queue; that is, the time elapsed between the time of his arrival at the queue and the time of his initial entry into the service organization.

(3) In-service time--The time elapsed between the time of his initial entry into the service organization and the time of his final departure from the service organization. (In the case of non-priority and the head-of-the-line priority discipline, this time is simply the service time of that particular customer.)

(4) Delay time--It is the sum total of the waiting time and the in-service time.

(5) Busy period--The busy period of a service organization is the elapsed time between the time when the service organization becomes fully occupied (busy) and the time when it ceases to be fully occupied. An example of a busy period is that due to the presence of higher priority customers in the queue(s) and in the server(s).

The busy period ends when higher-priority customers in the queue have been serviced and the service organization is again available to serve lower-priority customers.

(6) Output time--This is the interdeparture time interval of the customer.

(7) There are a number of less-often-used measures:

(1) The queue length per server

(2) The number of idle servers per total number of server

(3) Time elapsed between time when the number of customers in a queue exceeds a prescribed threshold

(4) The number of customers served during a busy period.

Table IV is a listing of these operational measures together with their assigned symbol. The less-often-used measures are grouped under "Others."

Table IV

OPERATIONAL MEASURES

Operational Measures	Symbol
Queue Length: In Queue(s)	L_q
In Queue(s) and Server(s)	L
Waiting Time	W
In-Service Time	T
Delay Time	D
Busy Period	B
Interdeparture Time	I
Others	O

To obtain a complete understanding of these measures, one would like to obtain closed analytical expressions for the probability density or the probability distributions of these measures. However, in many cases, especially in the more complex queuing models, closed analytical expressions are too clumsy to be of practical use or are not easily obtainable. In these latter cases the results of investigation are usually left in implicit form, in the form of the generating function (g.f.) or in the Laplace transform (L.T.) or in the Laplace-Stieltjes transform (L.S.T) of the desired probability function. It is possible, though, to obtain the first few moments, especially the first two moments, of the probability function from the g.f. or the transform. The symbolic representation for the results of the queuing investigation on operational measure X are:

(1) E(X)--the expected or the mean of X

(2) $E(X^n)$--the nth moment of X

(3) p(X=a)--the probability density of X, the probability that X takes on the value of a

(4) P(X≥a)--the probability distribution function of X, the probability that X is greater than a, $P(X \geq a) = \int_a^\infty p(X=a)dX$.

(5) G(Y)--the generating function of Y, where Y is the probability function

(6) $\tilde{Y}$--the L.T. or L.S.T. of Y.

For example, the notation $P(\tilde{W})$ is the transform of the probability distribution of the waiting time, and $G(p(L_q))$ is the g.f. of the probability function of the number of customers in the queue.

All the above operational measures may be investigated for either the steady-state or the transient situation. A small letter (t) is attached to the operational measure symbol to denote the transient case; for example, $P(L_q(t))$ means the probability distribution of queue length in the queue for the transient case.

D. QUEUING LITERATURE CLASSIFICATION SCHEME

Queuing literature can be classified according to the queuing model or models considered and the specific aspect of the operational measures investigated. In this section a general scheme of classification is developed and a simplified version of the general scheme is used to classify some priority queuing literature.

The format of the queue model classification is that used by Kendall, except that the three compartments are used to represent the characteristics of the input, the queue, and the service respectively.

Thus, the format is,

$$X/Y/Z$$

where X is the input characteristic

Y is the queue characteristic

Z is the service characteristic.

The format for a piece of queuing literature is $X/Y/Z\text{-}O_p$, where O_p stands for the operational measure or measures investigated.

Let us apply this scheme to a particular piece of queuing literature that considers a 'basic' queuing model with the following characteristics:

(1) Input--The source population is unlimited and contains several priority classes of customer. Customers from this population source arrive singly according to Erlangian distribution. Upon arrival the customer chooses to join the queue regardless of the condition of the queue.

(2) Queue--The queue contains two distinct queues and their size is unlimited. The queue is selected according to cyclic order and the content of the selected queue is completely served according to a dynamic priority discipline. The dynamic priority discipline is the conditional

pre-emptive type and the pre-empted customer is the lowest priority customer with the latest arrival time who are being serviced. The customer does not defect from the queue.

(3) Service--The service organization contains three parallel servers of the homogeneous type. Their service time distribution is constant. Customers are served in batches.

This particular article investigated the following operational measures in steady-state:

(1) The queue length distribution of each priority class

(2) The mean waiting time of each priority class

(3) The transform of the waiting time of each priority class.

The article, then can be symbolically represented by

US1E D/2UCE$_s$P(t)$_{pc1}$N/3$_p$1DS-P(L), E(W), $P\tilde{(W)}$.

It is seen that this generalized scheme of classification can be rather unwieldy. The amount of queuing literature existing today does not warrant such a detailed classification. Thus, a simpler version of the general scheme is used in this paper to classify the queuing literature, in particular the literature on priority queuing. This simplified scheme took into account the fact that certain input, queue, and service characteristics seldom exist in the queuing models that can be analytically analyzed while others occur in most information systems.

In this simplified scheme a queue model is classified by the pattern and the distributions of the input, the number of the discipline of the queue, and the number and service distribution of the server. Under the queue discipline, only the service order and the first and second level of priority modifiers are included. No distinction is made between the

multi-parallel and multi-serial server organization. The simplified queuing model classification parameters are shown in Table V. The format of the

Table V

SIMPLIFIED QUEUING MODEL CLASSIFICATION

Queuing Model Classification Parameters			
Primary	Secondary	Tertiary	Symbols
Input	Pattern	Singly	1
		Batches	S
	Distributions	Poissonian	M
		Regular	D
		Erlangian	E
		General/General Independent	G/GI
Queue	Number	Single	1
		Multiple	S
	Discipline	Ordered	O
		Priority-head-of-the-line	$P_h/P_h(t)$
		-interrupt	$P_i/P_i(t)$
		-pre-emption	$P_p/P_p(t)$
Service	Number	Single	1
		Multiple	S
	Distributions	Negative Exponential	M
		Constant	D
		Erlangian	E
		General	G

simplified queuing literature is given by the representation

$$X_1X_2/Y_1Y_2/Z_1Z_2\text{-}O_p$$

where X_1 is the pattern of arrival

X_2 is the arrival distributions

Y_1 is the number of queues

Y_2 is the queue discipline

Z_1 is the number of servers

Z_2 is the distribution of service time.

Thus, the simplified symbolic representation of the specified piece of queuing literature is

$$1E/2P(t)_p/3D\text{-}P(L),\ E(W),\ P\tilde{(W)}.$$

For illustration, this proposed method of classification will be applied to some priority queuing literature. Following is a list of articles published in English, together with their symbolic classification:

Serial Number	Article	Classification
1	Avi, Itzhak, B. and Naor, P., "On a Problem of Preemptive Priority Queuing," OR 9, pp. 664-672 (1961)	$1M/1P_i/1M\text{-}E(W),\ E(L_q^2)$
2	Barry, D. Y., "A Priority Queuing Problem," OR 4 pp. 385-386 (1956)	$1M/1P_i/1M\text{-}E(L_q)$
3	Burke, Paul J., "Priority Traffic with At Most One Queuing Class," OR 10, pp. 567-568 (1962)	$1M/1P_h/SM\text{-}P(W=o)$
4	Cobham, A., "Priority Assignment in Waiting Line Problem," OR 2, pp. 70-76 (1954) also OR 3, p. 547 (1955)	$1M/1P_h/1G\text{-}E(W)$ $1M/1P_h/SM\text{-}E(W)$
5	Cox, R. E.,"Traffic Flow in an Exponential Delay System with Priority Categories,"Proc IEEE, (London) 102, Pt.B, 1955, 815-818	$1M/1P_h/SM\text{-}E(W)$
6	Dressin, S. A., and Reich, E.,"Priority Assignment on a Waiting Line,"Quarterly of Applied Math. XV, 1957, 208-211	$1M/1P_h/1M\text{-}\tilde{P}(W)$

Serial Number	Article	Classification
7	Firstmann, S. I.,"Duration of a Countdown when Considered as an Interrupted Service Process," OR 11, 1963, 210-227	$1M/1P_i/1M\text{-}E(L_q)$
8	Gaver, D. P.,"A Waiting Line with Interrupted Service, Including Priorities,"J. Roy. Statist. Soc. (B) 24, 1962, 73-90	$1M/1P_i/1G\text{-}E(B)$, $E(L_q)$ $1M/1P_p/1G\text{-}E(B)$, $E(L_q)$
9	Heathcote, C. R.,"The Time-Dependent Problem for a Queue with Preemptive Priorities,"OR 7, 1959. 670-680	$1M/1P_i/1M\text{-}E(L_q(t))$
10	_____,"A Simple Queue with Several Pre-emptive Priority Classes," OR 8, 1960, 630-638	$1M/1P_i/1M\text{-}E(L_q)$
10a	_____,"Preemptive Priority Queuing,"Biometrika 48, 1960, 57-63	$1M/1P_i/1E_k\text{-}P(\tilde{L}_q)$ $1M/1P_i1D\text{-}E(L_q)$
11	Helly, Walter,"Two Doctrines for the Handling of Two-Priority Traffic by a Group of N Servers," OR 10, 1962, 268-269	$1M/1P_h/SM\text{-}P(W=0)$
12	Holly, J.,"Waiting Line Subject to Priorities," OR 2, 1954, 341-343	$1M/1P_h/1G\text{-}E(W)$
13	Jackson, J. R.,"Some Problems in Queuing with Dynamic Priorities,"UCLA Mngmt. Sci. Res. Rept 62, Nov. 1959	$1M/1P_h(t)/SM\text{-}E(W)$
14	_____,"Simulation of Queues with Dynamic Priorities,"UCLA Mngmt Sci Res. Rept. 71, Mar. 1961	$1M/1P_h(t)/SM\text{-}P(W\geq t)$
14a	_____,"Waiting-time Distribution for Queues with Dynamic Priorities,"NRLQ, 8, 1962, 31-36	$1M/1P_h(t)/1M\text{-}E(W)$
15	Jaiswal, N. Y., Preemptive Resume Priority Queue, OR 9, 1961, 732-770	$1M/1P_i/1G\text{-}E(L_q)$, $E(\tilde{L}_q(t)$, $P(\tilde{B})$.

Serial Number	Article	Classification
16	Jaiswal, N. Y.,"Time Dependent Solution of the 'Head of the Line' Priority Queue,"J. Roy. Statist. Soc. (B)24, 1962, 91-101	$1M/1P_h/1G$-$E(L_q(t))$.
17	Keilson, J.,"Queues Subject to Service Interruption," Ann Math. Statist., 33, 1962, 1314-1322	$1M/1P_i/1G$-$\tilde{P}(L)$, $\tilde{P}(B)$.
18	Kesten, H. and Runnenburg, J. Th.,"Priority in Waiting Line Problems, Koninkl, Ned, Akad, Weten-Schap," Proc. Ser A, 60, 1957, Pt I, 312-324 and Pt XI, 325-336	$1M/1P_h/1M$-$E(W^2)$, $\tilde{P}(W)$.
19	Miller, R. G.,"Priority Queues," Ann. Math. Statist. 31, 1960, 86-103	$1M/1P_h/1G$ $E(L_q^2)$, $E(T)$, $E(B^2)$ $1M/1P_i/1G$-
19a	Morse, P. M., Queues Inventories and Maintenance (Wiley, N.Y., 1958) Chap. 9.	$1M/1P_h/1M$-$E(L)$, $E(L_q)$, $E(W)$
20	Phipps, T. E.,"Machine Repairs as a Priority Waiting-Line Problem,"OR 4, 1956, 76-86	$1M/1P_h/1M$-$E(L_q)$, $E(W)$
21	Sandeman, P.,"Empirical Design of Priority Waiting,"OR 9, 1961, 446-455	$1M/1P_h/1M$-$P(W \geq t)$.
22	Stephan, F. F.,"Two Queues Under Preemptive Priority with Poisson Arrival and Service Rate," OR 6, 1958, 399-418	$1M/1P_i/1M$-$E(L_q^4)$, $E(W^3)$, $E(T^2)$
23	Thiruvengadam, K.,"Queuing with Breakdowns," OR 11, 1963, 62-71	$1M/1P_i/1G$-$E(L_q)$
24	Van der Zee, S. P.,"Priority Assignment in Waiting-Line Problems under Conditions of Misclassifications," OR 9, 1961, 875-885	$1M/1P_h/1M$-$E(W)$

Serial Number	Article	Classification
25	White, H. and Christie, L. S.,"Queuing with Preemptive Priorities or with Breakdown," OR 6, 1958, 79-95	$1M/1P_i/1M\text{-}P(L_q)$, $E(L_q^2)$, $E(T^2)$
26	Welch, P. D.,"Some Contribution to the Theory of Priority Queues,"Ph.D. Thesis, Columbia Columbia University, April 1963.	$1M/1P_i/1G\text{-}G(L_q)$, $P(\tilde{W})$.

III REVIEW OF ANALYTICAL DEVELOPMENT OF PRIORITY QUEUING

A. INTRODUCTION

A problem of considerable interest to service system designers is the study of the system performance characteristics under various combinations of external environments and internal organizations. In many cases the projected service system is physically large and involves extensive financial outlay. Only if the system designers understand the projected system performance before they undertake the physical implementation of the system can they avoid both the extravagance of over-design and the danger of under-design. They may gain such an understanding through

(1) The application of the experience gained from operating similar service systems, or through

(2) The study of a model of the service system.

Of the several types of models (scaled-down physical model, numerical or simulation model, and mathematical or analytical model), we are here mainly interested in the mathematical model. In particular, we are interested in the mathematical modeling of priority-queue service systems.

The analytical modeling of a service system is the characterization of the service system in terms of mathematical equations. These equations relate the various states of the service system. A state of a service system is the condition of the system at some specific instant of time. The condition of a service system can be described in various degrees of detail; i.e., the system can be described as in the empty state, where there is no customer in the queue or in the server(s), or as in the busy state, where all the servers are engaged in service. The busy state can be further qualified by describing the number of customers

in the queue. If the customers are distinguishable into classes, one may further qualify the "busy state with n customers in queue" to include the description of the make-up of the n customers in the queue by class, etc. These mathematical equations may then be solved for the desired system performance measures, see Morse.[5]

In many service systems especially in a system where customers arrive from many independent sources, the arrival distribution can be characterized by an analytical probability distribution function. Often the service requirements for these customers can also be characterized by the probability distribution function. Such a service system is called the stochastic service system. The theory of probability has been used to characterize this type of service system, and indeed this branch of probability theory is now called the theory of queues or queuing theory. The theory of queues has been successfully used to analyze a variety of service systems, such as the telephone system, the taxi system, the hospital system, the global communication system, and the job-shop system. A most comprehensive bibliography of this theory may be found in Saaty[3].

Although this analytical modeling method has been extensively used in the analysis of service systems, it is still limited to the analysis of relatively simple service systems, as was observed by Jackson.[6] Nevertheless, a knowledge of the solution to even a simplified version of the service system can often provide the system designer with a means of bounding the solution to a more complex one, as suggested by Camp.[7] Furthermore, it may enable him to obtain quick estimates of the desired

parameters of the service system under study.

It is, then, the purpose of this Section to review the present status of analytical-modeling methods of treating priority queuing, and in particular, the priority queue situations that can occur in information-handling systems, and to present them in such a way that the practical limitations of analytical methods can be established for the system designers. The historical development of priority queue is briefly outlined in Part B. The results of the analytical modeling method are systematized for ease of reference in Part C.

The analytical model of a service system shall be called a queuing model. The symbolic representation of the queuing models and their operational measures used in this report are those described in Section I.

B. DEVELOPMENT OF PRIORITY-QUEUE THEORY

In this Section it is assumed that the reader is familiar with the simplified symbolic representation of the queuing model that was introduced in Section II-D. However, as an aid to the reader this scheme of representation is summarized here. The format of the queuing model is given by the representation

$$X_1X_2/Y_1Y_2/Z_1Z_2$$

where X_1 is the pattern of arrival

X_2 is the arrival distribution

Y_1 is the number of queues

Y_2 is the queue discipline

Z_1 is the number of servers

Z_2 is the distribution of service time.

The queuing models discussed in this Section have the following common characteristics, the pattern of arrival is always simply, $X_1 = 1$, the

arrival distribution is all cases assumed to be Poissonian, $X_2 = M$ and there is only a single queue, $Y_1 = 1$. Hence, the queuing models have the common format of $1M/1Y_2/Z_1Z_2$. The specific symbols used for Y_2, the queue discipline, Z_1, the number of servers, and Z_2, the distribution of service time, are listed below:

$Y_2 = P_h$, head-of-the-line priority,

$= P_i$, interrupt priority,

$= P_p$, preemptive priority,

$= P_h(t)$, dynamic head-of-the-line priority,

$= 0$, ordered queue discipline,

$Z_1 = 1$, single server,

$= S$, multiple servers,

$Z_2 = M$, negative exponential distribution,

$= G$, general distribution,

$= E_k$, Erlangian distribution.

Analytical treatment of priority queues began with Cobham,[8] who derived the mean waiting time of the $1M/1P_h/1G$ and the $1M/1P_h/SM$ queuing models. Since then others have investigated other aspects of the head-of-the-line priority model. In 1957, Kesten and Runnenburg[9] obtained the Laplace-Stieltjes transform (L.S.T.) of the waiting time of the $1M/1P_h/1G$ queuing model and the waiting-time distribution* of the $1M/1P_h/1M$ queuing model with only two classes of customer. Cox[10] derived the waiting-time distribution of any class of customer in a $1M/1P_h/SM$ model. However, his distribution turns out to be incorrect because of an incorrect assumption, the assumption that the waiting-time distribution is exponential. Dressin

*The waiting time distribution formula (5.40) given by Kesten and Runnenburg contains an error. The corrected formula is presented in Section III-C of this report.

and Reich[11] independently derived the transform of the waiting time distribution of the $1M/1P_h/1M$ model. In 1957, Morse[12] obtained the generating function of the probability distribution on the number of customers in the queue, $G(L_q)$, for the $1M/1P_h/1M$ model. Miller[13] used the method of embedded Markov chain to obtain the $G(L_q)$ for the $1M/1P_h/1G$ model. In addition, Miller derived the generating function for the probabilities on the number of items served during a busy period.

The interrupt (pre-emptive resume) priority-queue model, $1M/1P_i/1-$, was first investigated by White and Christie[14] in 1958. They obtained the $G(L_q)$, $E(L_q)$ and $E(L_q^2)$ of the $1M/1P_i/1M$ model with identical service rate. In addition they obtained the $\tilde{P}(W)$, $E(W)$ and $E(W^2)$. Miller[13] derived the $G(L_q)$, $E(L_q)$ and $E(L_q^2)$ of the $1M/1P_i/1M$ model with different service rate for each of the two priority classes. Furthermore, he obtained the $\tilde{P}(W)$, $E(W)$, $E(W^2)$, $\tilde{P}(B)$, $E(B)$, $E(B^2)$ and the generating function for the probabilities on the number of items served during a busy period of the $1M/1P_i/1G$ model. In 1960, Heathcote[15] obtained the $G(L_q)$ and $E(L_q)$ of the $1M/1P_i/1E_k$ model where the $k = 1$ for the highest-priority customer. Avi-Itzhak and Naor[16] (1961) studied the $1M/1P_i/1M$ model where the population of the highest priority customer is limited and they obtained the $E(L_q)$. $E(W)$ and $E(W^2)$ of the model. Gaver[17] (1962) derived the $E(L_q)$ and $E(L_q^2)$ of the $1M/1P_i/1G$ model. All of the above authors considered the case where the source population contains two priority classes of customers. In 1963, Welch[18] obtained the $G(L_q)$ and $\tilde{P}(W)$ for the $1M/1P_i/1G$ model where the population is assumed to contain r priority classes of customers.

The time-dependent aspect of the $1M/1P_i/1_-$ model with two priority classes was investigated by Heathcote[19] in 1959 and by Jaiswal[20] in 1961.

Heathcote obtained the $G(L_q(t))$, $\tilde{P}(B)$, $E(B)$ and $E(B^2)$ for the exponential service time case. Jaiswal obtained the $P(\tilde{L}_q(t))$ and $\tilde{P}(B)$ for the general service time case.

The pre-emption (pre-emptive repeat) priority queue model $1M/1P_p/1_-$, was first discussed by White and Christie [14] in 1958, but it was Gaver [17] who, in 1962, gave a more rigorous treatment of the model. More specifically Gaver derived the $\tilde{P}(B)$, $\tilde{E}(B)$, $P(L_q)$, $G(L_q)$ and $E(L_q^2)$ for the $1M/1P_p/1G$ model.

The dynamic priority-queue model, $1M/1P_h(t)/1M$, has been investigated by Jackson. [21-23] He presented a procedure for computing the equilibrium waiting time distribution. Since the dynamic priority model, $1M/1P_h(t)/1M$, can be bounded by the $1M/1O/1M$ model on one hand and by the $1M/1P_h/1M$ on the other, the computing procedure is not presented in this paper.

This then constitutes most of the known analytical investigations of the priority queue models. The present state of the art of analytical modeling of priority queues is limited to models with the following characteristics:

(1) Input

(a) Population source--unlimited for lower priority customers

(b) Population characteristics--heterogeneous (In most cases, the population is assumed to contain two classes of priority customers.)

(c) Arrival pattern--single

(d) Arrival distribution--Poissonian for the highest-priority customers

(e) Policy--delayed

(2) Queue

(a) Number--one

(b) Size--unlimited

(c) Discipline--static and dynamic priority

head-of-the-line

interrupt (preemption resume)

preemption (preemption repeat).

(3) Service

(a) Number--one*

(b) Distribution--M, E_k and G

(c) Pattern--single

Even with these relatively simple priority-queue models, the probability distribution of the operational measures studies is most often left in implicit form, either as the generating function or the transform of the operational measure. The explicit form is either too difficult to obtain or is too cumbersome for practical application. The alternative is to obtain the first few moments (in particular the first two moments) of the probability distributions from the implicit form and use these moments to gain insight to the operational characteristics of the queuing model.

---

*There is one exception: Cobham[8] obtained the mean waiting time of the M/1P_h/SM model with identical service rate for all customers.

C. SOME USEFUL RESULTS OF PRIORITY QUEUING ANALYSIS

The explicit results of priority-queuing analysis are summarized and listed in this section. They are listed in the order of priority discipline: head-of-the-line, interrupt, and preemption. Within each priority discipline they are further ordered by the number of priority classes considered and by the service time distribution of the customers. The bracketed number and the set of numbers outside of the bracket at the right of the formula represent the serial number of the reference and the equation number in that reference from which the formula is obtained or derived. The references are listed at the end of Section I-D.

The symbols used here are defined as follows:

λ_i = mean arrival rate of the i-customer (ith-priority customer)

μ_i = mean service rate of the i-customers = $1/E(S_i)$.

S_i = service time of the i-customers.

L_i = number of i-customers in the system.

W_i = waiting time of the i-customers.

$E(X)$ = the expected or the mean of the measure X.

$E(X^2)$ = the second moment of the measure X.

$P(X \geq a)$ = the probability distribution function for the measure X.

ρ_i = offered load of i-customers = $\lambda_i E(S_i)$.

1. Head-of-the-Line Priority--Single Server

a. Identical Exponential Service Time--Two Priority Categories

REFERENCE

Eq. 1. $$E(L_1) = \frac{\rho_1(1 + \rho_2)}{1 - \rho_1}$$ [19a] 9.10

Eq. 2. $$E(W_1) = \frac{\rho_1 + \rho_2}{\mu(1 - \rho_1)}$$ [19] 3.4

Eq. 3. $$E(W_1^2) = \frac{2(\rho_1 + \rho_2)}{\mu^2(1 - \rho_1)^2}$$ [19] 3.5

Eq. 4. $$P(W_1 \geq t) = 1 - \rho e^{-(1-\rho_1)\mu t}$$ [18] 5.39

Eq. 5. $$E(L_2) = \rho_2 + \frac{\rho_2(\rho_1 + \rho_2)}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}$$ [19a] 9.23

Eq. 6. $$E(W_2) = \frac{\rho_1 + \rho_2}{\mu(1 - \rho_1)(1 - \rho_1 - \rho_2)}$$ [19] 3.12

Eq. 7. $$E(W_2^2) = \frac{2}{\mu^2(1 - \rho_1)^2(1 - \rho_1 - \rho_2)}$$

$$\left[\rho_1 + \rho_2 + \frac{\mu(\rho_1 + \rho_2)}{1 - \rho_1 - \rho_2} + \frac{\rho_1(\rho_1 + \rho_2)}{1 - \rho_1}\right]$$ [19] 3.13

Eq. 8. $$P(W_2 \geq t) = 1 - \frac{(\rho_1 + \rho_2)^2 - \rho_1}{\rho_2} e^{-\alpha\mu t}$$

$$- \frac{1 - \rho_1 - \rho_2}{2\pi} \int_{\alpha_1}^{\alpha_2} f(r)dr$$ [Appendix I] I-7

where

$$\alpha = \frac{\rho_2(1 - \rho_1 - \rho_2)}{\rho_1 + \rho_2}$$

$$\alpha_1 = (1 - \sqrt{\rho_1})^2$$

$$\alpha_2 = (1 + \sqrt{\rho_1})^2$$

$$f(r) = \frac{e^{-r\mu t}\sqrt{(\alpha_2 - r)(r - \alpha_1)}}{r(r - \alpha)}$$

b. Exponential Service Time--Two Priority Categories

Eq. 9. $$E(L_1) = \frac{\rho_1}{1 - \rho_1}\left[1 + \rho_2 \frac{\mu}{\mu_2}\right]$$ [19a] 9.23

Eq. 10. $$E(W_1) = \frac{1}{1 - \rho_1}\left[\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2}\right]$$ [19] 3.4

Eq. 11. $$E(W_1^2) = \frac{2}{(1 - \rho_1)^2}\left[\frac{\rho_1}{\mu_1^2} + \frac{\rho_2}{\mu_2^2} + \frac{\rho_1\rho_2}{\mu_2}\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)\right]$$ [19] 3.5

Eq. 12. $$P(W_1 \le t) = \frac{\lambda_1(1 - \rho_1 - \rho_2) + \lambda_2}{\mu_1(1 - \rho_1)} e^{-\mu_1(1-\rho_1)t}$$ [18] 5.33

Eq. 13. $$E(L_2) = \rho_2 + \frac{\lambda_2}{\mu_1 - \lambda_1}\left[\frac{\rho_1 + \rho_2 \frac{\mu_1}{\mu_2}}{1 - \rho_1 - \rho_2}\right]$$ [19a] 9.23

Eq. 14. $$E(W_2) = \frac{1}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}\left[\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2}\right]$$ [19] 3.12

Eq. 15. $$E(W_2^2) = \frac{2}{(1 - \rho_1)^2(1 - \rho_1 - \rho_2)}\left[\frac{\rho_1}{\mu_1^2} + \frac{\rho_2}{\mu_2^2} + \frac{\left(\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2}\right)}{(1 - \rho_1 - \rho_2)} + \frac{\frac{\rho_1}{\mu_1}\left(\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2}\right)}{(1 - \rho_1)}\right]$$ [19] 3.13

c. General Service Time--Two Priority Categories

Eq. 16. $$E(L_1) = \frac{\lambda_1}{\lambda_1 + \lambda_2}\left[\rho_1 + \rho_2 + \frac{\lambda_1[\lambda_1 E(S_1^2) + \lambda_2 E(S_2^2)]}{2(1 - \rho_1)}\right]$$ [19] 2.19

Eq. 17. $$E(W_1) = \frac{\lambda_1 E(S_1^2) + \lambda_2 E(S_2^2)}{2(1 - \rho_1)}$$ [19] 3.4

Eq. 18. $$E(w_1^2) = \frac{\lambda_1 E(s_1^3) + \lambda_2 E(s_2^3)}{3(1-\rho_1)} + \frac{\lambda_1 E(s_1^2)[\lambda_1 E(s_1^2) + \lambda_2 E(s_2^2)]}{2(1-\rho_1)^2}$$ [19] 3.5

Eq. 19. $$E(L_2) = \frac{\lambda_2}{\lambda_1 + \lambda_2}\left\{\rho_1 + \rho_2 + \frac{\lambda_1 E(s_1^2) + \lambda_2 E(s_2^2)}{2\mu_1} + \left[\frac{\lambda_2(\mu_1 + \mu_2) + \rho_1(1-\rho_1-\rho_2)}{(1-\rho_1)(1-\rho_1-\rho_2)}\right]\right\}$$ [19] 2.20

Eq. 20. $$E(w_2) = \frac{\lambda_1 E(s_1^2) + \lambda_2 E(s_2^2)}{2(1-\rho_1)(1-\rho_1-\rho_2)}$$ [19] 3.12

Eq. 21. $$E(w_2^2) = \frac{\lambda_1 E(s_1^3) + \lambda_2 E(s_2^3)}{3(1-\rho_1)^2(1-\rho_1-\rho_2)} + \frac{[\lambda_1 E(s_1^2) + \lambda_2 E(s_2^2)]^2}{2(1-\rho_1)^2(1-\rho_1-\rho_2)^2} + \frac{\lambda_1 E(s_1^2)[\lambda_1 E(s_1^2) + \lambda_2 E(s_2^2)]}{2(1-\rho_1)^3(1-\rho_1-\rho_2)}$$ [19] 3.13

d. <u>General Service Time--r-Priority Categories</u>

Eq. 22. $$E(W_k) = \frac{\sum_{i=1}^{r} \lambda_i E(s_i^2)}{2\left[1 - \sum_{i=1}^{k-1} \lambda_i E(s_i)\right]\left[1 - \sum_{i=1}^{k} \lambda_i E(s_i)\right]}$$ [18] 5.30

Eq. 23. $$E(w_k^2) = \frac{\sum_{i=1}^{r} \lambda_i E(s_i^3)}{3\left[1 - \sum_{i=1}^{k-1} \lambda_i E(s_i)\right]^2\left[1 - \sum_{i=1}^{k} \lambda_i E(s_i)\right]} + \frac{\sum_{i=1}^{r} \lambda_i E(s_i^2) \sum_{j=1}^{k} \lambda_j E(s_j^2)}{2\left[1 - \sum_{i=1}^{k-1} \lambda_i E(s_i)\right]^2\left[1 - \sum_{i=1}^{k} \lambda_i E(s_i)\right]^2} + \frac{\sum_{i=1}^{r} \lambda_i E(s_i^2) \sum_{j=1}^{k-1} \lambda_j E(s_j^2)}{2\left[1 - \sum_{i=1}^{k-1} \lambda_i E(s_i)\right]^3\left[1 - \sum_{i=1}^{k} \lambda_i E(s_i)\right]}$$ [18] 5.31

2. Interrupt Priority--Single Server

a. Exponential Service Time--Two Priority Categories

Eq. 24. $$E(L_1) = \frac{\rho_1}{1 - \rho_1}$$ [19] 2.3

Eq. 25. $$E(L_1^2) = \frac{\rho_1}{1 - \rho_1}\left[1 + \frac{2\rho_1}{1 - \rho_1}\right]$$ [19] 2.3

Eq. 26. $$E(W_1) = \frac{\rho_1}{\mu_1(1 - \rho_1)}$$ [19] 3.4

Eq. 27. $$E(W_1^2) = \frac{2\rho_1}{\mu_1^2(1 - \rho_1)^2}$$ [19] 3.5

Eq. 28. $$P(W_1 \le t) = 1 - \rho_1 e^{-\mu_1(1 - \rho_1)t}$$ [18] 5.39

Eq. 29. $$E(L_2) = \frac{\rho_2}{1 - \rho_1 - \rho_2}\left[1 + \frac{\mu_2}{\mu_1}\left(\frac{\rho_1}{1 - \rho_1}\right)\right]$$ [25] 13

Eq. 30. $$E(L_2^2) = \frac{2\rho_1(\lambda_2/\mu_1)^2}{(1 - \rho_1)^3(1 - \rho_1 - \rho_2)} + \frac{\rho_2^2}{(1 - \rho_1 - \rho_2)^2}\left[1 + \frac{\mu_2}{\mu_1}\left(\frac{\rho_1}{1 - \rho_1}\right)\right]^2 + \frac{\rho_2(1 - \rho_1)^3 + \rho_1^2(\lambda_2/\mu_1)^2 + \rho_1(1 - \rho_1)(1 - \rho_1 - \rho_2)(\lambda_2/\mu_1)}{(1 - \rho_1)^2(1 - \rho_1 - \rho_2)^2}$$ [19] 2.4

Eq. 31. $$E(W_2) = \frac{\rho_1 + \rho_2\frac{\mu_1}{\mu_2}}{\mu_1(1 - \rho_1)(1 - \rho_1 - \rho_2)}$$ [19] 3.12

Eq. 32. $$E(W_2^2) = \frac{2[\rho_1 + \rho_2(\mu_1/\mu_2)^2]}{\mu_1^2(1-\rho_1)(1-\rho_1-\rho_2)} + \frac{2[\rho_1 + \rho_2(\mu_1/\mu_2)]}{\mu_1^2(1-\rho_1)(1-\rho_1-\rho_2)}$$

$$\left[\rho_1 + \frac{\rho_1 + \rho_2(\mu_1/\mu_2)}{\mu_1(1-\rho_1)(1-\rho_1-\rho_2)}\right] \quad [19] \quad 3.13$$

b. General Service Time--Two Priority Categories

Eq. 33. $$E(L_1) = \rho_1 + \frac{\rho_1 + \lambda_1^2[E(S_1^2) - E(S_1)^2]}{2(1-\rho_1)} \quad [15] \quad 3.4$$

Eq. 34. $$E(W_1) = \frac{\lambda_1 E(S_1^2)}{2(1-\rho_1)} \quad [19] \quad 3.4$$

Eq. 35. $$E(W_1^2) = \frac{\lambda_1 E(S_1^3)}{3(1-\rho_1)} + \frac{[\lambda_1 E(S_1^2)]^2}{2(1-\rho_1)^2} \quad [19] \quad 3.5$$

Eq. 36. $$E(L_2) = \frac{1}{1-\rho_1}\left\{\rho_2 + \frac{\lambda_2[\lambda_2 E(S_2^2) + \lambda_1 E(S_1^2)]}{2(1-\rho_1-\rho_2)}\right\} \quad [15] \quad 35$$

Eq. 37. $$E(W_2) = \frac{\lambda_1 E(S_1^2) + \lambda_2 E(S_2^2)}{2(1-\rho_1)(1-\rho_1-\rho_2)} \quad [19] \quad 3.12$$

Eq. 38. $$E(W_2^2) = \frac{\lambda_1 E(S_1^3) + \lambda_2 E(S_2^3)}{3(1-\rho_1)(1-\rho_1-\rho_2)} + \frac{[\lambda_1 E(S_1^2) + \lambda_2 E(S_2^2)]^2}{2(1-\rho_1)^2(1-\rho_1-\rho_2)^2}$$

$$+ \frac{\lambda_1 E(S_1^2)[\lambda_1 E(S_1^2) + \lambda_2 E(S_2^2)]}{2(1-\rho_1)(1-\rho_1-\rho_2)} \quad [19] \quad 3.13$$

3. Preemptive Priority--Single Server

a. Exponential Service Time--Two Priority Categories

Eq. 39. $E(L_1)$ = same as Eq. 24

Eq. 40. $E(L_1^2)$ = same as Eq. 25

Eq. 41. $E(W_1)$ = same as Eq. 26

Eq. 42. $E(W_1^2)$ = same as Eq. 27

Eq. 43. $P(W_1 \leq t)$ = same as Eq. 28

Eq. 44. $$E(L_2) = \frac{\alpha_1 \rho_2}{1 - \rho_1} + \frac{(\alpha \rho_2^2)\beta}{2(1 - \rho_1)(1 - \rho_1 - \alpha\rho_2)} + \frac{\rho_1 \rho_2 (\mu_2/\mu_1)}{(1 - \rho_1)^2}$$ [8] 9.10

where $$\alpha = \frac{\mu_2}{\mu_2 - \lambda_1}$$

and $$\beta = \frac{2(\mu_2 - \lambda_1)^2}{\lambda_1^2}\left[\frac{2\lambda_1^2}{(\mu_2 - \lambda_1)(\mu_2 - 2\lambda_1)} + \frac{1}{\mu_1^2(1 - \rho_1)} - \frac{\lambda_1 \mu_2 (1 - \rho_1)}{(\mu_2 - \lambda_1)^2} - \rho_1 + 1\right]$$

b. General Service Time--Two Priority Categories

Eq. 45. $E(L_1)$ = same as Eq. 33

Eq. 46. $E(W_1)$ = same as Eq. 34

Eq. 47. $E(W_1^2)$ = same as Eq. 35

Eq. 48. $$E(L_2) = \frac{\alpha \rho_2}{1 - \rho_1} + \frac{(\alpha \rho_2)^2 \beta}{2(1 - \rho_1 - \alpha\rho_2)(1 - \rho_1)} + \frac{\rho_2 \lambda_1 E(S_1^2)}{2E(S_2)(1 - \rho_1)^2}$$

[8] 9.10

where $\alpha = \dfrac{E\left(e^{-\lambda_1 S_2}\right) - 1}{\lambda_1 E(S_2)}$

and $\beta = \left\{2E\left[\left(e^{\lambda_1 S_2} - 1\right)^2\right] + \left[\dfrac{\lambda_1^2 E(S_1^2)}{1 - \rho_1} + 2(1 - \rho_1)\right]\left[E\left(e^{\lambda_1 S_2}\right) - 1\right]\right.$

$\left. - 2\lambda_1(1 - \rho_1)E\left(S_2 e^{\lambda_1 S_2}\right)\right\}\left[E\left(e^{\lambda_1 S_2}\right) - 1\right]^{-2}$

4. Head-of-the-Line Priority--Multi-server (C-server)

For simplicity of writing let

$$\rho_i = \frac{\lambda_i}{c\mu} ,$$

$$\rho = \frac{1}{c\mu}\sum_{i=1}^{r}\lambda_i ,$$

and

$$P_D = 1\Big/\left\{1 + \frac{c!(1 - \rho)}{(c\rho)^c}\sum_{j=0}^{c-1}\frac{(c\rho)^j}{j!}\right\}$$

a. Identical Exponential Service Time--Two Priority Categories

Eq. 49. $E(W_1) = \dfrac{P_D}{c\mu(1 - \rho_1)}$ [4] 6

Eq. 50. $P(W_1 \leq t) = 1 - P_D e^{-(1-\rho_1)\mu t}$ [5] 26

Eq. 51. $E(W_2) = \dfrac{P_D}{c\mu(1 - \rho_1)(1 - \rho_1 - \rho_2)}$ [4] 6

Eq. 52. $$P(W_2 \le t) = 1 - \frac{P_D(\rho^2 - \rho_1)}{\rho\rho_2} e^{-\alpha t} - \frac{P_D(1-\rho)}{2\pi\rho} \int_{\alpha_1}^{\alpha_2} f(r)dr$$ [Appendix II] II-7

where

$$\alpha = \frac{\rho_2(1-\rho)}{\rho} c\mu$$

$$\alpha_1 = (1 - \sqrt{\rho_1})^2 c\mu$$

$$\alpha_2 = (1 + \sqrt{\rho_1})^2 c\mu$$

$$f(r) = \frac{e^{-r\mu t}\sqrt{(\alpha_2 - r)(r - \alpha_1)}}{r(r - \alpha)}$$

b. **Identical Exponential Service Time--r Priority Categories**

Eq. 53. $$E(W_k) = \frac{P_D}{c\mu\left(1 - \sum_{i=1}^{k-1} \rho_i\right)\left(1 - \sum_{i=1}^{k} \rho_i\right)}$$ [4] 6

for $\sum_{i=1}^{k} \rho_i < 1$.

5. **Interrupt Priority--Multi-server (C-server)**

a. **Identical Exponential Service Time--r Priority Categories**

Eq. 54. $$E(W_k) = \frac{P_{Dk}}{c\mu\left(1 - \sum_{i=1}^{k-1} \rho_i\right)\left(1 - \sum_{i=1}^{k} \rho_i\right)}$$ [Appendix II] II-9

for $\sum_{i=1}^{k} \rho_i < 1$,

where $P_{Dk} = 1/1 + \frac{c!(1 - \rho_k)}{(c\rho_k)^c} \sum_{j=0}^{c-1} \frac{(n\rho_k)^j}{j!}$

and $\rho = \frac{1}{c\mu} \sum_{i=1}^{k} \lambda_i$.

IV SIMULATION OF PRIORITY QUEUING SYSTEMS

A. INTRODUCTION

As indicated in Sect. III, the presently available analytical modelling techniques are mostly limited to the analysis of relatively simple queuing systems. In the case of the priority queue, this technique is further limited to the analysis of single-server systems. Even with these simple queuing systems, the analytical techniques are not capable of yielding some of the desirable operational measures, in particular the distributions of the measures. An alternative technique for analyzing a queuing system is the numerical (simulation) modelling technique. This technique involves the reduction of a queuing system into a series of numerical statements, which characterize the flow-sequence of events in the queuing system. Figure IV-1 shows a simplified logical flow-sequence diagram of a typical single-stage queuing system. Arrival events are generated by the "arrival event generator." The most recent event" is selected for action; in this particular case, the event is either an arrival or a departure.

In general, the simulator keeps track of the history of each arrival event, noting its arrival time, the time when it enters the server, and the time when it leaves the server. The simulator also samples the queue status at some arbitrary time. From the historical data of an arrival event, the waiting time and delay time can be computed. These data are accumulated in the form of histograms, which can then be analyzed to yield the desired operation measures. This flow process can be repeated until sufficient data has been obtained. The data can then be analyzed to yield the desired results.

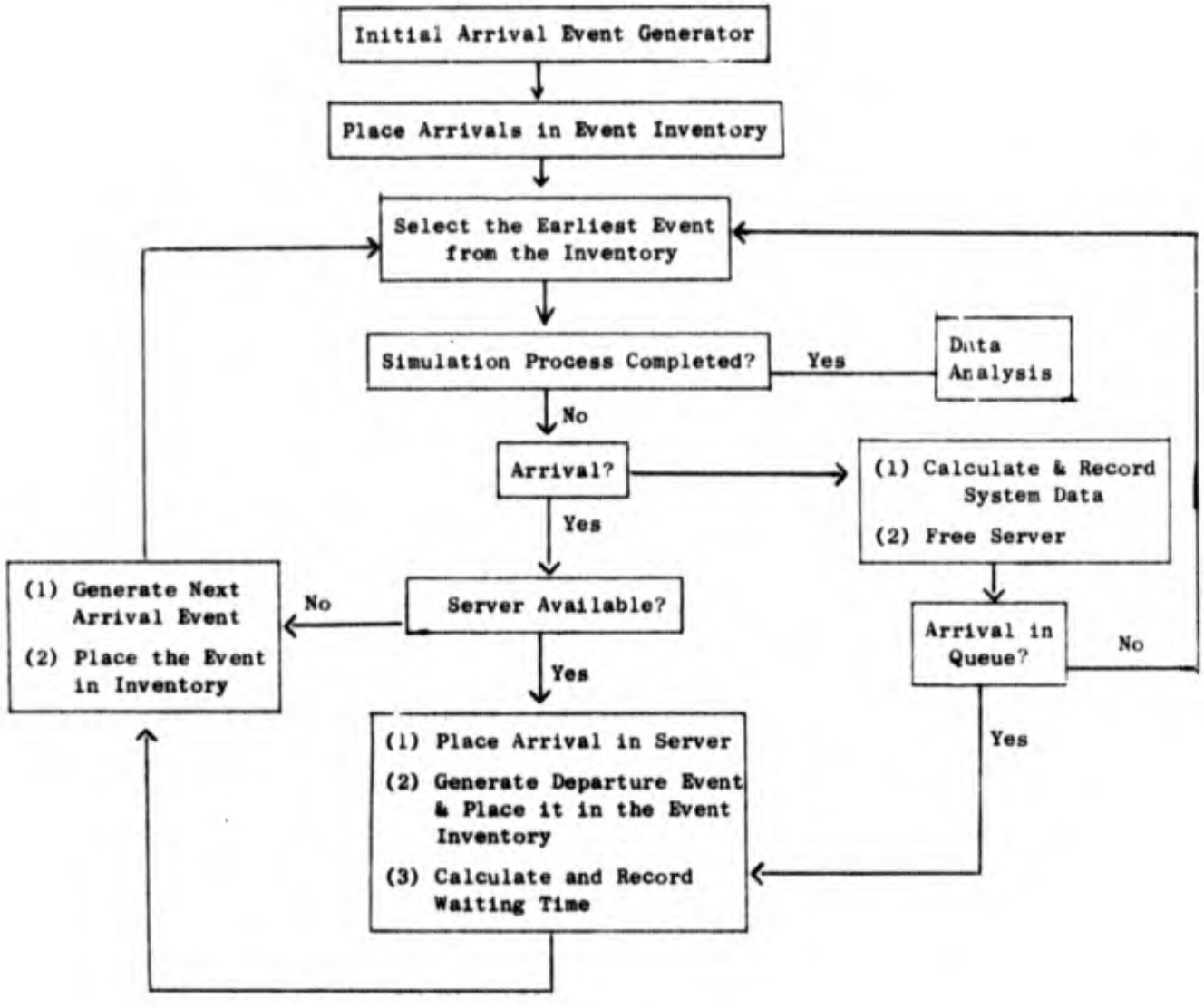


Fig. IV-1

LOGICAL FLOW-SEQUENCE DIAGRAM OF A TYPICAL SINGLE-STAGE QUEUING SYSTEM

The major advantage of the numerical modelling technique over the analytical technique is its inherent flexibility. It is possible to obtain empirical operational measures of a queuing system with any combinations of input, queue, and service characteristics. The user of the technique can obtain these measures under carefully controlled input, queue, and service characteristics. The principal disadvantage is that the validation of the numerical model can be rather time consuming, especially for a larger model.

The numerical model of a queuing system, like an experimental set-up of a physical system, can be used to yield any amount of empirical data, once it has been validated. The efficiency of a numerical model is partly a function of how well it is constructed and partly a function of how the experiment is carried out and how the results are analyzed. For a queuing system, such modelling is generally statistical in nature and the experiment usually involves a transient state. The questions usually asked by the experimenter are: (1) Which portion of the data is transient? (2) How often should one sample the data? (3) When should the simulation process be stopped? (4) How many simulation runs are required to yield sufficient representative data?

The numerical modelling technique has been applied to the modelling of a single-stage priority queuing system. The general characteristics of the priority queuing system model are described in Part B, and the process of checking out the model is detailed in Part C. Some typical results of this model are shown in Part D, together with remarks concerning the possible future development of the model. The model is currently programmed in the SUBALGOL (Stanford University BALGOL) language for the IBM 7090 computer.

B. PRIORITY-QUEUE SIMULATION MODEL

The queuing system that has been modelled in this section can be called the single-stage priority queuing system: schematically, it is as shown in Fig. IV-2.

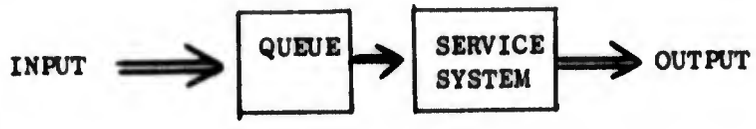


Fig. IV-2

SCHEMATIC DIAGRAM OF THE SINGLE-STAGE PRIORITY QUEUING SYSTEMS

Using the convention adopted in Sect. II, the queuing model is described in the order of its input, queue, and service characteristics.

Input--The input population source is assumed to be unlimited and heterogeneous. It contains r classes of customers. Class 1 is the highest-priority class and class r is the lowest-priority class. The interarrival time interval of each class of customer is distributed according to some given distribution with a given means. The programmed distribution of this model is the Poissonian distribution but it can easily be changed to other distributions. The customers are assumed to arrive singly, and to wait in the queue until their required service has been rendered. The amount of service, in terms of units of message, that is required by a specific class of customer is distributed according to some given distribution with a given mean. The programmed distribution is negative-exponential.

Queue--The queue is the single, unlimited queue. In an actual computer program, the queue size is limited by the size of computer

memory available. However, if a sufficiently large memory is available, the queue size becomes essentially infinite. The queue discipline is a mixed-priority discipline. Each of the r classes is assigned a priority queue discipline that can be head-of-the-line, P_h, or interrupt, P_i, or preemption, P_p priority discipline. Within each class of customer, ordered queue discipline is used. For example, if there are four classes of customers, the Class 1 customer is assigned the P_p discipline, the Class 2 customers is assigned the P_i discipline, and the Class 3 and Class 4 customers are assigned the P_h discipline. The arrival of a Class 1 customer will preempt from the lower-class customers who are in service, and the service on the preempted customer will resume from the beginning. The arrival of a Class 2 customer can only interrupt the service of Class 3 and Class 4 customers. Service on the interrupted customer will resume from the point of interruption. An arriving Class 3 customer will have precedence only over a waiting Class 4 customer. An arriving Class 4 (the lowest class) will merely join the end of the queue. All of the above arrival situations presume that a queue exists at the time of arrival.

Service--The service organization is assumed to contain n parallel homogeneous servers. The service rate of each server, in term of message units per unit time, is assumed to be constant. Thus, the service rate of each server, in term of the number of customers of a specific class per unit time, is distributed according to the "amount of service" distribution for that class of customer. The servers are numbered from 1 to n. If there is more than one server available, the server with the lowest number is the next server to render service to the customer in the queue. It is assumed that no time is lost during the selection of servers.

In summary the parameters of the simulated priority queuing system are as follows:

Input: Population size--unlimited

Number of classes of customer--up to six

Arrival pattern--single arrivals

Arrival distribution--Poissonian

Policy--delayed

Queue: Number of queues--one

Queue size--unlimited

Queue discipline--static priority

Head-of-the-line

Interrupt

Preemption

Service: Organization--multiple-parallel homogeneous, with up to 100 servers

Service-time distribution--negative exponential

Pattern--single service.

The logical flow chart of the simulation model is as shown in Fig. IV-3. The simulation started with the initial conditioning of the simulator, which involves reading in input-parameter cards and setting the simulator to the initial condition. The input parameters are:

The starting value of the random number, which is a 10-digit integer that is not divisible by 2 or 5.

Number of priority classes

Mean arrival rate for each priority class, λ_i in customers per unit time

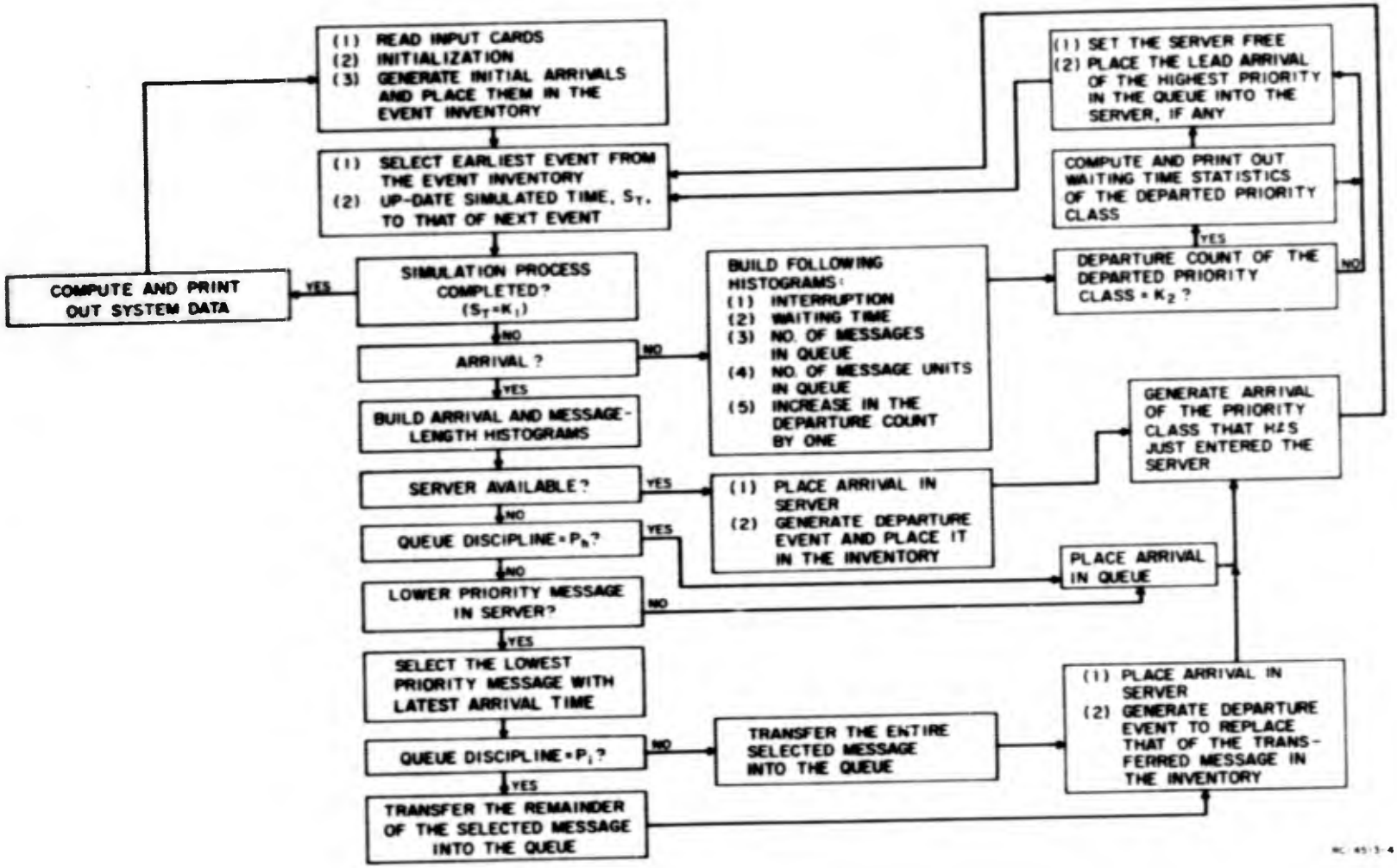


Fig. IV-3

LOGICAL FLOW CHART OF PRIORITY QUEUE SIMULATION MODEL

Mean service units for each priority class, l_i, in units of message per customer

Queue discipline for each priority class

Number of servers

Service rate of the server, in units of message per unit time

Number of departures when waiting-time statistics are to be computed and printed out, K_2,

Number of total departures before the simulation process ceases, K_1

The range and step size of the histograms for the following system data:

- Interarrival time, T_a
- Message length, ℓ
- Number of messages in queue, N_q
- Number of message units in queue, l_q
- Waiting time, t_w
- Number of interruptions, X_i
- Delay time, t_d
- Interdeparture time, T_d.

Once the initial conditioning is completed, the simulator proceeds to simulate the flow-sequence of events and to collect the necessary system data. There are two major events, arrival and departure. When an arrival occurs, the simulator will determine the availability of server and the queue discipline of the arriving customer. When the determination has been made, the simulator will place the arrival in the server, if the arrival is entitled to immediate service. In this

case, a departure event is generated for this arrival and placed in the inventory. The departure time for the arrival is the time of its entry into the server, plus the time required for service. If the arriving customer is not entitled to a server immediately, it is placed into the queue. In either case, a new arrival event of the same priority class as the last arrival event is generated and placed into the inventory. Next, the current earliest event is selected from the inventory for processing. If this event is a departure event, the various histograms are constructed. Furthermore, if the departure event is an integral modulo of n departures of that priority class, where n is any positive number, the mean waiting time and the cumulative probability distribution of the waiting time are printed out. In any case, the server is set free and the next event is selected from the inventory for processing. This flow-sequence of events is repeated until the total number of departures equals the preassigned value, K_1. The cumulative probability distributions for the system data T_a, ℓ, N_q, ℓ_q, T_w, X_i, t_d, and T_d are computed from the respective histograms and printed for each priority class. The simulation process will proceed to process the next priority queuing system if the input parameter cards are present. Otherwise the process ends.

C. PROCEDURE FOR VALIDATING THE MODEL

The simulation model just described is to be used as an experimental tool for obtaining empirical data concerning certain priority queue systems. It is therefore extremely desirable to validate the model. The validating procedure involves two steps. The first step is the "micro-logic" check and the second step is the "macro-process" check.

The "micro-logic" check is the validation of the logic of the mathematical modelling process, and is carried out by letting the simulator run through its program. At every event point, whether an arrival or a departure, the condition of the model is printed out for inspection. In this way the logical flow-sequence can be checked in detail. In particular, the histogram-building process is checked. The initial queue discipline employed is that of P_h; the P_i and P_p discipline are then checked independently. Finally, the mixed-priority queue discipline situation is checked. It is obvious that this checking procedure is rather time consuming. To minimize the expense of complete validation, the checking was carried out until most of the major logical branching and data accumulation processes are determined to be substantially valid. When the validity of the logical flow-sequence has been determined, the next step of validation is the "macro-process" check. The simulator is used to simulate a queue priority system whose mean waiting time for each priority class is known analytically. The simulated mean waiting time is sampled at every n departures and compared with the analytical mean waiting time. When the simulated mean waiting time approaches the analytical mean waiting time as the simulation progresses, it is a positive indication that the entire simulating process (the "macro-process") is functioning properly. An added check is to plot the simulated waiting time distribution against the analytical waiting time distribution, although this can only be done for the $1M/1P_h/1M$ queuing system, since the analytical waiting time distribution of other priority-discipline queuing systems is not available.

D. SIMULATION OF THE $1M/1P_1/SM$ SYSTEM

The simulation model was used to simulate the $1M/1P_1/SM$ system, where S=2, 4, 10. The input population was assumed to contain two priority categories--the first and second priority messages. The messages have identical arrival and length characteristics. Their inter-arrival distribution was Poissonian, with a mean arrival rate of unity, $\lambda_1=\lambda_2=$ one message per unit time. Their message length was negative exponentially distributed, with a mean length of unity, $l_1=l_2=$one message unit. The servers were identical in every respect and they can serve any waiting messages. The service rate of each server for each of three cases, S=2, 4, and 10, was assumed to be 1.25, 0.625 and 0.25 message units per unit time. This yielded a system loading factor, ρ, of 0.8 for each of the three cases.

Two independent simulation runs were made for each case. A different initial random number, which was used to generate the arrival and message length distributions, was employed for each of the two runs. At every 200 departures of a priority message, the mean waiting time, the waiting time distribution, and the length of queue distribution of that priority message were sampled. The simulation terminates when six thousand messages of either priority have been serviced. The samples within each run are not strictly statistically independent; nevertheless, they provide a guide to the variability of the data obtained from the simulation run.

Simulation started from an empty state, no message in the system--thus few initial departures would contain transient statistics. It is for this reason that the first two samples, the statistics for the first four hundred departures, were not included in calculating the simulated mean and variance of the waiting time. Table IV-1 lists the simulated mean and variance of the waiting time for each of the two runs and the combination of the two runs. The theoretical mean waiting times were

computed in accordance with the Eq. 53, given in Section III-C (p. 36). The simulated mean waiting time of individual Runs except for the S=10 case, are less than or equal to 15% from the theoretical mean (see Table IV-2). The simulated mean waiting time of the combination of the two runs are less than or equal to 10% from the theoretical mean, except in the case of S=10.

TABLE IV-1

SIMULATED MEAN WAITING TIME AND ITS VARIANCE OF THE $1M/1P_1/SM$ SYSTEM

Number of Servers-S	Priority	Theoretical Mean Waiting Time	Simulated Waiting Time					
			Run #1		Run #2		Run #1 and #2	
			Mean	Variance	Mean	Variance	Mean	Variance
2	1	0.152	0.154	0.056	0.151	0.085	0.152	0.071
	2	2.370	2.186	0.585	2.325	1.215	2.240	0.955
4	1	0.061	0.067	0.040	0.069	0.054	0.068	0.050
	2	1.988	1.730	0.593	1.939	1.132	1.835	0.909
10	1	0.006	0.004	0.006	0.007	0.012	0.005	0.009
	2	1.364	3.242	1.7745	1.284	0.893	2.263	1.696

TABLE IV-2

PERCENTAGE DEVIATION OF THE SIMULATED MEAN WAITING TIME FROM THE THEORETICAL MEAN WAITING TIME

Number of Servers-S	Priority	Run #1	Run #2	Run #1 and Run #2
2	1	1	1	0
	2	8	2	6
4	1	9	12	10
	2	15	3	8
10	1	40	19	10
	2	58	6	40

The simulated mean inter-arrival time and mean message length for the case S=10 are as follows:

Priority	Inter-Arrival Time	Message Length
1	0.991	1.007
2	0.995	0.999

It is seen that the simulated means deviated from the theoretical means by less than 1%. The simulated mean system loading factor is 0.797, which is only 0.4% lower than the theoretical loading factor of 0.8. Thus it is concluded that the excessive deviation of the simulated waiting time statistics from the theoretical value is not due to simulated inter-arrival time and message length distribution. The statistics on the probability of interruption for the two simulated runs for the S=10 case indicated a much higher frequency of interruption for the first run; the probability of interruption of second priority message is 0.59, and 0.39 for the first and second run. It seems to indicate that the arrival pattern of the first and second priority messages in the first run is such that they arrive in bunches. This resulted in higher probability of interruption and higher waiting time for the second priority. The relatively high deviation of the mean waiting time of the first simulation run of the S=10 case is most probably due to transient effect of the initial random number.

The simulated waiting time and queue length distributions for the case of S=2 are plotted in Figs. IV-4 and IV-5. The data are based upon 1400 departures of each priority category. The theoretical waiting time distribution of the first priority is plotted as the solid line in Fig. IV-4. The theoretical distributions of the second priority waiting

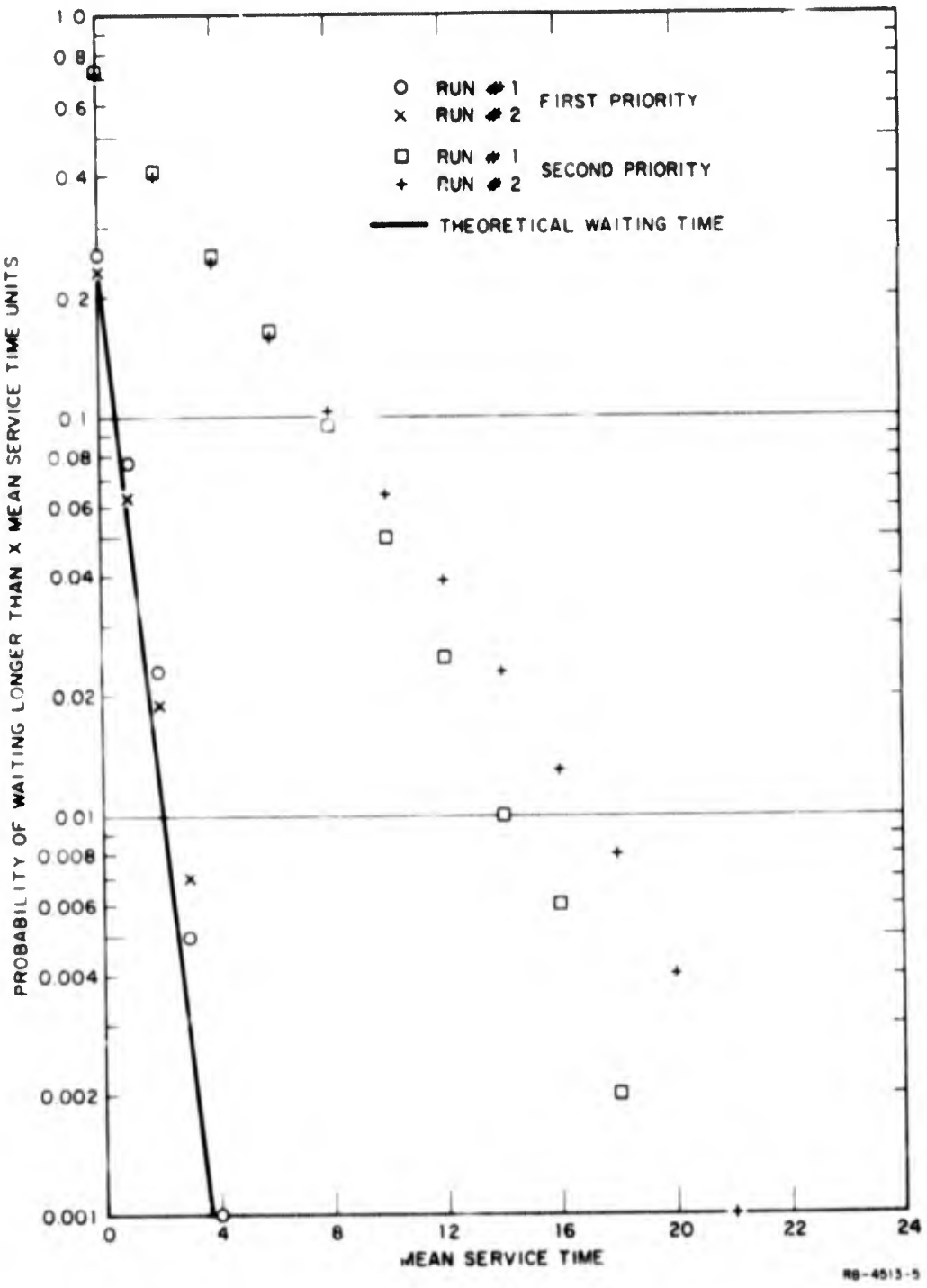


Fig. IV-4

SIMULATED WAITING TIMES DISTRIBUTION FOR 1M/1P$_i$/2M MODEL
(ρ = 0.8)

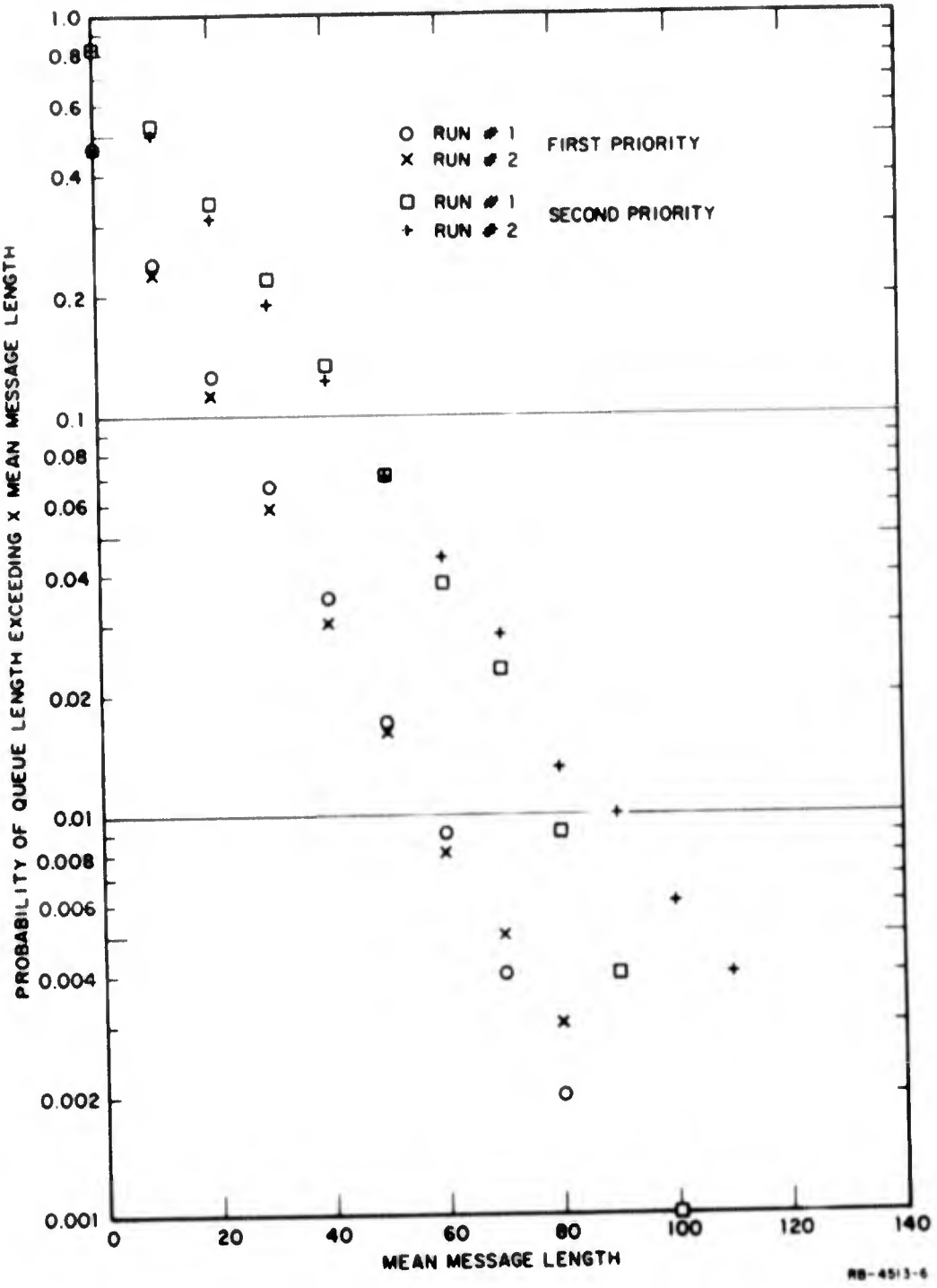


Fig. IV-5

SIMULATED QUEUE LENGTH DISTRIBUTION OF $1M/1P_1/2M$ SYSTEM

time and of the queue length are not available. A visual inspection of these figures seems to indicate that these distributions may be approximated by exponential distributions. As expected, the variance of data for higher values of the waiting time and queue length is larger than that at lower values. The theoretical probability and the simulated probability of having to wait, P_D, are as shown -

Priority	Probability of having to wait, P_D,		
	Theoretical	Run #1	Run #2
1	0.229	0.257	0.233
2	0.711	0.726	0.708

The simulated probabilities deviate from the theoretical probability by less than 15%. To obtain the empirical waiting time or queue length distributions from the simulation statistics, the standard method of testing hypotheses may be employed. It is known that computer programs for such testing are available, but due to the lack of time and funds this phase of study was not carried out.

Based upon the data obtained from very limited testing of the simulation model, it is concluded that the simulation model may be used to establish empirical formulas of queuing systems with sufficient accuracy (within ± 10%) for the purpose of preliminary system design. However, further testing of the simulation model is required to better calibrate the confidence level of the results of this model. In particular, further investigation is needed on the feasibility of integrating the program of the testing of hypotheses with this model.

V GUIDES AND PROCEDURES FOR APPLICATION OF PRIORITY QUEUING MODELS

A. INTRODUCTION

The current status of priority queuing analysis has been reviewed in Sec. III and the formulas developed for a few specific priority queuing systems were listed in Sec. III-C. The numerical modeling technique (simulation) was described in Sec. IV. The general procedure for applying the priority-queue modeling technique to the structuring of information systems is outlined in Sec. V-B. This procedure involves two major steps: (1) characterization of the information system, and (2) selection of the appropriate modeling technique. A specific example is selected to illustrate this procedure, the details of which are contained in Sec. V-C.

B. GENERAL PROCEDURE

1. Characterization of the Information System

Selection of the appropriate modeling techniques for a given information system begins with the characterization of the system in terms of its input, queue, and service processes, and the operational measures to be applied.

The parameters for characterizing these three processes for a single-stage queuing system were given in Sec. II, together with the most commonly encountered operational measures. The majority of the parameters can be characterized precisely, since they are in most cases deterministic. For example, the input population size may contain x sources and each of the x sources may generate y types of customers; the queue may consist of a single queue buffer of size n and the queue discipline may be the head-of-the-line priority queue; the service organization may contain c

identically parallel servers. The input (arrival) distribution and the service-time distribution are generally probabilistic in nature. It is thus necessary to accumulate and analyze the arrival and service time data. The analyzed data are then subjected to statistical tests* to determine the type of distribution (Poissonian, Erlangian, or exponential). If the data are not available, as in the case of newly designed systems, it might be possible to postulate a distribution based upon the predicted nature of input and service characteristics. For example, a Poissonian arrival distribution may be assumed if the arriving customers came from relatively large numbers of independent sources. The service time distribution may be assumed to be negative exponential if there is a great variety of customers, each with his own service requirement.

A large information system usually contains a number of interconnected information handling centers (IC). Customers (information messages) may be processed through one or more of these processing centers. At each center the customers may experience delay. If the input queue buffer of an information center is small with respect to its work load,** ρ, it is very probable that messages destined for this center

---

*For more detailed discussion on the subject of statistical testing one may refer to pp. 44-48 of Saaty (Ref. 3), or any standard statistical text book.

**Work load is defined as the ratio of the mean arrival rate to the mean service rate.

from other centers will queue up at the other centers. In this case the queues of these centers are no longer independent of each other, and the characterization of such a system requires the linking of single-stage queues. The single-stage queue model characterization scheme, in Sec. II, may be modified to characterize such a system. Figure V-1 shows a possible symbolic characterization of a hypothetical information system where there are three information-handling centers, A, B, and C. Centers A and B receive customers from direct sources, and x percent and y percent, respectively, of the customers processed at A and B enter C. In addition C receives customers from a direct source.

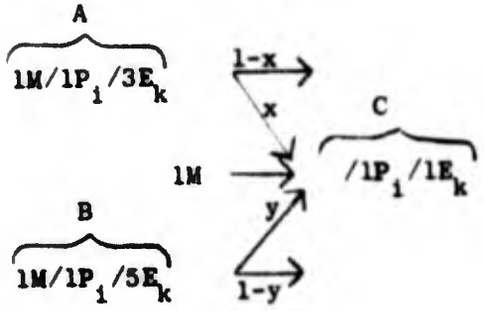


Fig. V-1

SYMBOLIC CHARACTERIZATION OF AN INFORMATION SYSTEM

It is seen that, for a system with a large number of interconnected centers, the symbolic characterization can be rather complex. However, in many practical information systems the queue buffer is relatively large, or can be assumed to be large for the purpose of preliminary analysis, hence it is possible to characterize the system as a collection of independent single-stage queues. The three centers in the hypothetical system can then be represented by three single-stage queues: $1M/1P_i/3E_k$, $1M/1P_i/5E_k$, and $1E_k/P_i/1E_k$.

2. Selection of Modeling Technique

When the system has been characterized, the next step is to select the appropriate modeling techniques. The sequence of modeling-technique selection is as follows:

a. Locate Available Analytical Models

If the available analytical models have been catalogued, and if the catalogue is indexed according to the classification scheme used to characterize the information system, it may then be possible to locate the standard analytical model, if it exists. No general catalogue is known to exist; there are a few specialized catalogues, such as the priority queuing model catalogue in Sec. III-C, and the non-priority queuing models catalogue of Shelton.[24] There are also a number of excellent references that contain information on available analytical models, such as, Morse,[12] Saaty,[3] and Syski.[24] In addition, there are a number of published tables, such as the "Delay Table for Finite and Infinite Source System," by Descloux,[26] and "Finite Queuing Tables" by Peck and Hazelwood.[27] Saaty,[3] in particular, has a most comprehensive collection of analytical queuing models that were developed prior to 1960. In the absence of a general catalogue, the search process involves the review of published queuing literature. Often the search ends with the location of analytical models that are only an approximation to the information system in question. In that case, one of two courses of action is open: (i) develop an exact analytical model by modifying the approximate model, or (ii) use the approximate models to create the bounds to the solution. Although it is most desirable that an exact analytical model be developed, an exact analytical model for a

relatively complex queuing system often does not yield easily to numerical evaluation. Indeed, in preliminary design of an information system, a reasonably close bound to the solution is often sufficient. Camp[7] has introduced some basic concepts of bounding practical queuing problems by analytic methods. In brief, the bounding is achieved by idealizing the information system in one or more ways, each of which yields an analytically manageable mathematical model of the system differing from the original in a known direction. Thus the original system is partially or completely "surrounded" by bounds obtained by analysis alone.

b. Develop Numerical Models

Should the analytical modeling technique fail to yield the desired solution, numerical modeling techniques may be employed. The numerical modeling technique is by far the most versatile and flexible. It can be made to model the original system in as much detail as desired, and to yield the operation measures as precisely as required. There is no theoretical limit as to how complex a system can be modeled and to what precision data can be produced. However, there is a practical limit in the cost and time required to construct and to operate such a model. Thus, in practice, the numerical model is invariably an approximation of the original system. Those system parameters that are expected to cause only minor perturbation to the required operational measures are usually deleted from the model. The evaluation of a numerical model requires so much computation that a digital computer is usually used. Several computer simulation languages have been developed, of which some of the better known are the General Purpose System Simulation (GPSS) of IBM, Simscript of RAND Corporation, and Control and Simulation Language (CSL) of Esso and IBM. The use of these simulator languages can reduce the total

effort of constructing the numerical model. Since the simulator languages generally make use of some intermediate compiler language, such as FORTRAN, the resulting simulation program is usually not as efficient as one programmed directly in the compiler language. For a small information system, such as the single-stage queuing system, it is believed that not much can be gained by programming it in the simulation languages.

C. SPECIFIC EXAMPLE

The hypothetical information system characterized in Fig. V-1 will be used as an illustrative example. The characteristics of the system will be amplified as follows:

(1) The source population at each of the three centers contain three priority classes of messages, $k = 1, 2, 3$.

(2) The message length (in words per message) is distributed in accordance with truncated exponential, with mean length $\bar{l}_k$ for the k-priority class. That is to say, there is a minimum and maximum message length.

(3) Each server at centers A and B has a service rate of x_a and x_b words per unit time. Thus, the service-time distribution for the k-priority message at A and B is a truncated exponential with mean service times $\bar{t}_{ka} = \bar{l}_k/x_a$ and $\bar{t}_{kb} = \bar{l}_k/x_b$, respectively.

(4) The service-time distribution for the k-priority message at C is truncated exponential with a mean service time $\bar{t}_{kc}$.

It is desired to investigate the expected waiting time of each priority class, and the size of queue buffer required to keep the probability of overflowing the buffer under a given traffic load less than some value, M.

The system as it is characterized cannot be modeled by existing analytical techniques. However, it be idealized by making the following assumptions:

(1) The queue buffer size at A, B, and C is assumed to be infinitely large. This assumption will produce an optimistic estimate of the waiting-time characteristics at A and B. The degree of underestimation is dependent upon the value of M; the smaller the value of M, the less will waiting time be underestimated.

(2) The message-length distribution is first assumed to be exponential, and then to be constant with a mean length $\bar{l}_k$ for the k-priority message. The first assumption will yield an upper bound, the second a lower bound of the expected waiting time.

With these idealizations, the system can now be represented by three independent systems for Centers A, B, and C: 1M/1Pi/3- for Center A, 1M/1Pi/5- for Center B, and 1M/1Pi/1- for Center C. From the analytical modeling point of view, the first two systems can be treated as a generalized multi-server model, 1M/1Pi/S-.

Before proceeding to locate the available analytical model(s), it is necessary to clarify the concept of queue length in many information systems. In a system where each message occupies a number of buffer locations (word positions), the queue length is defined as the number of

word positions occupied by the messages in the queue instead of the number of messages in the queue (the customary definition). This has been done for the queue buffers at A, B, and C. It has been shown by Wright[28] and by Davis[29] that the probability that the queue length exceeds 1 word, $P(L > 1)$, under the assumption of identical message-length distributions for each priority class, is

$$P(L > 1) = P_D e^{-(1-\rho)1/\bar{l}} \qquad (55)$$

where 1 is mean message length and P_D is the probability of any delay (as given on p. 44). It is to be noted that Eq. (55) is in effect the waiting-time distribution, of a $1M/1P_h/SM$ system, Eq. (50), where message-length distributions for each priority class are identical and where the service time of a message is proportional to its message length, $t = 1/x$ and $\bar{t} = \frac{1}{\mu} = \bar{l}/x$.

In the hypothetical system, a single-queue buffer is provided for all priority messages at each center; hence, it is not necessary to determine the queue length of each individual priority class. It follows that Eq. (54) may be used to determine the required queue-buffer size, l, for given ρ, $\bar{l}$, and $P(L > 1) = M$. An upper bound on the buffer size can be established by letting $\bar{l}$ be the longest mean message length among all priority classes.

A search of the catalogue of available analytical priority queuing models, Sec. III-C, reveals that the closest model is the $1M/1P_i/1G$ of Sec. III-C-2-b. The next question is how to use the equations in Sec. III to investigate the waiting time characteristics of each priority class at

each center. Center C will be analyzed first, since it is the simplest. Equations (34) and (37) allow one to obtain the expected waiting time of the first two priority class messages directly. The service of the third priority messages may be interrupted by the arrival in queue of higher-priority messages, regardless of the order of arrival of the higher-priority messages. Consequently, from the third-priority-message point of view, the first two priority messages are indistinguishable. Thus, Eq. (37) can be modified to yield the mean waiting time of the third-priority messages. This consists of combining the first- and second-priority messages into a single-priority category, and treating the third-priority messages as the second category. Thus Eq. (37) becomes

$$E(W_3) = \frac{[\lambda_1 E(S_1^2) + \lambda_2 E(S_2^2)] + \lambda_3 E(S_3^2)}{2(1 - \rho_1 - \rho_2)(1 - \rho_1 - \rho_2 - \rho_3)} \quad . \qquad (56)$$

By substituting $E(S_i^2) = \frac{1}{2\mu_i^2}$ and $E(S_i^2) = \frac{1}{\mu_i^2}$* for i = 1, 2, 3, into Eqs. (34), (37), and (56), one obtains the upper and lower bound of the expected waiting time.

---

*$E(S_i^2) = \frac{1}{2\mu_i^2}$ and $E(S_i^2) = \frac{1}{\mu_i^2}$ are the second moment of the exponential and constant service-time distribution with mean service time $\frac{1}{\mu_i}$

For Centers A and B, the only available analytical model is the $1M/1P_i/SM$ of Sec. III-C-5-a. This model can be used to obtain the upper bound of the expected waiting time. It is to be noted that the mean waiting time of a k-priority message is independent of messages whose priority is lower than k. It is also to be noted that Eq. (54) assumes an identical service rate for A and B, $\mu_1 = \mu_2$. To obtain the upper upper bound, one should use the highest service rate among the k or higher priority messages.

In summary, the available analytical modeling techniques allow one to investigate the following operational measures of the hypothetical system:

(1) The upper bound of the queue length distribution, which in turn yields the queue buffer size required.

(2) The upper and lower bound of the expected waiting times at Center C.

(3) The upper bound of the expected waiting times at Centers A and B.

It does not permit one to obtain the above measures under actual service-time distribution, nor does it provide the over-all delay-time distribution for those messages that traverse the system, through Centers A and C or Centers B and C. To obtain the more exact operational measures and the over-all delay-time distribution, one may have to resort to the numerical modeling techniques.

VI CONCLUSIONS AND RECOMMENDATIONS

A. CONCLUSIONS

A relatively simple scheme for characterizing single-stage queue systems can be realized. A slight modification to this scheme enables it to represent multi-stage (network of) queue systems symbolically. However, as the network becomes complex, this scheme of representation becomes too unwieldy. Some other scheme of network representation should be devised.

The queue system characterization scheme has been applied to classify priority queuing literature formulas. This same scheme may be used to classify more general queuing literature and formulas.

A review of priority queuing literature indicated that the present state of the art of analytical analysis of priority queues is limited to the analysis of the single-queue, single-server queuing system, where the input population is usually limited to two priority categories. Furthermore, mixture of priority queuing disciplines is not allowed. Even in these relatively simple priority queue systems, the analytical technique is not able to produce explicit results beyond the first few moments of the desired operational measures, such as the waiting time and the number in the queue. The attempt to advance the analytical analysis of priority queues is likely to progress slowly particularly for the multi-server, multi-priority-category (interrupt or preemptive) queuing system.

Preliminary tests of the simulation model indicated that it can yield data that is sufficiently accurate for use in preliminary system

design. In order to increase the confidence level of the simulated results, it is necessary to conduct further calibration tests of this model. Once calibrated, the model may then be used as an experimental (mathematical) tool to augment analytical techniques in investigating priority queuing systems that cannot be handled by analysis alone.

B. RECOMMENDATIONS

1. Classification

An initial attempt has been made to classify the useful results of analytical priority queuing analysis. It is recommended that this classification effort be kept up-to-date.

2. Simulation Model

Several improvements may be incorporated into the present simulation model to make it easier to use or more efficient in data gathering. It is recommended that the following modifications or improvements be incorporated into the present model:

(1) Make the assignment of range and step size of the histogram dependent upon the priority category.

(2) Automatically compute the mean input rate and mean message length at each sampling period.

(3) Automatically compute the mean of all the operational measures at the end of the simulation run.

(4) Investigate the possibility of integrating the simulation program with the testing of hypothesis program.

In the course of testing the simulation model, it was noted that the reliability of the simulation results is dependent upon the sampling scheme. The selection of the sampling scheme is in turn dependent upon the queuing system configuration, queuing discipline, system load factor, and the transient effect of the service system and of the random number generator. It is recommended that a series of test runs be made to establish the proper sampling scheme and to calibrate the model. This should include the use of Eq. 51 in Section III-C for calibrating the waiting time distribution aspect of the model.

3. Application of Priority Queues Modeling Techniques

The analytical and numerical priority queues modeling techniques gathered and developed in this study project have been tested on arbitrarily postulated single-stage priority queuing systems. It is recommended that these techniques be applied to a few specific priority queuing systems that are found in existing information systems to test the practicality of the basic concept of numerically augmented priority queuing analysis techniques.

REFERENCES

1. Brockneyer, E., Halstrom, H. L., and Jensen, A., The Life Works of A. K. Erlang, Trans. Danish Acad. Tech. Sci., 2, (1948).

2. Kendall, D. G., "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of the Imbedded Markov Chain," Ann. Math. Stat. 24, 338-354 (1953).

3. Saaty, T. L., Elements of Queuing Theory (McGraw-Hill Book Co., Inc., New York, 1961)

4. Saaty, T. L., "On the Problems of Jockeying Collusion, Scheduling, Optimization and Graph Theoretical Queues," paper presented at the First Joint National Meeting of ORSA and TIMS, (Nov. 1961).

5. Morse, P. M., Queues, Inventories and Maintenance (John Wiley and Sons, Inc., New York, 1958).

6. Jackson, R. R. P., and Adelson, R. M., "A Critical Survey of Queuing Theory, Part I," O. R. Quarterly (U.K.), Vol. 13 #1, pp. 13-22 (1962).

7. Camp, Glen D., "Bounding the Solution of Practical Queuing Problems by Analytic Methods," Operation Research for Management Vol. II. (The Johns Hopkins University Press, pp. 307-328 (1956).

8. Cobham, A., "Priority Assignment in Waiting Line Problem," OR 2, pp. 70-76 (1954).

9. Kesten, H., and Runnenburg, J. T., "Priority in Waiting Line Problems," Koninkl. Ned. Akad. Weten Schap Proc Ser A60, 1957, Pt I, pp. 312-324, and Pt II, pp. 325-336.

10. Cox, R. E., "Traffic Flow in an Exponential Delay System with Priority Categories," Proc. IEE (London) 102, Pt B, pp. 815-818 (1955).

11. Dressin, S. A., and Reich, E., "Priority Assignment on a Waiting Line," Quarterly of Applied Math. XV, pp. 208-211 (1957).

12. Morse, P.M., Queues, Inventories, and Maintenance (John Wiley and Sons, New York City, 1958).

13. Miller, R. G., "Priority Queues," Annals of Math. Statistics 31, pp. 86-103 (1960).

14. White, H., and Christie, L.S., "Queuing with Preemptive Priorities or with Breakdown," OR 6, pp. 79-95 (1958).

15. Heathcote, C. R., "A Simple Queue with Several Preemptive Priority Classes," OR 8, pp. 630-638 (1960).

16. Avi-Itzhak, B., and Naor, P., "On a Problem of Preemptive Priority Queuing," OR 9, pp. 664-672 (1961).

17. Gaver, D. P., "A Waiting Line with Interrupted Service, Including Priorities," J. Roy. Statistical Soc.(B) 24, pp. 73-90 (1962).

18. Welch, P. D., "Some Contribution to the Theory of Priority Queues," Ph.D. Thesis, Columbia University (April 1963).

19. Heathcote, C. R., "The Time-Dependent Problem for a Queue with Preemptive Priorities," OR 7, pp. 670-680 (1959).

20. Jaiswal, N. Y., "Preemptive Resume Priority Queue," OR 9, pp. 732-770 (1961).

21. Jackson, J. R., "Some Problems in Queuing with Dynamic Priorities," UCLA Management Sci. Res. Report 62 (November 1959).

22. Jackson, J. R., "Simulation of Queues with Dynamic Priorities," UCLA Management Sci. Res. Report 71 (March 1961).

23. Jackson, J. R., "Waiting-Time Distribution for Queues with Dynamic Priorities," NRLQ 9, pp. 31-36 (1962).

24. Shelton, J. R., "Solution Methods for Waiting Line Problems," The Journal of Industrial Engineering, pp. 293-303, 1960.

25. Syski, R., Introduction to Congestion Theory in Telephone Systems (Oliver and Boyd, Edinburgh, 1960).

26. Desclouse, A., Delay Tables for Finite and Infinite-Source Systems (McGraw-Hill Book Co., Inc. N. Y., N. Y. 1962).

27. Peck, L. G. and Hazelwood, R. N., Finite Queuing Tables (John Wiley and Sons, Inc., N. Y., N. Y. 1958).

28. Wright, E. P. G., "Basic Consideration in Calculating Storage for an Electronic Telegraph Switching Center," Electrical Communication, pp. 163-171 (1958).

29. Davis, R. H., et al., "Techniques for the Design and Utilization of Communication Networks," Quarterly Progress Report 3, prepared for ITT Communication Systems, Inc. Paramus, New Jersey, SRI Project 3453, September 1961.

APPENDIX A

WAITING-TIME DISTRIBUTION OF A 1M/1P_h/1M QUEUING SYSTEM

The Laplace transform of the waiting-time distribution of the second-priority customers, $W_2(t)$, in a 1M/1P_h/1M queuing system has been obtained by Dressin and Reich.[11] In this Appendix, the inverse of this transform is evaluated. The final result is presented in a closed form. The corresponding result as given in Kesten and Runnenburg[9] [Eq. 5.40,] is incorrect. Their result is corrected and given in Eq. (A-8). An alternative form is also given, Eq. (A-9). The corrected results of Kesten and Runnenburg may be integrated further, resulting in our form, Eq. (A-7), which thus furnishes the required result in its simplest form.

The Laplace transform as given in Dressin and Reich is

$$\tilde{W}_2(s) = (1 - \rho)\,[1 + 2A\,Q_2^{-1}(s)] \qquad \text{(A-1)}$$

where

$$Q_2(s) = s + \mu + \lambda_1 - 2A + \sqrt{(s + \mu + \lambda_1)^2 - 4\mu\lambda_1}$$

and

$$A = \lambda_1 + \lambda_2 \quad .$$

If the inverse Laplace transform of F(s) is defined by:

$$F(t) = L^{-1}\{F(s)\} = \frac{1}{2\pi i} \int_{Br_1} e^{st}\,\{F(s)/s\}\,ds,$$

Br_1 being the first Bromwich contour, Fig. A-1, then from Eq. (A-1),

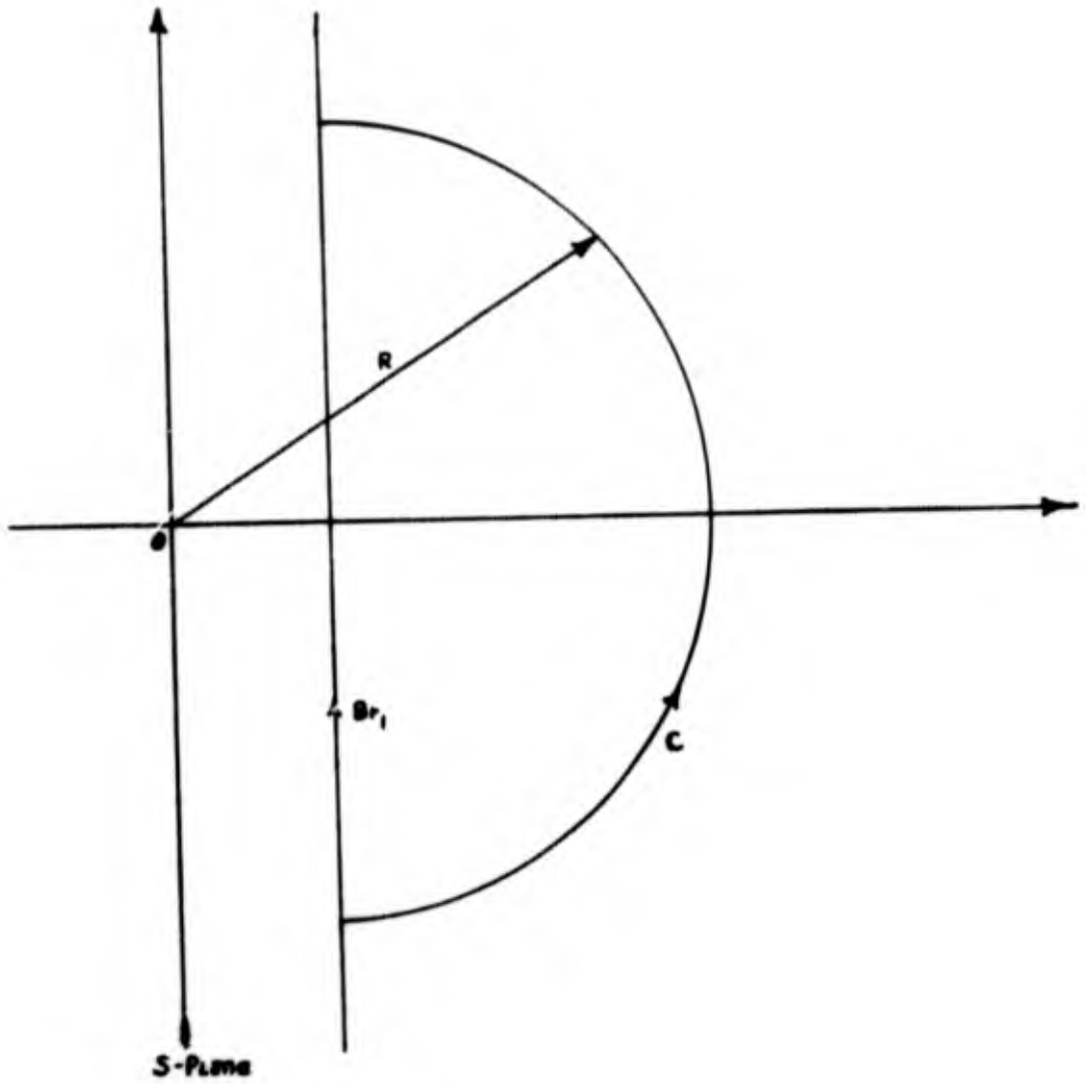


Fig. A-1

THE FIRST BROMWICH CONTOUR

$$W_2(t) = (1-\rho) + 2A(1-\rho)\, L^{-1}\{Q_2^{-1}(s)\} \qquad \text{(A-2)}$$

$$= (1-\rho) + 2A(1-\rho)\, \frac{1}{2\pi i} \int_{Br_1} \frac{e^{st} Q_2^{-1}(s)}{s}\, ds \ .$$

$Q_2^{-1}(s)$ has no singularities in the half plane to the right of Br_1. Therefore, we can deform the contour Br_1 to C and by Cauchy's integral theorem,

$$L^{-1}\{Q_2^{-1}(s)\} = \frac{1}{2\pi i} \int_C e^{st} \frac{1}{s} Q_2^{-1}(s)\, ds$$

$$= \frac{1}{2\pi i} \lim_{R\to\infty} \int_C e^{Rte^{i\theta}} Q_2^{-1}(Re^{i\theta})\, d\theta \ .$$

for large R, $|Q_2^{-1}(Re^{i\theta})| = O(\frac{1}{R})$, hence the integral for $t < o$ tends to zero for large R. Hence $L^{-1}\{Q_2^{-1}(s)\} = o$ for $t < o$.

To evaluate the inversion integral Eq. (A-2) for times greater than or equal to zero, the branch of the integral consistent with the condition Re. $s > o$ on Br_1 has to be continued analytically to its left and the path Br_1 be deformed to some suitable contour. The first condition is achieved by setting

$$s = R\, e^{i\theta}, \quad -\pi/2 < \theta < \pi/2$$

on Br_1 and then placing a proper cut along the negative real axis in the s - plane.

The integrand has branch points located at

$$s = \alpha_1, \ s = \alpha_2,$$

Where $\alpha_1 = (\sqrt{\mu} - \sqrt{\lambda_1})^2$

$$\alpha_2 = (\sqrt{\mu} + \sqrt{\lambda_1})^2$$

and poles at $s = o$ and $s = -\alpha$

where $\alpha = \dfrac{\lambda_2 (1-\rho)}{\rho}$. (A-3)

From Eq. (A-3) and the definition of 'α' it follows $\alpha < \alpha_1$. The integrand thus may be made analytical in the cut s - plane, the cut placed from $-\alpha$ to $-\infty$. A suitable contour is shown in Fig. A-2.

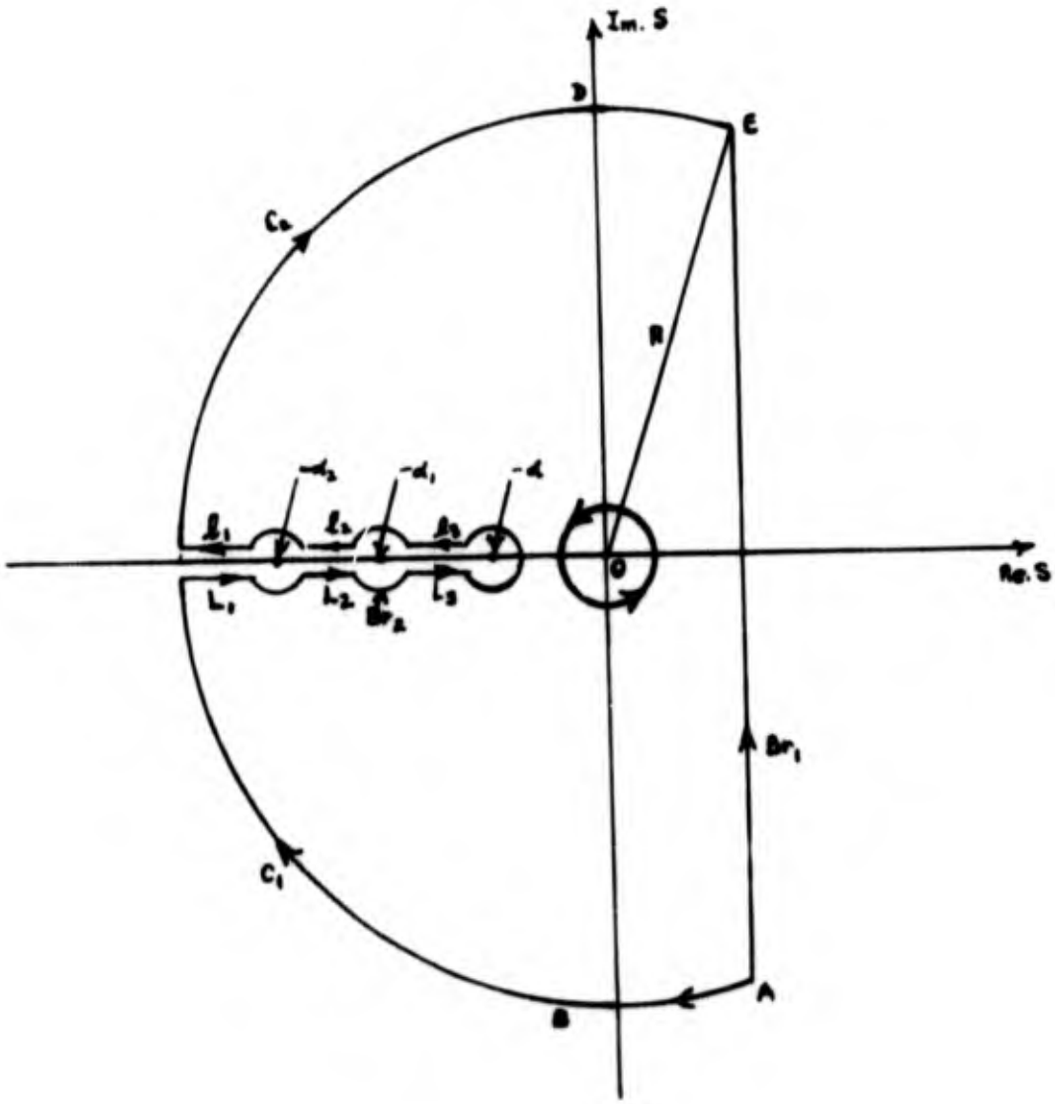


Fig. A-2
THE COMPOSITE CONTOUR TO WHICH CAUCHY'S INTEGRAL IS APPLIED IN THE DERIVATION OF THE WAITING TIME FUNCTION

where ABCDE is circular arc of radius R and Br_2 the second Bromwich contour. Again by Caughy's integral theorem,

$$L^{-1}\{Q_2^{-1}(s)\} = \frac{1}{2\pi i}\int_{C_1 + C_2 + Br_2} \frac{e^{st}}{s} Q_2^{-1}(s)\, ds \qquad (A\text{-}4)$$

now

$|Q_2^{-1}(s)| \leq 2/R$ for large R, and therefore by Jordan's Lemma,

$$\lim_{R\to\infty} \frac{1}{2\pi i} \int_{C_1 + C_2} \frac{e^{st}}{s} Q_2^{-1}(s)\, ds \to 0$$

and the inversion integral is given by

$$L^{-1}\{Q_2^{-1}(s)\} = \frac{1}{2\pi i}\int_{Br_2} \frac{e^{st}}{s} Q_2^{-1}(s)\, ds$$

$$= \frac{1}{2\pi i}\int_{Br_2} \frac{e^{st}\, ds}{s\{s + c_2 + \sqrt{(s+\alpha_1)(s+\alpha_2)}\}} \qquad (A\text{-}5)$$

where

$$= I_0 + I_1 + I_2 + I_3 + I_4 + I_5 + I_6$$

I_0 = contribution from the pole at s = 0

I_1 = contribution from the pole at $s = -\alpha$

I_2 = contribution from the lineal segments $L_3 + \ell_3$

I_3 = contribution from the branch point $s = -\alpha_1$

I_4 = contribution from the lineal segments $L_2 + \ell_2$

I_5 = contribution from the second branch point $s = -\alpha_2$

I_6 = contribution from the lineal segments $L_1 + \ell_1$,

and $c_2 = b + a_1 - 2A$.

The table of arguments for various elements of the integrand is as follows:

let $s = re^{i\theta}$

Path	θ	r	s	$s+\alpha$	$s+\alpha_1$	$s+\alpha_2$
Br_1	$-\frac{\pi}{2}$ to $\frac{\pi}{2}$					
L_1	$-\pi$	∞ to α_2	$-r$	$(r-\alpha)e^{-i\pi}$	$(r-\alpha_1)e^{-i\pi}$	$(r-\alpha_2)e^{-i\pi}$
L_2	$-\pi$	α_2 to α_1	$-r$	$(r-\alpha)e^{-i\pi}$	$(r-\alpha_1)e^{-i\pi}$	α_2-r
L_3	$-\pi$	α_1 to α	$-r$	$(r-\alpha)e^{-i\pi}$	α_1-r	α_2-r
ℓ_3	π	α to α_1	$-r$	$(r-\alpha)e^{i\pi}$	α_1-r	α_2-r
ℓ_2	π	α_1 to α_2	$-r$	$(r-\alpha)e^{i\pi}$	$(r-\alpha_1)e^{i\pi}$	α_2-r
ℓ_1	π	α_2 to ∞	$-r$	$(r-\alpha)e^{i\pi}$	$(r-\alpha_1)e^{i\pi}$	$(r-\alpha_2)e^{i\pi}$

By piecewise integration it can be shown that

$$L^{-1}\{Q_2^{-1}(s)\} = \frac{1}{2\mu(1-\rho)} - \frac{1}{2A(1-\rho)}\left[\frac{\rho_2-\rho_1}{\rho_2}\right]e^{-\alpha t}\frac{1}{4\pi A}\int_{\alpha_1}^{\alpha_2}\frac{\sqrt{(\alpha_2-r)(r-\alpha_1)}}{r(r-\alpha)}e^{-rt}dr \qquad \text{(A-6)}$$

Substituting Eq. (A-6) into Eq. (A-2), the waiting time function $W_2(t)$ is given as:

$$W_2(t) = 1 - \frac{\rho_2-\rho_1}{\rho_2}e^{-\alpha t} - \frac{(1-\rho)}{2\pi}\int_{\alpha_1}^{\alpha_2}\frac{e^{-rt}\sqrt{(\alpha_2-r)(r-\alpha_1)}}{r(r-\alpha)}dr \quad . \qquad \text{(A-7)}$$

The correct form of the Kesten-Runnenburg result, Eq. (5.40), is given by:

$$W_2(t) = 1 - \rho + \frac{A^2}{ba_2}\left[1 - e^{\frac{-a_2(1-\rho)}{\rho}t}\right]$$

$$-\sqrt{a_1 b}\,(1-\rho)\int_0^t \frac{-a_2(1-\rho)(t-\tau)}{\rho}\,d\tau \int_0^{t-\tau} \frac{I_1(2\sqrt{a_1 b s})}{s}\, e^{-\frac{a_1 + A\rho}{\rho}s}\,ds \quad \text{(A-8)}$$

An alternative form of (A-8) is given as:

$$W_2(t) = 1-\rho + \frac{A^2}{ba_2}\left[1 - \frac{-a_2(1-\rho)}{e^{\rho}}\,t\right] -$$

$$-\sqrt{\frac{a_1}{b}}\,\frac{A}{a_2}\int_0^t \frac{I_1(2a_1 b\tau)}{\tau}\, e^{-(a_1+b)\tau}\left\{1 - e^{\frac{-a_2(1-\rho)(t-\tau)}{\rho}}\right\}d\tau \quad . \quad \text{(A-9)}$$

APPENDIX B

WAITING-TIME DISTRIBUTION OF THE $1M/1P_h/SM$ SYSTEM

The queuing system considered here is one in which a number of priority classes present demands to a number of parallel, identical servers. The demands from each class arrive in a Poissonian process, and the servers have a negative exponentially distributed service time. If all servers are busy when a demand arrives, the demands are queued in a time-of-arrival-within-priority sequence. It is desired to find the probability distribution of waiting times for each priority class.

Let

λ_p = the arrival rate of demands in the p-th priority class

μ = the service rate for each of the parallel servers

c = the number of identical, parallel servers

$Wc_p(t)$ = the probability density of the waiting time T (= t) for a customer of the p-th priority class.

If less than c customers of <u>all</u> priorities are present in the system when a customer of priority p arrives, then $T = 0$. If exactly c such customers are present, then T is distributed as the busy period of a system of c servers serving only the arrivals of priority <u>higher</u> than p. This busy period in turn, is distributed as the busy period of a single

server serving only the higher-than-p traffic but with a servicing rate which is c times that of any of the c servers (i.e., with rate $c\mu$).[1]

If m demands of priority p or higher are on queue when a demand of priority p arrives in the system, (i.e., are ahead of the arriving demand) then the distribution of T is the m + 1-fold convolution of the distribution of the busy period of a c-server system servicing only the higher-than-p priority traffic.[2]

Let

$f_c(k:p)(t)$ = the conditional distribution density of T for c servers and priority higher-than-p traffic when an arriving demand finds all servers busy and k-1 demands on queue. This is obtained from the corresponding single-server distribution by substituting $c\mu$ for μ.

$P_c(k:p)$ = the probability that all c servers are busy and there are k-1 demands of priority p-or-higher on queue when a demand of priority p arrives (for $k \geq 1$)*

and $P_c(0:p)$ = the probability that there are less than c demands of any priority in the system.

Then

$$W_{cp}(t) = \sum_{k=0}^{\infty} P_c(k:p)\, f_c(k:p)(t) \quad .$$

---

*Note that since the priority p demands are arriving at random, this quantity, $P_c(k:p)$, is equal to the unconditional probability that the system is in the state given in the definition.

And the transform of this quantity is given by

$$\tilde{W}_{cp}(S) = \sum_{k=0}^{\infty} P_c(k:p)\ \tilde{f}_c(k:p)(S)$$

$$= \sum_{k=0}^{\infty} P_c(k:p)\ \tilde{f}_c^{k}(1:p)(S)$$

where[1]

$$f_c(1:p)(S) = \frac{S + \sigma_{p-1} + c\mu - \sqrt{(S + \sigma_{p-1} + c\mu)^2 - 4c\mu\sigma_{p-1}}}{2\,\sigma_{p-1}}$$

and

$$\sigma_p = \sum_{i=1}^{p} \lambda_i \quad .$$

For $k \geq 2$, the steady-state equations, obtained in the usual way, are

$$P_c(k:p) = c\mu dt\ P_c(k+1:p) + \sigma_p dt\ P_c(k-1:p) + (1-c\mu dt-\sigma_p dt)\ P_c(k:p)$$

or

$$(1+\rho_p)\ P_c(k:p) = P_c(k+1:p) + \rho_p\ P_c(k-1:p) \qquad (B-1)$$

where

$$\rho_p = \sigma_p / c\mu \quad .$$

It follows from these equations that

$$P_c(k:p) = K\rho_p^{k-1} \quad , \quad k \geq 2 \qquad (B-2)$$

for some constant K to be determined.

In order to obtain the state equation for $P_c(1:p)$ it is necessary to introduce P_{c-1} and P_c, the respective probabilities that there are exactly c-1 and c customers of any priority in the system. Then, in equilibrium,

$$P_c(1:p) = c\mu dt\ P_c(2:p) + \sigma dt\ P_{c-1} + (1-\sigma_p dt - c\mu dt)\ P_c(1:p)$$

$$+ c\mu dt\ (P_c(1:p) - P_c)^*$$

or

$$\rho_p P_c(1:p) = P_c(2:p) + \rho P_{c-1} - P_c \qquad (B-3)$$

where

$$\sigma = \sum_{i=1}^{r} \lambda_i \quad \text{and} \quad \rho = \sigma/c\mu \quad .$$

From the well-known non-priority delay theory for multi-servers

$$P_c = \rho P_{c-1} \quad .$$

So from Eq. (B-3)

$$\rho_p\ P_c(1:p) = P_c(2:p) \quad .$$

Whence from Eq. (B-2)

$$P_c(1:p) = K$$

and therefore

$$P_c(k:p) = K\rho_p^{k-1} \quad \text{for} \quad k \geq 1 \quad . \qquad (B-4)$$

In order to determine $P_c(0:p)$, let

P_{cj} = the non-priority probability that there are exactly j demands in the c-server system.

Then

$$P_c(0:p) = \sum_{j=0}^{c-1} P_{cj} \quad .$$

It remains to determine the value of K to use in Eq. (B-4). This can be found from the condition

$$\sum_{k=0}^{\infty} P_c(k:p) = 1 \quad .$$

That is,

$$1 = P_c(0:p) + K \sum_{k=1}^{\infty} \rho_p^{k-1} = P_c(0:p) + \frac{K}{(1-\rho_p)} \quad .$$

From this,

$$K = (1-\rho_p)\left(1 - \sum_{j=0}^{c-1} P_{cj}\right) \quad .$$

Letting

$$P_D = 1 - \sum_{j=0}^{c-1} P_{cj} = \text{the probability of delay in a non-priority, c-server system with arrival rate } \sigma,$$

$$= [1 + \{(1-\rho)c!/(c\rho)^c\} \sum_{j=0}^{c-1} (c\rho)^j/j!]^{-1} \quad ,$$

then

$$P_c(0:p) = 1 - P_{cd} \quad ,$$

and

$$K = P_D\ (1 - \rho_p) \quad .$$

Thus

$$\tilde{W_{cp}}(S) = 1 - P_D + P_D(1-\rho_p)\ \tilde{f_c}(1:p)/(1-\rho_p \tilde{f_c}(1:p))$$

$$= 1 - P_D + P_D\ (1-\rho_p)\ 2\ (c\mu)/Q_{cp}(S) \qquad (B-5)$$

where

$$Q_{cp}(S) = S + \sigma_{p-1} + c\mu - 2\sigma_p + \sqrt{(S + \sigma_{pd} + c\mu)^2 - 4\ c\mu\ \sigma_{p-1}} \quad .$$

For the case p = 2, Eq. (B-5) becomes

$$\tilde{W_{c2}}(S) = 1 - P_D + P_D\ (1-\rho)2(c\mu)/Q_{c2}(S) \quad . \qquad (B-6)$$

It is noted that Eq. (B-6) has the same form as Eq. (A-1) of Appendix A. Thus, by proper substitution, one gets

$$W_{c2}(t) = 1 - \frac{P_D\ (\rho^2-\rho_1)}{\rho\ \rho_2}\ e^{-\alpha t} - \frac{P_D\ (1-\rho)}{2\pi\rho} \int_{\alpha_1}^{\alpha_2} f(r)\ dr \qquad (B-7)$$

where

$$\alpha = \frac{\rho_2(1-\rho)}{\rho} c\mu$$

$$\alpha_1 = (1-\sqrt{\rho_1})^2 c\mu$$

$$\alpha_2 = (1+\sqrt{\rho_1})^2 c\mu$$

$$f(r) = \frac{e^{-r\mu t}\sqrt{(\alpha_2-r)(r-\alpha_1)}}{r(r-\alpha)} \quad .$$

An approximation for the multiserver, preemptive priority system may be obtained by replacing P_D by the probability, P_{Dp}, of delay in a c-server system subjected to traffic of priority p or higher (i.e., σ_p). P_{Dp} is given by

$$P_{Dp} = [1 + ((1-\rho_p)c!/a_p^c) \sum_{j=0}^{c-1} a_p^j/j!]^{-1} \qquad (B-8)$$

where

$$a_p = c\,\rho_p = \sigma_p/\mu \quad .$$

From the form of $\tilde{W}_{cp}(s)$ it can be seen that the mean waiting time, $E_{cp}(W)$, for the p-th priority class and mean service rate μ is given by

$$E_{cp}(W) = \frac{d\,\tilde{W}_{cp}(S)}{ds}\Big|_{s=0} = \frac{P_{Dp}}{c\mu(1-\rho_{p-1})(1-\rho_p)} \quad . \qquad (B-9)$$

It is to be noted that Eq. (B-9) is similar to the mean waiting time for the $M/1P_h/SM$ system, Eq. (53) except that P_{Dp} replaces P_D.

APPENDIX C

THE PRIORITY QUEUING SIMULATION MODEL

The general description of the priority queuing simulation model was presented in Sec. IV-B. In this appendix the detailed computer program, input data card format and typical simulation output are presented.

A. Detailed Computer Program

To aid the understanding of the simulation program some of the more important terms used in the program are defined as follows:

A(i,1) = arrival time of the i-th priority message

A(i,2) = length of the i-th priority message

ARRIVH = inter-arrival time histogram

Count = computed number of departures

DELAYH = delay time histogram

DISCIP(i) = queuing discipline of the i-th priority category;
1 = P_h, 2 = P_i, and 3 = P_p discipline

DLAST = last departure of any priority

DQNUMH = number of messages in queue histogram

DQLENH = length of queue (message units) histogram

INTERH = interruption histogram

KZ2 = number of departures at which time system data is sampled

LENGTH = length of arriving messages histogram

LAM(i) = mean arrival rate of the i-th priority category

MML(i) = mean message length of the i-th priority category

MST(1) = mean service time of the i-th priority category

NDEP = the duration of simulation run (number of departures)

NP = number of priority category

NRD = starting random number

NSERV = number of servers

Q(i,j,1) = arrival time of the j-th message of i-th priority category

Q(i,j,2) = message length of the j-th message of the i-th priority category

Q(i,j,3) = number of interruption of the j-th message of i-th priority category

Q(i,j,4) = waiting time of the j-th message of i-th priority category

Range = the range of the histograms

SPDEPH = inter-departure time histogram

SRATE = service rate of each server

STEP = the step size of the histogram

WAITH = waiting time histogram

SERVER(i,j) = the j-th property of the message in the i-th server

j = 1, the priority

j = 2, the number of interruption

j = 3, the arrival time

j = 4, the message length

j = 5, the time of the initial entry into service

j = 6, the departure time

j = 7, the waiting time

```
* C360SKIBALGOL    6    MIN    3000  VAHE SAGHERIAN  S.R.I. E132
2INTEGER                                          COUNT,NDEP,K22,LK,MAXSS,INT,
2M,IJK,                                       LKE,MAXNQ,MAXIT,MAXIL,MAXLQ,MAXWT,NRD,
2MAXIN,MAXDL,MAXTD,RANGE,DICSIP,IQ,I,J,K,L,LL,KK,JJ,NP,II,NSERV$
2ARRAY ARRIVH(5,102),LENGTH(5,102),DQNUMH(5,52),MW2(3,52),MW3(3,52),
2LKE(5),WAB(2,102),WAITZ(5,102),MW1(3,52),DQLENH(5,52),WAITH(5,102),
2INTERH(5,52),DELAYH(5,102),SPDEPH(5,102),STEP(8),RANGE(8),LAM(5),
2MML(5),DICSIP(5),SERVER(10,7),Q(5,750,4),A(5,3),QQ(5,2),IQ(5),MST(5),
2RQ(5),INT(5),WTIME(5),LK(5),MW(5),DQLEZ(5,52),ZAB(2,102)$
2EXTERNAL PROCEDURE RANDOM($RFIX,RFLT)$
2STAR..FOR I=(1,1,5)$
2       FOR J=(1,1,102)$BEGIN ARRIVH(I,J)=LENGTH(I,J)=WAITH(I,J)=
2WAITZ(I,J)=DELAYH(I,J)=SPDEPH(I,J)=0  END$
2FOR J=(1,1,52)$BEGIN DQNUMH(I,J)=DQLENH(I,J)=INTERH(I,J)=
2DQLEZ(I,J)=0 END$
2FOR I=(1,1,5)$ BEGIN RQ(I)=INT(I)=WTIME(I)=LK(I)=MW(I)=MST(I)=IQ(I)=
2MML(I)=DICSIP(I)=LKE(I)=LAM(I)=0 END$
2FOR I=(1,1,3)$FOR J=(1,1,52)$ BEGIN MW1(I,J)=MW2(I,J)=MW3(I,J)=0 END$
2FOR J=(1,1,10)$FOR I=(1,1,7)$ SERVER(J,I)=0$
2FOR J=(1,1,5)$FOR I=(1,1,750)$FOR K=(1,1,4)$  Q(J,I,K)=0$
2FOR I=(1,1,2)$FOR J=(1,1,102)$ BEGIN WAB(I,J)=ZAB(I,J)=0 END$
2FOR I=(1,1,5)$FOR J=(1,1,3)$ A(I,J)=0$
2READ($$DATA)$
2INPUT DATA(NSERV,NP,K22, NRD,SRATE,NDEP,FOR I=(1,1,NP)$(LAM(I),MML(I),
2            DICSIP(I)),FOR I=(1,1, 8)$(RANGE(I),STEP(I)))$
2KK=NP+1$                    BIG=9**25$
2FOR I=(1,1,NSERV)$BEGIN SERVER(I,1)=KK$
2                                          FOR J=(2,1,7)$SERVER(I,J)=0     $
2              SERVER(I,6)=BIG  END$
2FOR I=(1,1,NP)$ FOR J=(1,1,750)$ FOR K=(1,1,4)$ Q(I,J,K)=0 $
2                         MAXIT=FIX(RANGE(1)/STEP(1)  )$
2                         MAXIL=FIX(RANGE(2)/STEP(2)  )$
2                         MAXNQ=FIX(RANGE(3)/STEP(3)+1)$
2                         MAXLQ=FIX(RANGE(4)/STEP(4)+1)$
2                         MAXWT=FIX(RANGE(5)/STEP(5)+1)$
2                         MAXIN=FIX(RANGE(6)/STEP(6)+1)$
2                         MAXDL=FIX(RANGE(7)/STEP(7)  )$
2                         MAXTD=FIX(RANGE(8)/STEP(8)  )$
2SUM=SUMM=COUNT=0$
2FOR I=(1,1,NP)$ BEGIN MST(I)=MML(I)/SRATE$ SUMM=SUMM+LAM(I).MML(I)$
2   SUM=SUM+LAM(I)$ AVE=SUMM/(SUM.SRATE)$ ENTER GENERATE$ INT(I)=0 END$
2CHOOSE.. TMIN=BIG$     IJK=L=0$
2 FOR I=(1 1,NSERV)$ BEGIN IF TMIN GTR SERVER(I,6)$
2            BEGIN  L=I$       TMIN=SERVER(I,6)  END END$
2    IF L GTR 0$  BEGIN  LL=L+1$
2   FOR K=(LL,1,NSERV)$  BEGIN
2IF SERVER(K,6) EQL  SERVER(L,6)$  BEGIN IF SERVER(K,1) LSS SERVER(L,1)$
2     L=K END  END  END$
2FOR I=(1,1,NP)$ BEGIN IF TMIN GTR A(I,1)$ BEGIN II=I$
2       IJK=1$  TMIN=A(I,1)  END  END$
2IF IJK EQL 1$  BEGIN SIMT=A(II,1)$J=II$   GO SKIP END$
2     SIMT=SERVER(L,6)$
2SKIP..
2 IF (COUNT GTR NDEP) OR (IQ(NP) GEQ 748)$ BEGIN
2                         WRITE($$HEAD,FMT)$ OUTPUT HEAD(COUNT,SIMT)$
```

```
2FORMAT FMT(*NUMBER OF DEPARTURES*,I5    ,B5,*SIMULATED TIME=*,F15.8,W3)$
2FOR M=(1,1,NP)$ WRITE($$ONE,TWO)$
2                OUTPUT ONE(M,FOR J=(1,1,LKE(M))$(MW1(M,J),MW2(M,J),
2MW3(M,J)))$
2FORMAT TWO(*PRIORITY=*,I1,*WAITING TIME AVERAGES*,W0,(3F20.8,W0))$
2FOR I=(1,1,NP)$ BEGIN
2SUM=NUMER=0$
2FOR K=(1,1,MAXIT+1)$ SUM=SUM+ARRIVH(I,K)$
2FOR K=(1,1,MAXIT+1)$ BEGIN NUMER=NUMER+ARRIVH(I,K)$
2                             ARRIVH(I,K)=1.0-NUMER/SUM  END$
2SUM=NUMER=0$
2FOR K=(1,1,MAXIL+1)$ SUM=SUM+LENGTH(I,K)$
2FOR K=(1,1,MAXIL+1)$ BEGIN NUMER=NUMER+LENGTH(I,K)$
2                             LENGTH(I,K)=1.0-NUMER/SUM  END$
2SUM=NUMER=0$
2FOR K=(1,1,MAXNQ+1)$ SUM=SUM+DQNUMH(I,K)$
2FOR K=(1,1,MAXNQ+1)$ BEGIN NUMER=NUMER+DQNUMH(I,K)$
2                             DQNUMH(I,K)=1.0-NUMER/SUM  END$
2SUM=NUMER=0$
2FOR K=(1,1,MAXLQ+1)$ SUM=SUM+DQLENH(I,K)$
2FOR K=(1,1,MAXLQ+1)$ BEGIN NUMER=NUMER+DQLENH(I,K)$
2                             DQLENH(I,K)=1.0-NUMER/SUM  END$
2SUM=NUMER=0$
2FOR K=(1,1,MAXWT+1)$ SUM=SUM+WAITH(I,K)$
2FOR K=(1,1,MAXWT+1)$ BEGIN NUMER=NUMER+WAITH (I,K)$
2                             WAITH(I,K)=1.0- NUMER/SUM  END$
2SUM=NUMER=0$
2FOR K=(1,1,MAXIN+1)$ SUM=SUM+INTERH(I,K)$
2FOR K=(1,1,MAXIN+1)$ BEGIN NUMER=NUMER+INTERH(I,K)$
2                             INTERH(I,K)=1.0-NUMER/SUM  END$
2SUM=NUMER=0$
2FOR K=(1,1,MAXDL+1)$ SUM=SUM+DELAYH(I,K)$
2FOR K=(1,1,MAXDL+1)$ BEGIN NUMER=NUMER+DELAYH(I,K)$
2                             DELAYH(I,K)=1.0-NUMER/SUM  END$
2SUM=NUMER=0$
2FOR K=(1,1,MAXTD+1)$ SUM=SUM+SPDEPH(I,K)$
2FOR K=(1,1,MAXTD+1)$ BEGIN NUMER=NUMER+SPDEPH(I,K)$
2                             SPDEPH(I,K)=1.0-NUMER/SUM   END$
2                        WRITE($$HIST1 ,GRAM1 )$
2                         WRITE($$HIST2 ,GRAM2 )$
2                         WRITE($$HIST4 ,GRAM4 )$
2                         WRITE($$HIST8 ,GRAM8 )$
2                         WRITE($$HIST9 ,GRAM9 )$
2                         WRITE($$HIST10,GRAM10)$
2                         WRITE($$HIST11,GRAM11)$
2                         WRITE($$HIST12,GRAM12)  END$
                                                  GO TO FIN   END$
2    OUTPUT HIST1 (I,FOR J=(1,1,MAXIT+1)$ ARRIVH(I,J))$
2    OUTPUT HIST2 (FOR J=(1,1,MAXIL+1)$ LENGTH(I,J))$
2    OUTPUT HIST4 (FOR J=(1,1,MAXNQ+1)$ DQNUMH(I,J))$
2    OUTPUT HIST8 (FOR J=(1,1,MAXLQ+1)$ DQLENH(I,J))$
2    OUTPUT HIST9 (FOR J=(1,1,MAXWT+1)$  WAITH(I,J))$
2    OUTPUT HIST10(FOR J=(1,1,MAXIN+1)$ INTERH(I,J))$
2    OUTPUT HIST11(FOR J=(1,1,MAXDL+1)$ DELAYH(I,J))$
2    OUTPUT HIST12(FOR J=(1,1,MAXTD+1)$ SPDEPH(I,J))$
```

```
2FORMAT GRAM1(*PRIORITY*,I3,H2,*CUMULATIVE PROBABILITY DISTRIBUTION*,W4,
2              *INTER-ARRIVAL TIME*,W4,W0,(20X6.3,W0))$
2  FORMAT GRAM2(W4,*MESSAGE LENGTH*,W4,W0,(20X6.3,W0))$
2FORMAT GRAM4(W4,*MESSAGES IN QUEUE AT DEPARTURE TIME*,W4,(20X6.3,W0))$
2FORMAT GRAM5(W4,*LENGTH OF QUEUE AT DEPARTURE TIME*,W4,W0,(20X6.3,W0))$
2  FORMAT GRAM9(W4,*WAITING TIME*,W4,W0,(20X6.3,W0))$
2  FORMAT GRAM10(W4,*INTERRUPTIONS*,W4,W0,(20X6.3,W0))$
2  FORMAT GRAM11(W4,*DELAY TIME*,W4,W0,(20X6.3,W0))$
2  FORMAT GRAM12(W4,*INTER-DEPARTURE TIME OF SAME PRIORITY*,W4,W0,
2                (20X6.3,W0))$
2IF IJK EQL 1$ BEGIN          IF INT(J) NEQ 0$
2BEGIN FK=A(J,3).LAM(J)/STEP(1)$  K=FIX(FK)$
2IF (FK-K) GTR 0$             K=K+1$
2                       IF K GTR MAXIT   $ K=MAXIT   +1$
2                       ARRIVH(J,K)=ARRIVH(J,K)+1 END$
2              INT(J)=1$
2FK=A(J,2)/(MML(J).STEP(2))$  K=FIX(FK)$
2IF (FK-K) GTR 0$             K=K+1$
2                IF K GTR MAXIL   $ K=MAXIL   +1$
2                LENGTH(J,K)=LENGTH(J,K)+1$
2FOR II=(1,1,NSERV)$BEGIN IF SERVER(II,1) EQL KK$ GO BELOW END$
2        IF DICSIP(J) EQL 1$ BEGIN
2UP..IQ(J)=IQ(J)+1$ K=IQ(J)$ Q(J,K,1)=A(J,1)$ Q(J,K,2)=A(J,2)$
2Q(J,K,3)=Q(J,K,4)=0$
2                     QQ(J,1)=QQ(J,1)+A(J,2)$
2    I=J$ ENTER GENERATE$  GO CHOOSE END$
2I=J$ LL=0$ FOR K=(1,1,NSERV)$ BEGIN
2           IF SERVER(K,1) GTR I$ BEGIN LL=K$ I=SERVER(K,1) END END$
2IF LL EQL 0$  GO UP$
2L=LL+1$ FOR K=(L,1,NSERV)$
2          BEGIN IF SERVER(K,1) EQL SERVER(LL,1)$
2  BEGIN IF SERVER(K,3) GTR SERVER(LL,3) $  LL=K  END  END$
2K=SERVER(LL,1)$
2FOR I=(IQ(K),-1,1)$ BEGIN L=I+1$ Q(K,L,1)=Q(K,I,1)$ Q(K,L,2)= Q(K,I,2)$
2Q(K,L,4)=Q(K,I,4)$          Q(K,L,3)=Q(K,I,3)   END$
2                 Q(K,1,1)=SERVER(LL,3)$ Q(K,1,3)=SERVER(LL,2)+1.0$
2  Q(K,1,4)=SERVER(LL,7)$
2IF DICSIP(J) EQL 2$ BEGIN
2  Q(K,1,2)=(SERVER(LL,6)-SIMT).SRATE$  GO DOWN  END$
2       Q(K,1,2)=SERVER(LL,4)$
2DOWN..                                              IQ(K)=IQ(K)+1$
2       QQ(K,1)=QQ(K,1)+Q(K,1,2)$ SERVER(LL,1)=J$
2       SERVER(LL,3)=SERVER(LL,5)=SIMT$ SERVER(LL,4)=A(J,2)$
2  SERVER(LL,7)=0$
2       SERVER(LL,2)=0$ I=J$ ENTER GENERATE$ DT=SERVER(LL,4)/SRATE$
2  SERVER(LL,6)=SIMT+DT$      GO CHOOSE$
2BELOW..  SERVER(II,7)=0$
2SERVER(II,1)=J$  SERVER(II,3)=SERVER(II,5)=SIMT$ SERVER(II,4)=A(J,2)$
2SERVER(II,2)=0$  I=J$ ENTER GENERATE$
2DT=SERVER(II,4)/SRATE$
2 SERVER(II,6)=SIMT+DT$
2    GO CHOOSE  END $
2 J=L$  COUNT=COUNT+1$
2             M=SERVER(J ,1)$ SERVER(J ,1)=KK$ K=SERVER(J ,2)+1$
2    IF K GTR MAXIN   $ K=MAXIN   +1$ INTERH(M,K)=INTERH(M,K)+1$
```

```
2FK=SERVER(J,7)/(MST(M).STEP(5))+1$  K=FIX(FK)$
2 IF (FK-K) GTR 0$ K=K+1$  IF K GTR MAXWT$  K=MAXWT+1$
2WAITZ(M,K)=WAITZ(M,K)+1$
2  WAITH(M,K)=WAITH(M,K)+1$
2FOR L=(1,1,NP)$ BEGIN I=IQ(L)+1$
2   IF I GTR MAXNQ   $ I=MAXNQ   +1$
2         DQNUMH(L,1)=DQNUMH(L,1)+1 END$
2FOR II=(1,1,NP)$ RQ(II)=QQ(II,1)$
2FOR II=(1,1,NSERV)$BEGIN
2K=SERVER(II,1)$ IF K NEQ KK$ BEGIN DT=SIMT-SERVER(II,5)$
2LONG=SERVER(II,4)-DT,SRATE$ RQ(K)=RQ(K)+LONG                    END END$
2FOR L=(1,1,NP)$ BEGIN FK=RQ(L)  /MML(L)+1$  K=FIX(FK)$
2IF (FK-K) GTR 0$            K=K+1$
2        IF K GTR MAXLQ   $ K=MAXLQ   +1$
2   DQLEZ(L,K)=DQLEZ(L,K)+1 $
2    DQLENH(L,K)=DQLENH(L,K)+1 END$
2LK(M)=LK(M)+1$ WTIME(M)=WTIME(M)+SERVER(J,7)$ K=LK(M)$
2IF MOD(K,KZ2 ) EQL 0$ BEGIN LKE(M)=LKE(M)+1$   L=LKE(M)$
2MW1(M,L)=WTIME(M)/KZ2 $ MW(M)=MW(M)+WTIME(M)$  MW2(M,L)=MW(M)/K$
2MW3(M,L)=MW(M)/(K-WAITH(M,1))$ WTIME(M)=0    $
2SUM=0$ FOR JJ=(1,1,MAXWT+1)$BEGIN SUM=SUM+WAITZ(M,JJ)$
2WAITZ(M,JJ)=0$ WAH(1,JJ)=1.0-SUM/KZ2  END$
2NUMER=0.0$FOR JJ=(1,1,MAXLQ+1)$NUMER=NUMER+DQLEZ(M,JJ)$
2SUM=0$ FOR JJ=(1,1,MAXLQ+1)$ BEGIN SUM=SUM+DQLEZ(M,JJ)$
2DQLEZ(M,JJ)=0$ ZAH(1,JJ)=1.0-SUM/NUMER  END$
2SUM=0$ FOR JJ=(1,1,MAXWT+1)$ BEGIN SUM=SUM+WAITH(M,JJ)$
2WAH(2,JJ)=1-SUM/K END$ MAXSS=MAXWT+1$
2NUMER=0.0$FOR  JJ=(1,1,MAXLQ+1)$NUMER=NUMER+DQLENH(M,JJ)$
2SUM=0$ FOR JJ=(1,1,MAXLQ+1)$ BEGIN SUM=SUM+DQLENH(M,JJ)$
2ZAR(2,JJ)=1-SUM/NUMER  END$
2WRITE($$NINE,TEN)$
2WRITE($$ELEV,TWEL)$OUTPUT ELEV(FOR  JJ=(1,1,MAXSS)$WAH(2,JJ))$
2MAXSS=MAXLQ+1$
2WRITE($$FTEEN,TWEL)$ OUTPUT FTEEN(FOR JJ=(1,1,MAXSS)$ZAB(1,JJ))$
2WRITE($$STEEN,TWEL)$ OUTPUT STEEN(FOR JJ=(1,1,MAXSS)$ZAB(2,JJ))$
2FORMAT TWEL(WO,(2UX6.3,WO))$
2OUTPUT NINE(K,M,SIMT,FOR JJ=(1,1,MAXSS)$(WAB(1,JJ)                  ))$
                                                                  I=M$MAXSS=J$
2
2WRITE($$HIST9,GRAM9)$
2J=MAXSS END$
2FORMAT TEN(*DEPARTURE=*,I4,*PRIORITY=*,I1,*SIMT=*,F15.8,WO,
2(20X6.3,WO))$
2 DELAY=SERVER(J,6)-SERVER(J,3)$
2   FK=DELAY/(MST(M).STEP(7))$  K=FIX(FK)$
2IF (FK-K) GTR 0$            K=K+1$
2        IF K GTR MAXDL   $ K=MAXDL   +1$ DELAYH(M,K)=DELAYH(M,K)+1$
2IF QQ(M,2) NEQ 0$ BEGIN SPD=SERVER(J,6)-QQ(M,2)$
2   FK=SPD/(MST(M).STEP(8))$  K=FIX(FK)$
2IF (FK-K) GTR 0$            K=K+1$
2    IF K GTR MAXTD   $ K=MAXTD   +1$SPDEPH(M,K)=SPDEPH(M,K)+1 END$
2   QQ(M,2)=SERVER(J,6)$
2   SERVER(J,6)=BIG$
2  FOR I=(1,1,NP)$ BEGIN IF IQ(I) GEQ 1$ BEGIN SERVER(J ,1)=I$
2                SERVER(J ,5)=SIMT$ SERVER(J ,4)=Q(I,1,2)$
2  SERVER(J,3)=Q(I,1,1)$
```

```
2 DT=SERVER(J,4)/SRATE$  SERVER(J,6)=DT+SIMT$
2 SERVER(J,7)=Q(I,1,4)$
2  SERVER(J ,2)=Q(I,1,3)$
2 IF SERVER(J,2) EQL 0$  SERVER(J,7)=SIMT-Q(I,1,1)$
2FOR L=(1,1,IQ(I))$ BEGIN Q(I,L,1)=Q(I,L+1,1)$ Q(I,L,2)=Q(I,L+1,2)$
2  Q(I,L,4)=Q(I,L+1,4)$
2   Q(I,L,3)=Q(I,L+1,3) END$
2  IQ(I)=I (I)-1$ QQ(I,1)=QQ(I,1)-SERVER(J ,4)$
2                          GO CHOOSE END  ENDS
2LEIL..
2  GO TO CHOOSES
2SUBROUTINE  GENERATES
2BEGIN  RANDOM($NRD,Y)$
2       T=-LOG(Y)/LAM(I)$
2        A(I,1)=A(I,1)+T$
2           RANDOM($NRD,F)$
2     A(I,2)=-LOG(F).MML(I)$
2           A(I,3)=T$
2               RETURN ENDS
2FIN..GO STAR$
2FINISH$
<3 (S2  >       (S$A
AU "2(-A    (  )2U9BU9(   ( 6< >4 >)UB54( 3(-
2     FINISH$
```

B. Input Data Card Format

For each simulation run it is necessary to specify the following input data:

Number of servers (NSERV)

Number of priority category (NP)

Number of departures at which time system data are sampled (KZ2)

Service rate of each server (SRATE)

Duration of simulation-number of departures (NDEP)

Mean arrival rate (LAM)
Mean message length (MML)
Queuing discipline (DISCIP)
} of each priority category

Range
Step size
} of each histogram

These data are contained in a sequence of input data cards. The format of these cards is shown in Fig. C-1. The cards are to be arranged in the order of the card number. The format of each input data is specified as either integer (I) or floating point (FP).

Card No.	Item No.	Data Description	Format	Remarks
1	1	NSERV	I	
	2	NP	I	
	3	KZ2	I	
	4	NRD	I	10 digit number not divisible by 2 or 5
	5	SRATE	I	
	6	NDEP	I	
2	1	LAM(1)	FP	First Priority Parameter Card
	2	MML(1)	FP	
	3	DISCIP(1)	I	
NP + 1	1	LAM(N)	FP	Nth Priority Parameter Card
	2	MML(N)	FP	
	3	DISCIP(N)	I	
NP + 2	1	RANGE	I	Range and Step for following histograms: ARRIVH, LENGTH, DQNUMH, DQLENH, WAITH, INTERH, DELAYH and SPLEPH respectively.
	2	STEP	FP	
NP + 9	1	RANGE	I	
	2	STEP	FP	

Fig. C-1

INPUT DATA CARD FORMAT

C. Typical Simulation Output

The output of the 1M/1Pi/Sm, where S = 2, 4, 10, simulation runs is used for illustration purpose. In this series of simulation runs the system data, the waiting time and length of queue, is sampled at every 200 departures of a priority category. A typical output at this sampling time is as shown in Fig. C-2. It is for the first run of the 1M/1Pi/2M case. The sample is taken at the 2200 departure of the second priority message and the simulated time is 2189.1565 time unit. The four rows of numbers represents the cumulative probability distribution of the waiting time for the last 200 departures; cumulative probability distribution of the waiting time for the entire 2200 departures, cumulative probability distribution of the length of queue for the last 200 departures and cumulative probability distribution of the length of queue for the entire 2200 departures.

At the end of the simulation run, after 6000 departures and at simulated time of 2963.5438 time units, the printout is as shown in Fig. C-3. The average waiting time for each priority category is computed for: (1) the last 200 departures, (2) all the departures and (3) all departures that have to wait at each sampling time. The cumulative probability distributions of the input parameters and various operational measures for each priority category are printed in the order of priority level.

```
DEPARTURE=2200PRIORITY=2SIMT=  2.1891565, 01
  .710  .505  .385  .310  .245  .190  .170  .140  .120  .080  .055  .025  .020  .000  .000  .000  .000  .000  .000  .000
  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000

  .728  .549  .431  .345  .270  .210  .172  .132  .104  .079  .055  .039  .030  .019  .013  .010  .008  .004  .002  .000
  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000

  .824  .666  .493  .411  .339  .272  .188  .146  .129  .106  .074  .040  .020  .015  .010  .010  .002  .000  .000  .000
  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000

  .833  .678  .550  .443  .355  .301  .238  .181  .141  .106  .077  .057  .040  .031  .020  .016  .011  .008  .006  .004
  .001  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000
```

Fig. C-2

TYPICAL OUTPUT OF THE SIMULATION AT SAMPLING TIME

```
NUMBER OF DEPARTURES 6001      SIMULATED TIME=  2.9635438E 03
PRIORITY=1WAITING TIME AVERAGES
        3.3209004E-01          3.3209004E-01          1.0888198E 00
        6.0114259E-02          1.9610215E-01          8.1709230E-01
        1.0730647E-01          1.6650356E-01          6.9861632E-01
        1.6100136E-01          1.6512801E-01          7.0642997E-01
        1.3976196E-01          1.6005440E-01          6.6138346E-01
        9.2955208E-02          1.4487153E-01          6.1602014E-01
        1.8200767E-01          1.5360527E-01          6.1267058E-01
        1.6359238E-01          1.5485366E-01          6.1480360E-01
        2.2253578E-01          1.6237389E-01          6.4094957E-01
        2.0259514E-01          1.6639602E-01          6.5125642E-01
        2.3937651E-01          1.7306243E-01          6.6434788E-01
        7.4849854E-02          1.6495054E-01          6.5395257E-01
        1.6960144E-01          1.6521600E-01          6.4740587E-01
        5.7892609E-02          1.5755504E-01          6.2460472E-01
PRIORITY=2WAITING TIME AVERAGES
        1.8824239E 00          1.8829249E 00          2.6704141E 00
        1.8330695E 00          1.8579967E 00          2.5452010E 00
        1.9074497E 00          1.8744977E 00          2.5795960E 00
        1.2644451E 00          1.7232146E 00          2.4750227E 00
        3.4649164E 00          2.0714409E 00          2.9093974E 00
        2.6753707E 00          2.1721376E 00          3.0308896E 00
        2.3653590E 00          2.1997406E 00          3.0371172E 00
        1.9701242E 00          2.1710386E 00          3.0153314E 00
        3.1908382E 00          2.2739052E 00          3.1446968E 00
        2.2000908E 00          2.2713237E 00          3.1143674E 00
        2.0491110E 00          2.2516680E 00          3.0961097E 00
        1.5285496E 00          2.1916081E 00          3.0122443E 00
        1.8591075E 00          2.1658465E 00          2.9889602E 00
        2.1283924E 00          2.1560294E 00          2.9767651E 00
        1.9447527E 00          2.1419433E 00          2.9516304E 00

PRIORITY 1  CUMULATIVE PROBABILITY DISTRIBUTION

INTER-ARRIVAL TIME

  .802  .664  .537  .446  .369  .298  .242  .195  .160  .132  .105  .084  .070  .057  .047  .039  .034  .029  .024  .021
  .014  .012  .010  .009  .009  .008  .007  .005  .005  .004  .004  .004  .003  .002  .002  .001  .000  .000  .000

MESSAGE LENGTH

  .812  .667  .543  .449  .362  .292  .234  .204  .167  .135  .113  .095  .078  .067  .055  .043  .036  .028  .023  .020
  .016  .013  .011  .009  .007  .005  .004  .004  .004  .004  .003  .003  .003  .002  .002  .001  .001  .001  .000

MESSAGES IN QUEUE AT DEPARTURE TIME
  .128  .053  .020  .007  .003  .001  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000
  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000

LENGTH OF QUEUE AT DEPARTURE TIME

  .463  .238  .125  .067  .035  .017  .009  .004  .002  .001  .001  .001  .000  .000  .000  .000  .000  .000  .000  .000
  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000

WAITING TIME

  .257  .077  .023  .005  .001  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000
  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000  .000
                                                                                                                  SC-4043-8
```

Fig. C-3

TYPICAL PRINTOUT OF THE SIMULATION AT THE END OF SIMULATION RUN

INTERRUPTIONS

.000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000

DELAY TIME

.445 .190 .082 .033 .012 .005 .003 .001 .001 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000
.000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000
.000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000

INTER-DEPARTURE TIME OF SAME PRIORITY

.445 .190 .081 .038 .020 .008 .003 .003 .001 .001 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000
.000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000

PRIORITY 2 CUMULATIVE PROBABILITY DISTRIBUTION

INTER-ARRIVAL TIME

.824 .664 .542 .441 .356 .287 .230 .185 .150 .120 .100 .080 .072 .060 .047 .038 .032 .026 .023 .019
.016 .013 .011 .010 .009 .008 .006 .005 .004 .003 .002 .002 .002 .002 .002 .001 .001 .001 .000

MESSAGE LENGTH

.819 .677 .560 .459 .369 .298 .245 .205 .166 .134 .106 .087 .070 .058 .047 .038 .031 .024 .020 .016
.014 .011 .009 .008 .006 .004 .004 .002 .002 .002 .001 .001 .001 .001 .001 .001 .000 .000 .000

MESSAGES IN QUEUE AT DEPARTURE TIME

.730 .591 .465 .361 .267 .199 .149 .111 .086 .064 .052 .038 .029 .020 .013 .010 .007 .006 .004 .002
.001 .001 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000

LENGTH OF QUEUE AT DEPARTURE TIME

.831 .675 .538 .429 .340 .280 .218 .168 .132 .097 .071 .053 .038 .028 .018 .012 .009 .006 .004 .003
.001 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000

WAITING TIME

.726 .538 .419 .323 .254 .201 .164 .123 .095 .071 .050 .033 .025 .015 .010 .008 .006 .003 .002 .000

.000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000

INTERRUPTIONS

.330 .110 .034 .011 .003 .001 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000

DELAY TIME

.778 .618 .489 .392 .309 .249 .202 .158 .123 .092 .064 .047 .031 .022 .014 .011 .008 .004 .002 .002
.001 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000
.000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000 .000

RC-4013-9

Fig. C-3

(Continued)