STOCHASTIC SERVICE SYSTEMS
WITH PRIORITIES

by

L. Durr, B.A., M.A.

A thesis submitted to the Faculty of Science in
partial fulfillment of the requirements for the
degree of Master of Science

Department of Electrical Engineering
Faculty of Pure and Applied Science
The University of Ottawa
Ottawa, Canada
1969

ABSTRACT

This project is concerned with the effects of input priorities on the service given by stochastic service systems. By examining these effects a more efficient method of assignment can be found.

Investigations are carried out into a number of service schemes in use or proposed for future use. The service disciplines discussed offer a variety of techniques by which the waiting times of the different classes of inputs can be manipulated or adjusted to meet a set of operational requirements.

Single and multiple server systems with either non-preemptive or preemptive priority disciplines are considered. A variety of queue disciplines are also taken into consideration.

ACKNOWLEDGMENTS

The author wishes to express his gratitude and appreciation to Professor G. S. Glinski, Chairman of the Department of Electrical Engineering, for his direction and encouragement during the course of this research project.

The author is indebted to the Northern Electric Company for the financial support through which this thesis was made possible and the members of the staff who offered valuable and helpful criticism.

The author is indebted to Miss C. A. Matthews for typing this thesis.

TABLE OF CONTENTS

Chapter I

Introduction

1. Stochastic System

A stochastic system can be defined as a system whose load or input and whose internal operations are subject to random events. These random events constitute stochastic processes when the time in which they occur is considered.

There are many methods by which the performance of complex stochastic systems can be analysed. The methods can be grouped into three main categories:

a) Mathematical modeling and analysis.

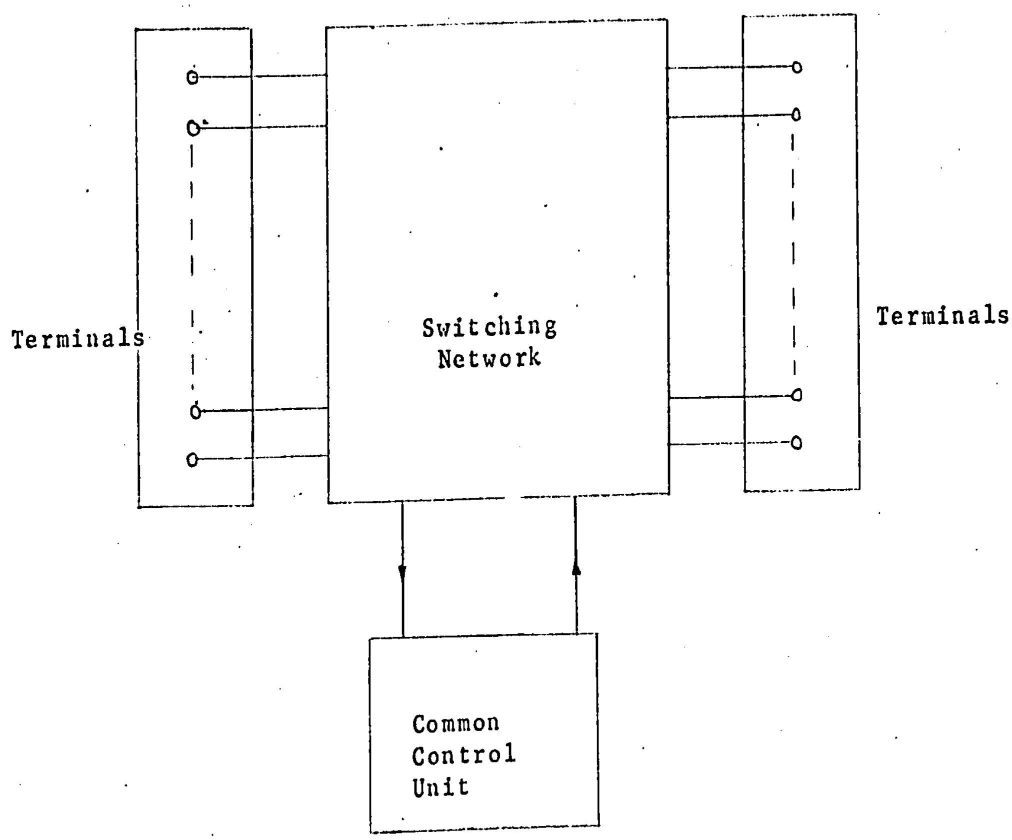
b) Simulation models and their subsequent analysis

c) The analysis of data obtained by measurements on the system during productive operation.

Each method introduces its own disadvantages and ideally a combination of all three methods should be used. In the initial design stage, however, the physical system is not available so that measurements cannot be made. Also as the systems considered become more complex the potential cost and accuracy or realism of representation of the three methods usually increases in the order that they have been mentioned.

In consideration of developing a general theory that may be carried over to other applications we will study the stochastic systems by means of mathematical modeling and analysis.

In this research, the stochastic systems of interest to us are telephone switching systems and the subscribers receiving service from them.



Telephone Switching System
Fig 1.1

As shown in Fig 1.1 a telephone switching system can be subdivided into:

a) A set of terminals

b) Common control equipment which processes requests for connection

c) A switching network through which connections are made.

The system operates in the following manner:

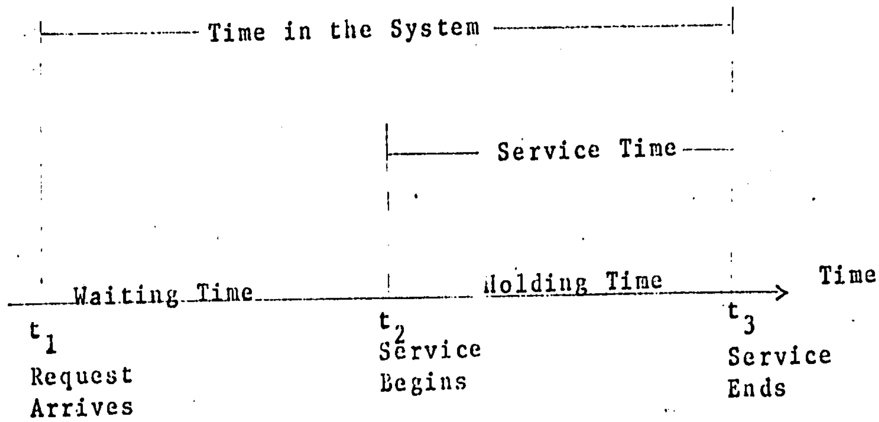
a) Calls or requests for connection between idle terminals arise.

b) Requests are processed by the control unit and desired connections are completed, if possible, through the network.

c) Calls remain in the network until communications ends.

d) Terminals return to an idle condition when a call terminates.'

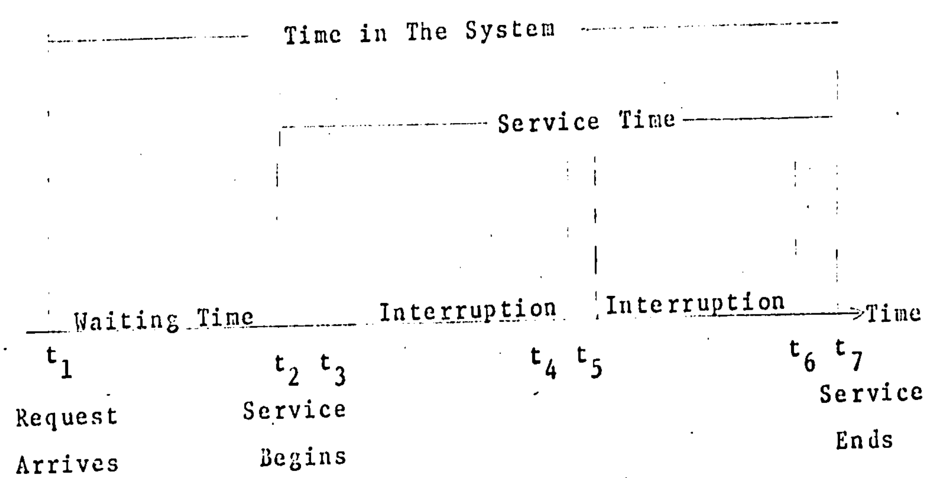
Most modern telephone systems follow this basic pattern. The call processing can thus be split into two distinct types of models: loss and delay systems. All connections through the network are made on a loss basis that is if a connection cannot be made at the first attempt no further attempt is made. Connections to the control unit are made on a delay basis that is if a request for service arrives when the control unit is busy, the request is placed in a queue (waiting line) and waits until the control unit becomes available. If the request is the only item in the queue it then siezes the control unit and uses it for a full holding time. If the request is not the only item in the queue, the control unit may or may not serve it first depending upon the queue discipline. Thus, requests for service (calls) which arrive when the control unit is busy are said to meet congestion and are forced to wait. Contrary to the loss system these calls are not discarded. A call arriving in the system, at time t_1, will be serviced at time t_2, where $t_2 \geq t_1$ and will have completed its service at time t_3, where $t_3 \geq t_2$. This is illustrated in Fig 1.2. The time interval (t_2-t_1) is known as the waiting time. The time interval (t_3-t_2) has been denoted as the service time. The time spent by the control unit in servicing a call is known as the holding time. Fig 1.2 shows a system in which there are no interruptions of service so that the holding time and the service time are the same. Fig 1.3 illustrates the case where the call in service is interrupted and resumed when the interruption finishes. In this type of system, the holding time is now the sum of the intervals (t_3-t_2), (t_5-t_4) and (t_7-t_6).



Relationship Between Waiting and Holding Times

Without Interruptions

Fig 1.2



Relationship Between Waiting and Holding Times

With Interruptions

Fig 1.3

Previously the majority of the effort has been concentrated into loss systems [1-5] and although a great wealth of literature is available on delay systems [1, 6-10] relatively few are directly applicable to telephone systems. In view of this, the specific subject of analysis in which we are interested will be certain delay systems which are of interest in serving calls in telephone systems. The basic model is distinguished by the method in which calls are processed by the control unit of the system.

The control unit consists of parts that are arranged in a manner reflecting their function and are determined by the operations necessary to establish a connection and by the philosophy of the design and the technology which are basic to the system. To establish a connection the control unit performs some or all of the following steps:

a) If there is more than one request for service, decides by means of a priority system which request to serve first.

b) Identifies the physical location of the terminal or calling party.

c) Translates the information received from the calling party into the location of the called party.

d) Tries to complete the connection.

As can be seen from this the control unit is basically a data processing system. It collects information about desired connections, digests it, makes routing decisions and issues orders for completing requested calls in the network.

Its performance can be described by the probability distribution of delay in setting up a desired connection.

The control unit can thus be modeled as a queuing system with priority discipline. The model considered is of the type where the control is represented only in a macroscopic manner. That is operation of the control as a whole is modeled not the specific operations of the various sections of the control.

2. Summary

In the research, the effect of input priorities on the service given by a control system with various queue disciplines is investigated. Mathematical analysis of a number of possible service schemes in use or proposed for future use are presented with the aim that the examination of the effects of assigning priorities will lead to a more efficient method of assignment.

The investigation is carried out as follows:

In chapter II, the subject of the work is placed in a practical setting. The parameters of study that is the input process, the queue discipline and the holding time distributions assumed are discussed. Chapter III is a review of the priority system models considered previously in the literature. In Chapter IV, the effects of assigning priorities when the queue discipline is last-in first-out (LIFO) are considered. The priority disciplines analysed are (1) non-preemptive and (2) preemptive resume. Systems in which requests for service are taken from the queue at random are considered in Chapter V. Chapter VI extends the discussion to multiserver systems. Future lines of investigation are suggested in Chapter VII.

Chapter II

Parameters of the Model

1. Introduction

When choosing a mathematical model for a physical system, one is confronted with two requirements which generally oppose each other. One would like to describe the system exactly yet if this is done, the model becomes so complicated that it becomes impossible to analyse. The model analysed may therefore involve departures from reality; a model that is sufficiently amenable to mathematical analysis often results only after one has introduced admittedley false assumptions, ignored certain effects and correlations and generally simplified the system to be studied.

Since we are interested in telephone traffic it seems reasonable that we should take account of the following significant factors

(a) Demand for service or the input process,

(b) The length of the holding time or the service mechanism,

(c) How a request for service is chosen from amongst a number of simultaneous requests or the queue discipline,

(d) The manner in which a - c are interrelated.

Unfortunately the mathematical complexity of such a realistic model precludes investigation. Therefore the model used only considers a - c.

2. Demand for Service

The demand for telephone traffic is usually made precise by describing a stochastic process that represents the way in which requests for telephone service occur in time. A realistic description will take account of the facts that the demand is not constant but has daily extremes and that in small systems the demand may be materially lessened when many calls are simultaneously in progress. This is usually referred to as the "finite sources" effect. In this research, to reduce the complexity of the mathematics we have ignored these two facts with the proviso that the results we derive are only applicable to systems for which the demand is nearly constant and which have a large number of subscribers.

With these assumptions a mathematically representative description of the demand is specified by the condition that the time intervals between requests for service have lengths that are mutually independent positive random variables with a negative exponential distribution. This is equivalent to saying that the call arrivals form a Poisson process as can be seen from the following: Denote by X_1, $X_2 \ldots$ mutually independent random variables with the common exponential distribution

$$F(t) = 1-e^{-\lambda t} \qquad t \geq 0 \qquad (2.1)$$

Let the sum $S_n = X_1 + X_2 + \ldots + X_n$ have density g_n and distribution function G_n

Then $$g_n = \frac{\lambda(\lambda t)^{n-1} e^{-\lambda t}}{(n-1)!} \qquad t>0 \tag{2.2}$$

and $$G_n = 1 - e^{-\lambda t}\left(1+\lambda t+ \dots + \frac{(\lambda t)^{n-1}}{(n-1)!}\right) \qquad t>0 \tag{2.3}$$

Now introduce a family of new random variables $N(t)$ as follows:

$N(t)$ is the number of indices $k \geqslant 1$ such that $S_k \leqslant t$. The event $\{N(t) = n\}$ occurs if and only if $S_n \leqslant t$ but $S_{n+1}>t$. As S_n has the distribution G_n the probability of this event equals $G_n(t) - G_{n+1}(t)$

ie. $$P\{N(t) = n\} = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \tag{2.4}$$

ie. the random variable $N(t)$ has a Poisson distribution with expectation λt.

3. Holding Time

The time that a request for service occupies a unit of equipment is called the holding time of the unit. The holding times of successive calls are assumed to be statistically independent of one another and of the input process. The distribution of the holding time is considered to be negative exponential except in Chapter IV where it has been possible to ease this restriction and use a general holding time distribution for the model considered.

It is important to note that when a negative exponential distribution is assumed for the holding time then the residual lifetime also has an exponential distribution. This 'memoryless' property of the exponential distribution can be shown in the following manner:

Assume that we have a random variable t with the distribution:

$$F(t) = \begin{cases} 0 & t < 0 \\ 1 - e^{-\mu t} & t \geq 0 \end{cases} \tag{2.5}$$

where μ is a real number such that $\mu>0$. The corresponding density function exists and is given by

$$f(t) = \frac{dF(t)}{dt} = \begin{cases} 0 & t < 0 \\ \mu e^{-\mu t} & t \geq 0 \end{cases} \tag{2.6}$$

Suppose that we are given that $t > T$. For example, consider a call with an exponentially distributed holding time and suppose it is known a priori or a posteriori that its holding time exceeds T secs. It is required to find the distribution of the interval remaining, $t_1 = (t-T)$. The distribution required is:

$$H(x,T) = Pr\{t_1 < x / t > T\}$$

$$= \frac{Pr\{T+x > t > T\}}{Pr\{t > T\}}$$

ie.

$$H(x,T) = \frac{\int_T^{T+x} \mu e^{-\mu t}\, dt}{\int_T^{\infty} \mu e^{-\mu t}\, dt}$$

$$= 1 - e^{-\mu x} \qquad (2.7)$$

which is independent of T. In effect, the above equation states that no matter how much of an exponentially distributed holding time has elapsed, the time remaining is also exponentially distributed with the same mean.

4. The Queue Discipline

There are two main types of telephone systems: electronic and electro-mechanical. The term electronic is perhaps a misnomer in the sense that there still may be electromechanical components in an electronic system. The essential difference between the two types of system lies in the type of logic used. The electronic system relies on a large digital memory to aid in processing calls whereas the electromechanical system depends upon "wired" memory. The electronic system control thus has the ability to remember the order in which requests for service arise. Thus the most appropriate queue disciplines to study are first-in first-out (FIFO) and last-in first-out (LIFO). The electromechanical system on the other hand has no memory of the order of arrival of requests for service and thus, the more appropriate queue discipline to study, is the one where requests are taken from the queue at random. There are also a number of variations of the priority discipline which should be studied to try to optimize the assignment of the priorities. An example of a priority discipline would be that a control unit may give incoming calls priority over originated calls. In the electronic system, calls can be interrupted, the information stored and when the interrupt service is finished, control is returned to the interrupted service. This is known as preemptive resume priority.

The priority disciplines considered in this thesis are:

a) Non preemptive - policy: If a call of higher priority than the one being served arrives in the system the current service is never interrupted.

b) Preemptive - resume policy: The current service can be interrupted if a call of higher priority arrives in the system. The interrupted call is continued from the point at which it was interrupted after no calls of higher priority remain in the system.

CHAPTER III

A REVIEW OF SINGLE SERVER PRIORITY SYSTEMS

1. Introduction

The problem of the optimum assignment of priorities in service systems has appeared in many forms and used in many different applications. The problem is certainly not unique to telephone systems. In the past, the problem has arisen in road traffic studies, in transport systems such as aircraft waiting to land, in hospital waiting rooms, in repair and maintenance and in many other activities. The results found from each of the models considered can usually be applied very widely.. With this in mind we review the results of previous work and record it in the terminology of queueing systems without reference to the application.

2. The Review

The problem of priorities was first considered by Cobham[11,12]. He considered a single server system with a non preemptive priority scheme. The queue discipline within a priority group was first-in first-out (in this chapter it will be understood that the queue discipline is first-in first-out (FIFO) unless otherwise stated). In a simple manner Cobham derived the equilibrium expected waiting times. He assumed that the calls arrive at random but with the calling rates λ_p depending upon the priority

$p(p = 1, 2, \ldots, k)$ such that the total calling rate λ is given by

$$\lambda = \sigma_k \tag{3.1}$$

where

$$\sigma_p = \sum_{m=1}^{p} \lambda_m \tag{3.2}$$

Each priority group was assumed to have its own distribution function $H_p(t)$ of holding time with mean $1/\mu_p$. The traffic intensity ρ_p is defined by:

$$\rho_p = \sum_{m=1}^{p} \lambda_m/\mu_m \tag{3.3}$$

The combined holding time distribution was defined by:

$$H(t) = \frac{1}{\lambda} \sum_{p=1}^{K} \lambda_p H_p(t) \tag{3.4}$$

The average waiting time W_p of priority p type calls was found to be:

$$W_p = \frac{\lambda/2 \int_0^{\infty} t \, dH(t)}{(1-\rho_p)(1-\rho_{p-1})} \tag{3.5}$$

It can be seen from this that the lower the priority number p (i.e. the higher the actual priority of the call) the lower the value of W_p.

In deriving this result Cobham used an infinite iterative procedure. Holley [13] established the same result by means of a finite process. He also showed that the average length of the queue, Q, was:

$$Q = \sum_{p=1}^{k} \lambda_p W_p \tag{3.6}$$

Dressin and Reich [14] considered a special case of Cobham's model. They assumed that the holding time distribution was negative exponential. The state equations were set up in the form of differential equations. Using these, they showed that the density function for the probability that a test call of priority p will have to wait, conditional on the assumption that there are (n-1) calls of priority p or higher waiting before the instant of the arrival of the test call, is the n fold convolution of the corresponding density when n=1.

The rigorous analysis of the priority problem for the case of arbitrary holding time distribution was performed by Kesten and Runnenberg [15]. They considered the conditional distribution function $W_{p,n}(t)$ of the waiting time of the nth. departing call, given that this call is of the priority p type and showed that this distribution tends to a limit which is independent of the queue present when the channel opens. The limiting distribution $W_p(t)$ of the pth priority group has the following Laplace-Stieltjes Transform:

$$W_p^*(s) = \int_0^\infty e^{-st}\, dW_p(t)$$

$$= \frac{(1-\rho)z_p + \sum_{m=p+1}^{k} \lambda_m(1-H_p^*(Z_p))}{S+\lambda_p-\lambda_p H_p^*(Z_p)} \qquad (3.7)$$

where Z_p satisfies

$$Z_p = s+\sigma_{p-1} - \sum_{i=1}^{p-1} \lambda_i H_i^*(Z_p) \qquad (3.8)$$

Kesten and Runnenberg also found the moments of the distribution function. The first moment or the average waiting time is the same as that found by Cobham.

In [16], Chapter 9, Morse considered the single server queue with two priority types each with a different exponential holding time distribution. He found the generating function for the probabilities of the number of priority and non priority items in the system.

Defining P(m,n) as the probability that there are m priority 1 type calls with average holding time $1/\mu_1$ and n priority 2 type calls with average holding time $1/\mu_2$ in the system, Morse showed that

$$G(s,t) = \sum_{m,n} P(m,n)\, s^m t^n$$

$$= \frac{\mu_1(s-1)(1-\lambda_1/\mu_1-\lambda_2/\mu_2)}{s(1-s)+\lambda_2 s(1-t)+\mu_1(s-1)}$$

$$+ \frac{B(t)x\{\mu_1 t(s-1)-\mu_2 s(t-1)\}}{t\{x-\lambda_1 s\}\{\lambda_1(1-s)s+\lambda_2 s(1-t)+\mu_1(s-1)\}} \qquad (3.9)$$

where

$$x = \lambda_1+\lambda_2(1-t)+\mu_2 \qquad (3.10)$$

$$B(t) = \frac{t\{1-\lambda_1/\mu_1-\lambda_2/\mu_2\}\{x-\mu_2-\mu_1\alpha(t)\}\{x(\mu_1-\mu_2)-\lambda_1\mu_1\}}{x\{\lambda_1\mu_1 t-(\mu_1-\mu_2)tx+\mu_2(\mu_1(1-\alpha(t))-\mu_2)\}} \qquad (3.11)$$

where

$$\alpha(t) = \frac{x+\mu_1-\mu_2 - \sqrt{x^2-4\lambda_1\mu_1}}{2\mu_1} \tag{3.12}$$

From this Morse showed that

$$\sum_{n,m} nP(m,n) = \frac{\lambda_2}{\mu_2} + \frac{\lambda_2(\lambda_1\mu_2{}^2+\lambda_2\mu_1{}^2)}{(\mu_1-\lambda_1)(\mu_1\mu_2-\lambda_1\mu_2-\lambda_2\mu_1)\mu_2} \tag{3.13}$$

and

$$\sum_{n,m} mP(m,n) = \frac{\lambda_1}{(\mu_1-\lambda_1)} \left\{ \frac{\mu_2{}^2+\lambda_2\mu_1}{\mu_2{}^2} \right\} \tag{3.14}$$

The first published results for a preemptive priority discipline were by White and Christie [17]. An extension of the results in that paper was made by Stephan [18].

White and Christie found the generating function G(s,t) and thereby the first and second moments:

$$G(s,t) = \frac{\mu_2(1-t)(1-\lambda_1/\mu_1-\lambda_2/\mu_2)}{(t\mu_1\alpha(t)-\lambda_1 t-\lambda_2 t(1-t)-\mu_2(t-1))(1-s\alpha(t))} \tag{3.15}$$

where $\alpha(t)$ is given by (3.12).

The first moments of the number in the queue are:

$$\sum_{m,n} nP(m,n) = \frac{\lambda_2(\mu_2\lambda_1+\mu_1{}^2-\mu_1\lambda_1)}{(\mu_2\mu_1-\lambda_2\mu_1-\lambda_1\mu_2)(\mu_1-\lambda_1)} \tag{3.16}$$

and

$$\sum_{m,n} mP(m,n) = \lambda_1 | (\mu_1-\lambda_1) \tag{3.17}$$

This latter result is the same as that for priority items in isolation. This can be seen to be true if one considers the fact that under the pre-emptive service discipline the presence of calls of lower priority does not influence the stochastic law of the higher priority process. Therefore to find the stationary distribution of the queue size or the waiting time for the higher priority calls, one can consider that the lower priority calls are not present in the system.

The method employed by both Morse and, White and Christie made use of the fact that for Poisson distributed arrivals and an exponential holding time distribution the queueing process can be described by a continuous time parameter Markov process. If P is the stationary distribution of the queueing process and A is the infinitesimal matrix of the forward steady state equations then P is a solution of these equations. This can be represented by PA=0. If the system of equations PA=0 has a unique solution and a stationary distribution is assumed to exist then algebraic manipulations of equations PA=0 will yield a characterization of the stationary distribution or its generating function. Once the generating function has been derived the moments can be found very easily.

This method only applies however when we have exponentially distributed holding times because of the 'memorylessness' of this the exponential function as demonstrated in Chapter II, Section 3.

A stochastic process can only be said to be of the Markov type if a knowledge of the present value of the state of the stochastic system makes all information about its past history irrelevant to a prediction of its future behaviour. The process may become non-Markovian if part of the information contained in the description of the state of the system is suppressed, for a knowledge of the previous behaviour may enable some of the suppressed information about the present state to be recovered. The Markov property can be restored by reducing the continuous time parameter process to discrete time. A discrete Markov process is generated if the queueing process is observed only at the points in time which are the termination points of the service period (regeneration points), priority or nonpriority. This technique was originated by Palm and put on a rigorous basis by Kendall [19]. Miller [20] used this method to generalize Morse's results to general holding time distributions. It is interesting to note that Miller's results do not reduce to Morse's results for the exponential holding time showing that the stationary distribution for the imbedded Markov Chain and general time t are identical only for simpler queues.

Gaver [21] later introduced the concept of completion time to aid the analysis of priority assignments. This was used later by Ari-Itzhak and Naor [22] to study the multipurpose service stations problem.

Completion time is defined as the duration of the period that elapses between the instant at which the call first leaves the queue for service and the instant at which the call departs the system. In a non-preemptive priority system this is just the holding time of the call. In a preemptive priority system the completion time would be the sum of the holding time and any additional durations resulting from higher priority calls arriving in the system. The completion time of a call in the highest priority group (lowest numbered priority) is, of course, just the holding time because no call can interrupt it.

Takacs [23] considered a single server system with both preemptive and non-preemptive priority serve. His work is of interest in Chapters IV and V and will be referred to then.

CHAPTER IV

Priority Systems with Last-In First-Out Queue Discipline

1. Introduction

As indicated in Chapter III, there exists a substantial amount of literature on priority systems. Unfortunately non of this can be used when the system considered has other than FIFO for queue discipline. There are cases in the literature which consider other than FIFO but which, however, do not consider the effect of assigning priorities.

In this chapter we introduce and analyse a model that is descriptive of a single control unit which has LIFO queue discipline and considers the effect of assigning priorities. Our main purpose is to investigate the model with the aim of obtaining specific measures of its performance to compare these under different operational schemes and loading assumptions.

The basic model is presented in the next section and is analysed in section 4 under two different priority disciplines and the assumptions of a Poisson input process and a general holding time distribution which has a different mean for the demands in each priority group. In section 5, this restriction of different holding time distributions is eased. As will be seen the results are given in the form of conventional performance measures (e.g. waiting time distribution, mean and variance of the waiting time). The chapter concludes with examples and remarks concerning the model and conclusions pertinent to the study of the present model.

2. The Model

Calls with different priorities are arriving at a control unit in accordance with a Poisson process of intensity λ. The calls are served in order of priority and for each priority in the reverse order of arrival (LIFO).

Calls of lower numbered priority are served first. We shall consider two types of service:

Case (i): Non-preemptive type

The current service is never interrupted. This is true even if the arriving call has priority over the call in service at that time.

Case (ii): Preemptive-resume type.

The current service can be interrupted. If the arriving call has a lower priority number than the call being served, then the server interrupts the current service and immediately starts serving the call of higher priority. The interrupted service is resumed when there are no calls with a lower priority number in the queue. The interrupted service is considered to have priority over calls in the same priority class. An alternative queue discipline, which is not considered, would be to place the interrupted service back into the queue such that it competes for service with new arrivals of the same priority class.

3. Notation

The holding times of calls are assumed to be mutually independent random variables and independent of the inter-arrival times. Denote by $H_p(t)$ (p=1, 2, ..., K) the probability that the service time of a call with priority p, is $\leq$t.

Let
$$\alpha_p^{(n)} = \int_0^\infty t^n \, dH_p(t) \qquad (4.1)$$

and in particular,

$$\alpha_p = \alpha_p^{(1)}$$

P(p) is the probability that an arriving call has priority $\leq$p. The arrivals with priority $\leq$p follow a Poisson process of intensity σ_p where

$$\sigma_p = \lambda P(p) = \sum_{i=1}^{p} \lambda_i \qquad (4.2)$$

and in particular

$$\lambda = \sigma_k = \sum_{i=1}^{K} \lambda_i \qquad (4.3)$$

Under the condition that an arriving call has priority $\leq$p the probability that its service time is $\leq$t is given by:

$$A_p(t) = \frac{1}{P(p)} \int_0^p H_y(t)dP(y) \tag{4.4}$$

and

$$a_p^{(n)} = \int_0^\infty t^n \, dA_p(t) \tag{4.5}$$

$$\sigma_p A_p(t) = \sum_{i=1}^p \lambda_i \, H_i(t) \tag{4.6}$$

$$\sigma_p a_p^{(n)} = \sum_{i=1}^p \lambda_i \alpha_i^{(n)} \tag{4.7}$$

and in particular, $a_p = a_p^{(1)}$, $a^{(n)} = a_k^{(n)}$, and $a = a^{(1)}$. In order that a stationary distribution of the waiting times may exist, we have assumed $\sigma_p a_p < 1$.

$W_p(t)$ = the waiting time distribution for a priority p type call. The waiting time is the time that elapses between the instant at which the call appears at the server and that instant at which its service begins.

$Q_p(t)$ = the distribution of the time elapses before the first priority p type call is served, after the test call of priority p arrives in the system.

$M_p(t)$ = the distribution of the time to serve the calls with priority <p waiting when the test call arrives (including the remaining part of the current service time).

$F_p(t)$ = the distribution of the completion time for priority p type calls.

$G_p(t)$ = the distribution of the busy period in a single server system with Poisson input of intensity λ_p and with identically distributed mutually independent service times having the distribution $F_p(t)$.

$D_p(t)$ = the distribution of the busy period in a single server system with Poisson input of intensity σ_{p-1}, and with identically distributed, mutually independent service times having the distribution $A_{p-1}(t)$.

If U(t) is the distribution of a non-negative random variable, and the nth moment is given by:

$$U^{(n)} = \int_0^\infty t^n \, dU(t)$$

let

$$U^*(s) = \int_0^\infty e^{-st} \, dU(t)$$

4. Analysis

COMPLETION TIMES AND BUSY PERIODS

We consider first, the completion time distribution $F_p(t)$ for priority p type calls. Gaver[21] shows that:

$$F_p{}^*(s) = H_p{}^*(s+\sigma_{p-1} - \sigma_{p-1}D_p{}^*(s)) \tag{4.8}$$

The moments are obtained by differentiating at s=0

Therefore,

$$F_p{}^{(1)} = \alpha_p{}^{(1)} \;(1 + \sigma_{p-1}D_p{}^{(1)}) \tag{4.9}$$

$$F_p{}^{(2)} = \alpha_p{}^{(1)} \;\sigma_{p-1}D_p{}^{(2)} + \alpha_p{}^{(2)}(1 + \sigma_{p-1}D_p{}^{(1)})^2 \tag{4.10}$$

The busy period $D_p(t)$, of service times with the distribution $A_{p-1}(t)$ and arrival rate σ_{p-1} can be found from Takács [24]. He proved that $D_p{}^*(s)$ can be found from the expression:

$$D_p{}^*(s) = A_{p-1}{}^*(s + \sigma_{p-1} - \sigma_{p-1}\, D_p{}^*(s)) \tag{4.11}$$

Therefore $\quad D_p{}^{(1)} \;= a_{p-1}/(1-\sigma_{p-1}a_{p-1})$ (4.12)

$$D_p{}^{(2)} \;= a_{p-1}^{(2)}/(1-\sigma_{p-1}a_{p-1})^3 \tag{4.13}$$

Similarly $\quad G_p{}^*(s) = F_p{}^*(s+\lambda_p-\lambda_pG_p{}^*(s))$ (4.14)

and $\quad G_p{}^{(1)} \;= F_p{}^{(1)}/(1-\lambda_pF_p{}^{(1)}) = \alpha_p{}^{(1)}/(1-\sigma_pa_p)$ (4.15)

$$G_p{}^{(2)} \;= F_p{}^{(2)}/(1-\lambda_pF_p{}^{(1)})^3$$

$$= \alpha_p{}^{(1)} \;\sigma_{p-1}a_{p-1}^{(2)}/(1-\sigma_pa_p)^3 + \alpha_p{}^{(2)}(1-\sigma_{p-1}a_{p-1})/(1-\sigma_pa_p)^3 \tag{4.16}$$

THE DETERMINATION OF $W_p(t)$

The test call of priority p arrives in the system at t_o. In both cases (i) and (ii), its waiting time can be considered as consisting of two parts. The first is the time elapsed before the first priority p type service starts after t_0 and has distribution function $Q_p(t)$. Because of the reverse order of arrival queue discipline, this first priority p type call served after t_0 must have arrived in the system after the test call, that is in the first interval. If no calls of priority p arrive in this first interval then the first priority p type call served will of course be the test call. Assume that j, (j=0,1,...), calls of priority p arrive in this first interval then the second interval is identical with the total length of j independent busy periods in a single server queueing process, with Poisson input of intensity λ_p, and with identically distributed mutually independent service times having the same distribution as the priority p type completion time. The distribution function of the length of the busy period in this process has been denoted by $G_p(t)$.

Then

$$W_p(t) = \sum_{j=0}^{\infty} \int_0^t \left(\frac{e^{-\lambda_p y} (\lambda_p y)^j}{j!} dQ_p(y)\right) ** G_{p,j}(t)$$

Where ** denotes convolution and $G_{p,j}(t)$ is the jth iterated convolution of $G_p(t)$ with itself.

Therefore

$$W_p*(s) = \int_0^{\infty} e^{-(s+\lambda_p-\lambda_p G_p*(s))t} dQ_p(t)$$

$$= Q_p*(s+\lambda_p-\lambda_p G_p*(s)) \qquad (4.17)$$

It remains only to find $Q_p(t)$.

DERIVATION OF $Q_p(t)$

$Q_p(t)$ is the distribution of the time interval starting at t_0 and finishing when the first priority p type call is served. In this interval, then, only the calls of priority <p waiting at t_0 or arriving after t_0 are served. In both cases (i) and (ii), the interval can be considered as being composed of two parts. The first is the time needed to complete the service of all the calls of priority <p which are in the system at t_0 (including the remaining part of the current service). The second interval is the time to serve the calls of priority <p which arrive in the two intervals. The first interval has the distribution function $M_p(t)$ and, if j, (j=0,1,...) calls of priority <p arrive in this interval, the distribution of the second interval is that of the total length of j independent busy periods, each with distribution $D_p(t)$.

ie. $$Q_p(t) = \sum_{j=0}^{\infty} \int_0^t \left(\frac{e^{-\sigma_{p-1}y}(\sigma_{p-1}y)^j}{j!} dM_p(y)\right) ** D_{p,j}(t)$$

where $D_{p,j}(t)$ is the jth iterated convolution of $D_p(t)$ with itself.

Therefore,

$$Q_p*(s) = \int_0^{\infty} e^{-(s+\sigma_{p-1}-\sigma_{p-1}D_p*(s))t}\, dM_p(t)$$

$$= M_p*(s+\sigma_{p-1}-\sigma_{p-1}D_p*(s)) \tag{4.18}$$

The moments of $W_p(t)$ can now be obtained by using Fas di Bruno's formula[25] for the higher derivatives of a function of a function.

Thus from (4.17) and (4.18)

$$W_p^{(1)} = M_p^{(1)} \, (1 + \sigma_{p-1} D_p^{(1)})(1+\lambda_p G_p^{(1)})$$

$$= M_p^{(1)}/(1-\sigma_p a_p) \tag{4.19}$$

$$W_p^{(2)} = M_p^{(2)}\left[1+\sigma_{p-1}D_p^{(1)}\right]^2\left[1+\lambda_p G_p^{(1)}\right]^2 + M_p^{(1)} \, \sigma_{p-1}D_p^{(2)}\left[1+\lambda_p G_p^{(1)}\right]$$

$$+\lambda_p G_p^{(2)} M_p^{(1)}\left[1+\sigma_{p-1}D_p^{(1)}\right]$$

$$= M_p^{(2)}/(1-\sigma_p a_p)^2 + M_p^{(1)}\sigma_{p-1}a_{p-1}^{(2)}/(1-\sigma_p a_p)^3(1-\sigma_{p-1}a_{p-1})^2$$

$$+ \left[\lambda_p M_p^{(1)}/(1-\sigma_p a_p)^3 \, (1-\sigma_{p-1}a_{p-1})\right]\left[\alpha_p^{(1)}\sigma_{p-1}a_{p-1}^{(2)} + \alpha_p^{(2)}(1-\sigma_{p-1}a_{p-1})\right] \tag{4.20}$$

Thus we require $M_p^{(1)}$ and $M_p^{(2)}$ for both cases (i) and (ii)

DERIVATION OF $M_p(t)$

Case (i): The distribution of $M_p(t)$ may be found from Takács.[23] $M_p(t)$ is the distribution of the time to service all the calls with priority $< p$ present in the system at t_0. In Takács study, the distribution of the time needed to complete service of all calls of priority p or less present in the system at t_0 is derived. Although the queue disciplines in each priority level in Takacs study and the present study are different, the queue length distributions in the two cases are the same and therefore the waiting time to service calls with priority less than p present at time t_0 in this study can be obtained from Takacs results by replacing p by p-1.

Therefore when $\lambda a < 1$

$$M_p^*(s) = \frac{(1-\lambda a)s+\lambda(1-A_k^*(s))-\sigma_{p-1}(1-A_{p-1}^*(s))}{s-\sigma_{p-1} + \sigma_{p-1} A_{p-1}^*(s)} \tag{4.21}$$

and

$$M_p^{(1)} = \lambda a^{(2)}/2(1-\sigma_{p-1}a_{p-1}) \tag{4.22}$$

$$M_p^{(2)} = \lambda a^{(3)}/3(1-\sigma_{p-1}a_{p-1}) + \lambda a^{(2)}\sigma_{p-1}a_{p-1}^{(2)}/2(1-\sigma_{p-1}a_{p-1})^2 \tag{4.23}$$

when $\lambda a > 1$ but $q = \inf(p: \sigma_p a_p \geq 1)$

then

$$M_p^*(s) = \frac{(1-\sigma_{q-1}a_{q-1})(1-H_q^*(s))+A_q^*(s)(\lambda(1-A_k^*(s))-\sigma_{p-1}(1-A_{p-1}^*(s)))}{A_q^*(s)(s-\sigma_{p-1} + \sigma_{p-1}A_{p-1}^*(s))} \tag{4.24}$$

and

$$M_p^{(1)} = \frac{\sigma_{p-1}a_{q-1}^{(2)}\alpha_q + (1-\sigma_{q-1}a_{q-1})a_q^{(2)}}{2(1-\sigma_{p-1}a_{p-1})\alpha_q} \tag{4.25}$$

$$M_p^{(2)} = \frac{\sigma_{q-1}a_{q-1}^{(3)}\alpha_q + (1-\sigma_{q-1}a_{q-1})\alpha_q^{(2)}}{3(1-\sigma_{p-1}a_{p-1})\alpha_q}$$

$$+ \frac{\sigma_{q-1}a_{q-1}^{(2)}\ \sigma_{p-1}a_{p-1}^{(2)}\alpha_q + (1-\sigma_{q-1}a_{q-1})\sigma_{p-1}a_{p-1}^{(2)}\ \alpha_q^{(2)}}{2(1-\sigma_{p-1}a_{p-1})^2\alpha_q} \tag{4.26}$$

Case (ii): The presence of calls with priority >p does not influence the system. Therefore, to find $M_p(t)$, we can assume that calls of priority >p will be immediately removed from the system on their arrival. The formula for $M_p^*(s)$ can be found from (13) by replacing λ by σ_p.

i.e.

$$M_p^*(s) = \frac{(1-\sigma_p a_p)s + \sigma_p(1-A_p^*(s))-\sigma_{p-1}(1-A_{p-1}^*(s))}{s-\sigma_{p-1} +\sigma_{p-1}A_{p-1}^*(s)} \tag{4.27}$$

and $$M_p^{(1)} = \sigma_p a_p^{(2)}/2(1-\sigma_{p-1}a_{p-1}) \tag{4.28}$$

$$M_p^{(2)} = \sigma_p a_p^{(3)}/3(1-\sigma_{p-1}a_{p-1}) + (\sigma_p a_p^{(2)}\sigma_{p-1}a_{p-1}^{(3)})/2(1-\sigma_{p-1}a_{p-1})^2 \tag{4.29}$$

Therefore, <u>in case (i)</u>, if <u>$\lambda a < 1$</u> from (4.19) and (4.22)

$$W_p^{(1)} = \lambda a^{(2)}/2(1-\sigma_{p-1}a_{p-1})(1-\sigma_p a_p) \tag{4.30}$$

and from (4.20), (4.22) and (4.23)

$$W_p^{(2)} = \lambda a^{(3)}/3(1-\sigma_{p-1}a_{p-1})(1-\sigma_p a_p)^2 + \lambda a^{(2)}\sigma_{p-1}a_{p-1}^{(2)}/2(1-\sigma_p a_p)^2(1-\sigma_{1-p}a_{1-p})^2$$

$$+ \lambda a^{(2)}\sigma_p a_p^{(2)}/2(1-\sigma_p a_p)^3(1-\sigma_{p-1}a_{p-1}) \tag{4.31}$$

The distribution of waiting times when the queue discipline is first come first served has been investigated by a number of authors. In particular, Cobham[11] found the first moment ${}_kW_p^{(1)}$ and Kesten and J.Th. Runnenburg[15] found the second moment ${}_kW_p^{(2)}$.

$${}_kW_p^{(1)} = \lambda a^{(2)}/2(1-\sigma_{p-1}a_{p-1})(1-\sigma_p a_p) = W_p^{(1)}$$

$${}_kW_p^{(2)} = \lambda a^{(3)}/3(1-\sigma_{p-1}a_{p-1})^2(1-\sigma_p a_p) + \lambda a^{(2)}\sigma_{p-1}a_{p-1}^{(2)}/2(1-\sigma_p a_p)\ (1-\sigma_{p-1}a_{p-1})^3$$

$$+ \lambda a^{(2)}\sigma_p a_p^{(2)}/2(1-\sigma_p a_p)^2(1-\sigma_{p-1}a_{p-1})^2$$

$$= \frac{(1-\sigma_p a_p)}{(1-\sigma_{p-1}a_{p-1}} \quad W_p^{(2)}$$

$$\sigma_p a_p > \sigma_{p-1}a_{p-1}$$

$$\therefore \quad (1-\sigma_p a_p)/(1-\sigma_{p-1}a_{p-1}) < 1$$

so that the variance of the waiting time distribution is always greater under the discipline'last come first served' although the expected waiting times are the same. This is an extension of the results of Wishart[27] for the same systems without priorities. In case (i) when $\lambda a \geq 1$ and $\sigma_p a_p < 1$ then by (4.19) and (4.25)

$$W_p^{(1)} = \frac{\alpha_q \sigma_{q-1} a_{q-1}^{(2)} + (1-\sigma_{q-1}a_{q-1})\alpha_q^{(2)}}{2(1-\sigma_p a_p)(1-\sigma_{p-1}a_{p-1})} \tag{4.32}$$

where q is the smallest index such that $\sigma_q a_q \geq 1$. $W_p^{(2)}$ can be found in a similar fashion from (4.20) and (4.26)

<u>In case (ii)</u>, if $\sigma_p a_p < 1$, then from (4.19) and (4.28)

$$W_p^{(1)} = \sigma_p a_p^{(2)}/2(1-\sigma_{p-1}a_{p-1})(1-\sigma_p a_p) \tag{4.33}$$

and from (4.20) and (4.29)

$$W_p^{(2)} = \sigma_p a_p^{(3)}/3(1-\sigma_{p-1}a_{p-1})(1-\sigma_p a_p)^2 + \sigma_p a_p^{(2)}\sigma_{p-1}a_{p-1}^{(2)}/2(1-\sigma_p a_p)^2(1-\sigma_{p-1}a_{p-1})^2$$
$$+ \left(\sigma_p a_p^{(2)}\right)^2/2(1-\sigma_p a_p)^3(1-\sigma_{p-1}a_{p-1}) \tag{4.34}$$

This case with the first come first served queue discipline was investigated by Miller[20]. He found the first two moments ${}_M W_p^{(1)}$ and ${}_M W_p^{(2)}$:

$$_M W_p^{(1)} = \sigma_p a_p^{(2)}/2(1-\sigma_{p-1}a_{p-1})(1-\sigma_p a_p) = W_p^{(1)}$$

$$_M W_p^{(2)} = \frac{(1-\sigma_p a_p)}{(1-\sigma_{p-1}a_{p-1})} \; W_p^{(2)}$$

so that once again it can be seen that although the variance is greater for the last come first served discipline the average waiting times are the same.

5. THE CASE WHEN THE PRIORITIES HAVE THE SAME SERVICE TIME DISTRIBUTIONS

In this case

$$F_p(t) = D_p(t) \tag{4.35}$$

and,

$$A_i(t) = A(t) \quad \text{for } i = 1, \ldots, k.$$

From (4.11), (4.14) and (4.35)

$$G_p^*(s) = D_p^*(s+\lambda_p-\lambda_p G_p^*(s)) = A^*(s+\sigma_p-\sigma_p G_p^*(s)) \tag{4.36}$$

Therefore from (4.17) and (4.18)

$$W_p^*(s) = M_p^*(s+\sigma_p-\sigma_p G_p^*(s)) \tag{4.37}$$

In case (i) when $\lambda a < 1$, using (4.21), (4.35), (4.36) and (4.37)

$$W_p^*(s) = 1-\lambda a + \frac{\lambda(1-\sigma_{p-1}a)(1-G_p^*(s))}{s+\lambda_p-\lambda_p G_p^*(s)}$$

Therefore using Lagrange's expansion[26]:

$$W_p^*(s) = 1-\lambda a + \lambda(1-\sigma_{p-1}a)\left[\frac{1}{(s+\lambda_p)} - \sum_{n=1}^{\infty} \frac{(-1)^n \lambda_p^n}{n!} \frac{d^{n-1}}{ds^{n-1}} \left(\frac{A(s+\lambda_p)^n}{(s+\lambda_p)^2}\right)\right]$$

and upon inversion

$$W_p(x) = 1-\lambda a + \lambda(1-\sigma_{p-1}a) \sum_{j=1}^{\infty} e^{-\lambda_p x} (\lambda_p x)^{j-1}/j! \int_0^x (1-{}_jH(u))du \tag{4.38}$$

where

${}_jH(u)$ is the jth iterated convolution of H(u) with itself.

Similarly in case (ii),

When $\sigma_p a_p < 1$, using (4.27), (4.35), (4.36), (4.37) and inverting

$$W_p(x) = 1-\sigma_p a_p + \sigma_p(1-\sigma_{p-1}a) \sum_{j=1}^{\infty} e^{-\lambda_p x} (\lambda_p x)^{j-1}/j! \int_0^x (1-{}_jH(u))du$$

6. <u>EXAMPLES</u>

Two cases will be considered:

1) Constant Holding Time Distribution

2) Exponential Holding Time Distribution

These two distributions have been chosen as examples because, for a given occupancy, the pair of curves define a band containing all LIFO delay distributions for that occupancy where the holding time distribution is of the gamma type in which the coefficient of variation is not greater than unity. (In particular, the χ^2 distributions with two or more degrees of freedom are of this type).

Example 1

Consider the case where the number of priority groups is 2 and H(t) is given by:

$$H(t) = \begin{cases} 0 & t \leq h \\ 1 & t > h \end{cases}$$

$$\therefore \quad \int_0^x (1-{}_jH(t))dt = \min({}_jh,x)$$

$\therefore$ in case (i), with a non-preemptive priority discipline and, assuming that $\lambda h < 1$, we have from (4.38):

$$W_1(x) = 1-\lambda h + \lambda \sum_{j=1}^{\infty} \frac{e^{-\lambda_1 x}(\lambda_1 x)^{j-1}}{j!} \min(jh,x)$$

and

$$W_2(x) = 1-\lambda h + \lambda(1-\lambda_1 h) \sum_{j=1}^{\infty} \frac{e^{-\lambda_2 x}(\lambda_2 x)^{j-1}}{j!} \min(jh,x)$$

In case (ii), with a preemptive priority discipline in a similar manner from (4.39):

$$W_1(x) = 1-\lambda_1 h + \lambda_1 \sum_{j=1}^{\infty} \frac{e^{-\lambda_1 x}(\lambda_1 x)^{j-1}}{j!} \min(jh,x)$$

and

$$W_2(x) = 1-\lambda h + \lambda(1-\lambda_1 h) \sum_{j=1}^{\infty} \frac{e^{-\lambda_2 x}(\lambda_2 x)^{j-1}}{j!} \min(jh,x)$$

These formulae have been computed for various values of ρ, ρ_1 and ρ_2. The results are shown in Figs. 4.1 - 4.7. The results are plotted in terms of $(1-W_p(x/h))$ against x/h.

Example 2

Consider the case when the number of priority classes is two and H(t) is given by:

$$H(t) = \begin{cases} 0 & t < 0 \\ 1-e^{-\mu t} & t \geq 0 \end{cases}$$

$$\therefore \quad {}_jH(t) = 1- \sum_{v=0}^{j-1} \frac{e^{-\mu t}(\mu t)^v}{v!}$$

In the non-preemptive priority case then,

$$W_1(x) = 1-\lambda/\mu + \lambda/\mu \sum_{j=1}^{\infty} \sum_{v=0}^{j-1} \frac{e^{-\lambda_1 x}(\lambda_1 x)^{j-1}}{j!} \frac{\Gamma x\mu(v+1)}{v!}$$

and

$$W_2(x) = 1-\lambda/\mu + \frac{\lambda}{\mu} (1-\lambda_1/\mu) \sum_{j=1}^{\infty} \sum_{v=0}^{j-1} \frac{e^{-\lambda_2 x}(\lambda_2 x)^{j-1}}{} \frac{\Gamma x\mu(v+1)}{v!}$$

Where $\Gamma_{x\mu}(v+1)$ is the incomplete gamma function and is defined as

$$\Gamma_y(z) = \int_0^j t^{z-1} e^{-t} dt \qquad R(z) > 0$$

In the preemptive priority case,

$$W_1(x) = 1-\lambda_1/\mu + \lambda_1/\mu \sum_{j=1}^{\infty} \sum_{v=0}^{j-1} \frac{e^{-\lambda_1 x}(\lambda_1 x)^{j-1}}{j!} \frac{\Gamma x\mu(v+1)}{v!}$$

and

$$W_2(x) = 1-\lambda/\mu + \lambda(1-\lambda_1/\mu) \sum_{j=1}^{\infty} \sum_{v=0}^{j-1} \frac{e^{-\lambda_2 x}(\lambda_2 x)^{j-1}}{j!} \frac{\Gamma x\mu(v+1)}{v!}$$

These formulae have been computed for various values of ρ, ρ_1 and ρ_2 and are recorded in Fig. 4.8 - Fig. 4.14.

7. DISCUSSION

Figs. 4.1 to 4.14 all demonstrate that if the prime requirement is to reduce the probability of delay for one particular class of call, then this class should be given priority. This course of action will, however, increase the probability of delay of the remaining class (or classes). The introduction of a preemptive priority discipline will reduce the probability of delay of the preferred class even further and with no corresponding increase in delay for the remaining class. There is the penalty, of course, that a storage facility must now be provided for the interrupted calls.

It may not be important that the delay of one particular class be reduced. The criterion may simply be to cut down the total number of waiting calls of either type. In other words, the delay in queue of all calls may be the important criterion.

Let us first consider case (i) and two classes of calls, each with exponentially distributed holding times with averages $1/\mu_1$ and $1/\mu_2$ respectively. We can find the average delay W for all calls from the equation:

$$W = \frac{\lambda_1 W_1^{(1)} + \lambda_2 W_2^{(1)}}{\lambda}$$

Then from 4.30

$$W_1^{(1)} = \frac{(\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2)}{(1-\lambda_1/\mu_1)}$$

and $$W_2^{(1)} = \frac{(\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2)}{(1-\lambda_1/\mu_1)(1-\lambda_1/\mu_1-\lambda_2/\mu_2)}$$

$$\therefore \quad W = \frac{(\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2)(\lambda_2 + \lambda_1(1-\lambda_1/\mu_1-\lambda_2/\mu_2))}{\lambda(1-\lambda_1/\mu_1)(1-\lambda_1/\mu_1-\lambda_2/\mu_2)}$$

From Morse[16] the equivalent value of delay for the case when neither class is assigned priority is given by

$$W_M = \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{1-\lambda_1/\mu_1-\lambda_2/\mu_2}$$

If the ratio W/W_M is larger than 1, the delay is longer if priorities are imposed than if no priorities are imposed. If the ratio is less than 1 imposing priorities shortens the mean delay. As there is a linear relationship between the mean queue size and the mean delay then we can also say that if the ratio is larger or smaller than 1, there will be more or less calls waiting than if no priorities had been imposed.

The ratio $\frac{W}{W_M} < 1$

if

$$1-\lambda_1/\lambda(\lambda_1/\mu_1 + \lambda_2/\mu_2) < 1-\lambda_1/\mu_1$$

ie. $$\lambda\mu_2 < \lambda_1\mu_2 + \lambda_2\mu_1$$

ie. $$\frac{1}{\mu_1} < \frac{1}{\mu_2}$$

ie. When the non-priority units take longer to be served imposing priorites will lengthen the overall delay and queue length.

Consequently when it is required to reduce the total number waiting or to shorten the overall mean delay in queue, then if there is any way of distinguishing, on the average, between calls according to the prospective lengths of their holding times, it helps to give priority to that class of calls which tends to have the shorter mean holding time even if there is a wide spread in individual holding times. This result is independent of the ratio λ_1/λ_2. The degree of advantage does, however, depend upon the value of the ratio λ_1/λ_2. The degree of reduction in mean delay obtained by imposing priorities depends upon the numerical magnitude of W/W_M whereas the fact that there is a reduction depends upon whether μ_1 is greater than μ_2 or not.

In case (ii), from 4.33

$$W_1^{(1)} = \frac{\lambda_1\mu_1^2}{(1-\lambda_1/\mu_1)}$$

and $$W_2^{(1)} = \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{(1-\lambda_1/\mu_1)(1-\lambda_1/\mu_1-\lambda_2/\mu_2)}$$

$$\therefore \quad W = \frac{\{\lambda_1^2/\mu_1^2(1-\lambda_1/\mu_1-\lambda_2/\mu_2)+\lambda_2(\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2)\}}{\lambda(1-\lambda_1/\mu_1)(1-\lambda_1/\mu_1-\lambda_2/\mu_2)}$$

The ratio $W/W_M < 1$

if $$\lambda_1^2/\mu_1^2(1-\lambda_1/\mu_1-\lambda_2/\mu_2)+\lambda_2(\lambda_1/\mu_1^2+\lambda_2/\mu_2^2)<\lambda(1-\lambda_1/\mu_1)(\lambda_1/\mu_1^2+\lambda_2/\mu_2^2)$$

ie. $$\mu_1^2(\mu_1-\lambda) + \lambda_1\mu_2(\mu_1-\mu_2) > 0$$

Now in order that the system remains unsaturated

$$1-\lambda_1/\mu_1-\lambda_2/\mu_2 > 0$$

and if we now consider $\mu_1>\mu_2$

$$1-\lambda_1/\mu_1 - \lambda_2/\mu_1 > 0$$

ie. $\mu_1>\lambda$

$$\therefore \quad \mu_1^2(\mu_1-\lambda)+\lambda_1\mu_2(\mu_1-\mu_2)>0$$

ie. if $\mu_1>\mu_2$, $W<W_M$ as expected.

Thus a decrease in the overall delay can be achieved by assigning priority to the class of calls with the smaller average holding time.

Fig. 4.15 shows a comparison of the two holding time distributions considered. In order to show the association between the various queueing disciplines Fig. 4.16 and 4.17 have been included.

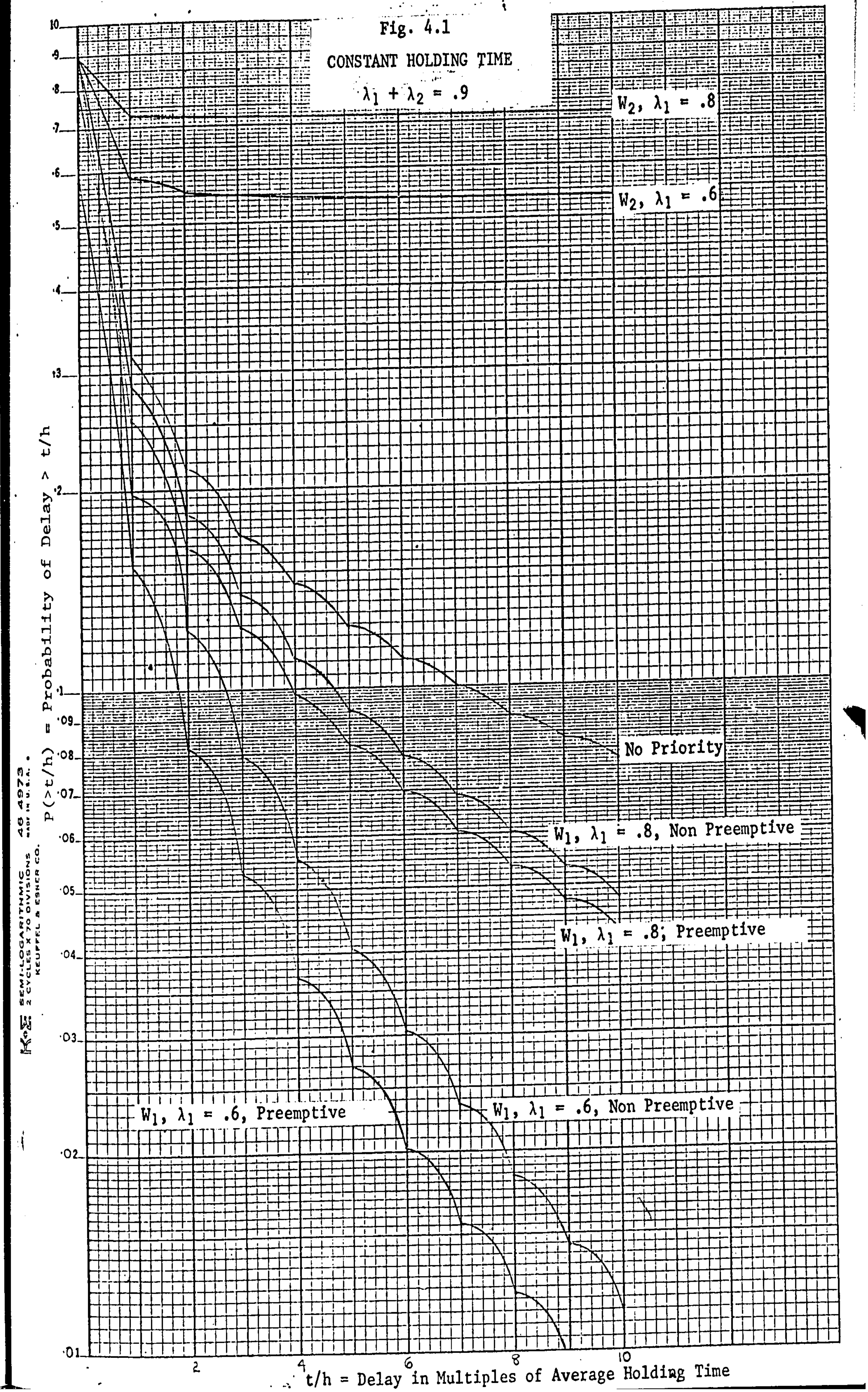


Fig. 4.1
CONSTANT HOLDING TIME
$\lambda_1 + \lambda_2 = .9$
W_2, λ_1 = .8
W_2, λ_1 = .6
No Priority
W_1, λ_1 = .8, Non Preemptive
W_1, λ_1 = .8, Preemptive
W_1, λ_1 = .6, Preemptive
W_1, λ_1 = .6, Non Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
SEMI-LOGARITHMIC 46 4973
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

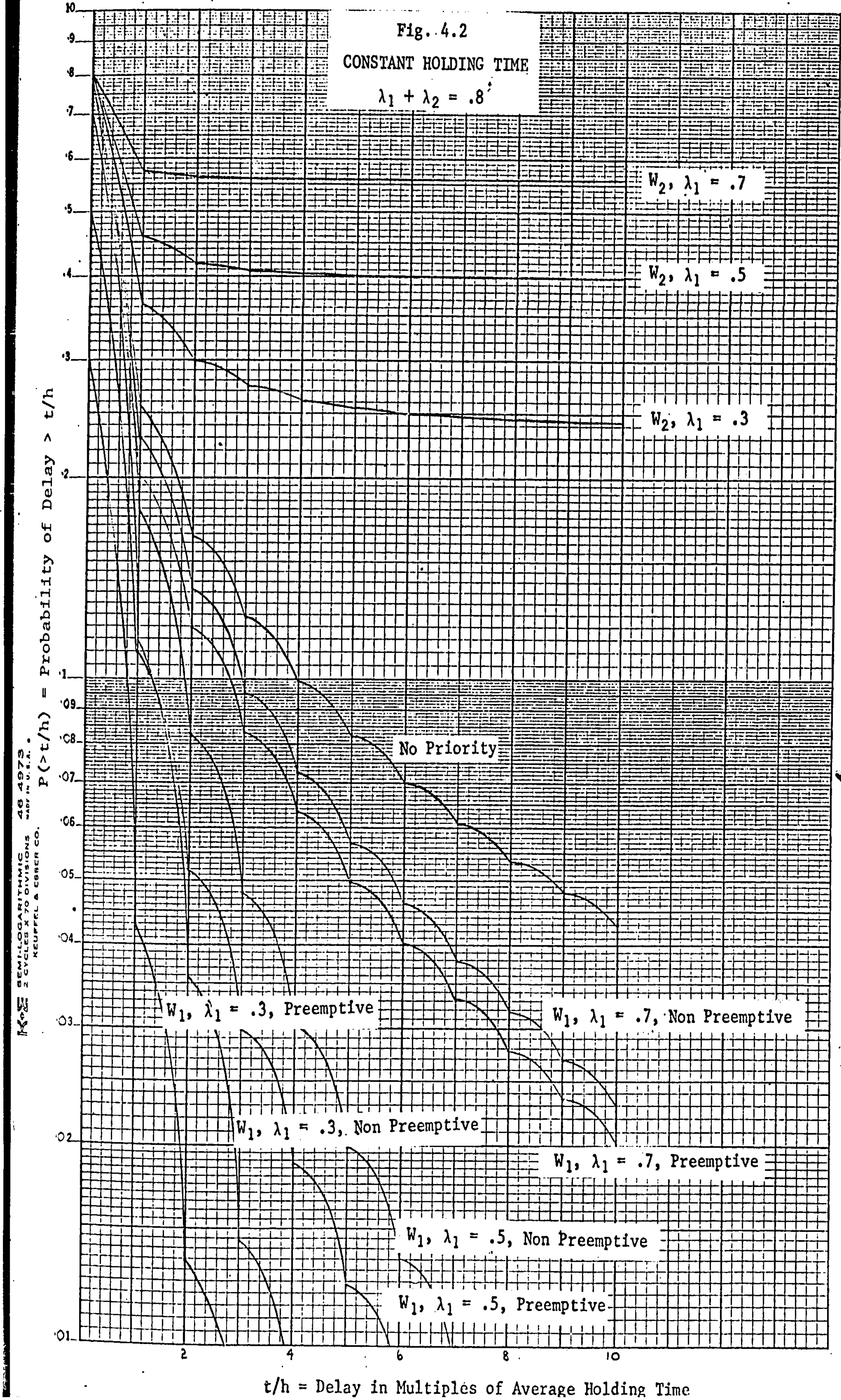


Fig. 4.2
CONSTANT HOLDING TIME
$\lambda_1 + \lambda_2 = .8$
$W_2, \lambda_1 = .7$
$W_2, \lambda_1 = .5$
$W_2, \lambda_1 = .3$
No Priority
$W_1, \lambda_1 = .3$, Preemptive
$W_1, \lambda_1 = .7$, Non Preemptive
$W_1, \lambda_1 = .3$, Non Preemptive
$W_1, \lambda_1 = .7$, Preemptive
$W_1, \lambda_1 = .5$, Non Preemptive
$W_1, \lambda_1 = .5$, Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
K+E SEMI-LOGARITHMIC 46 4973
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

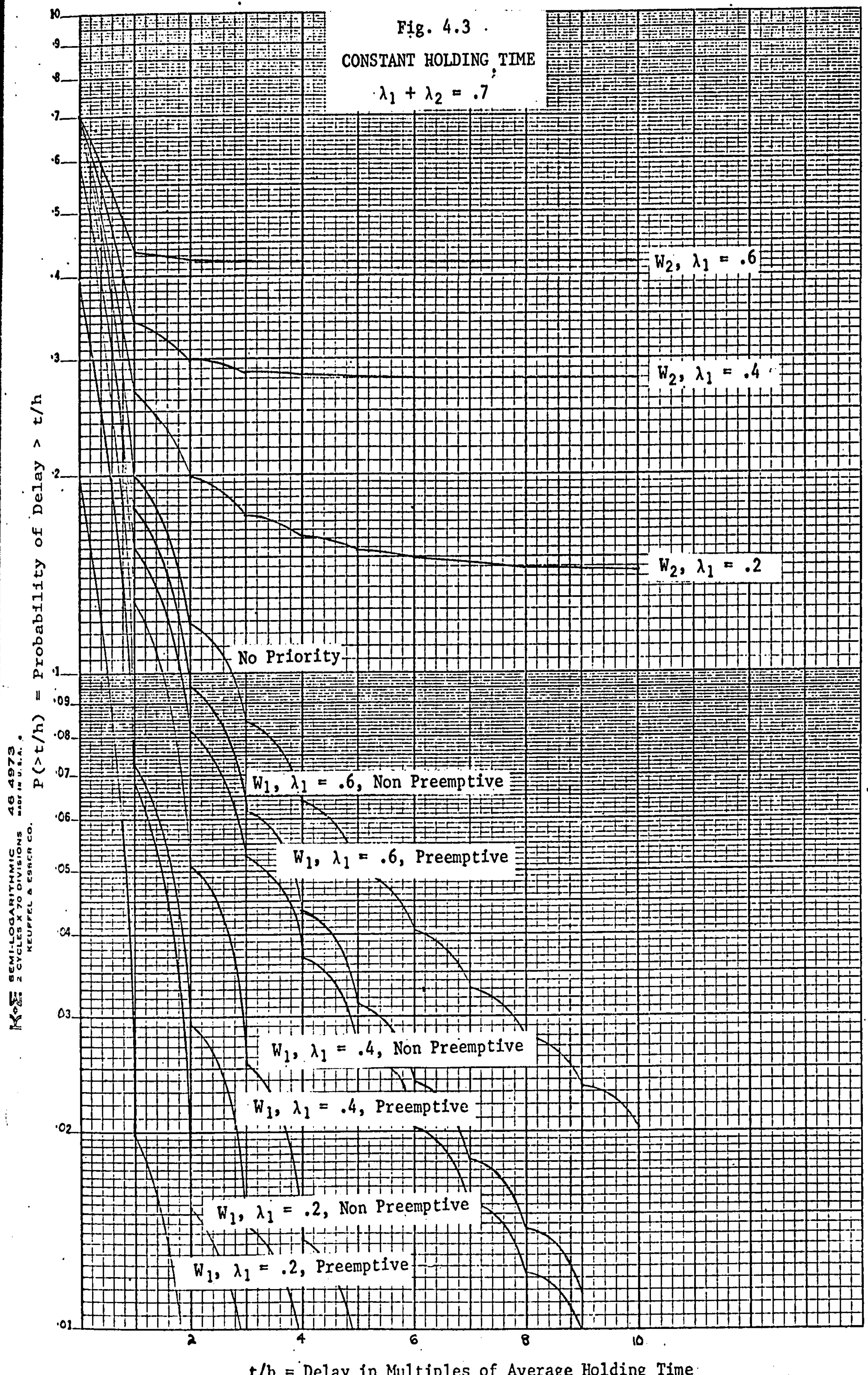


Fig. 4.3
CONSTANT HOLDING TIME
$\lambda_1 + \lambda_2 = .7$
W_2, λ_1 = .6
W_2, λ_1 = .4
W_2, λ_1 = .2
No Priority
W_1, λ_1 = .6, Non Preemptive
W_1, λ_1 = .6, Preemptive
W_1, λ_1 = .4, Non Preemptive
W_1, λ_1 = .4, Preemptive
W_1, λ_1 = .2, Non Preemptive
W_1, λ_1 = .2, Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
K&E SEMI-LOGARITHMIC 46 4973
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.
.01
.02
.03
.04
.05
.06
.07
.08
.09
.1
.2
.3
.4
.5
.6
.7
.8
.9
10
2
4
6
8
10

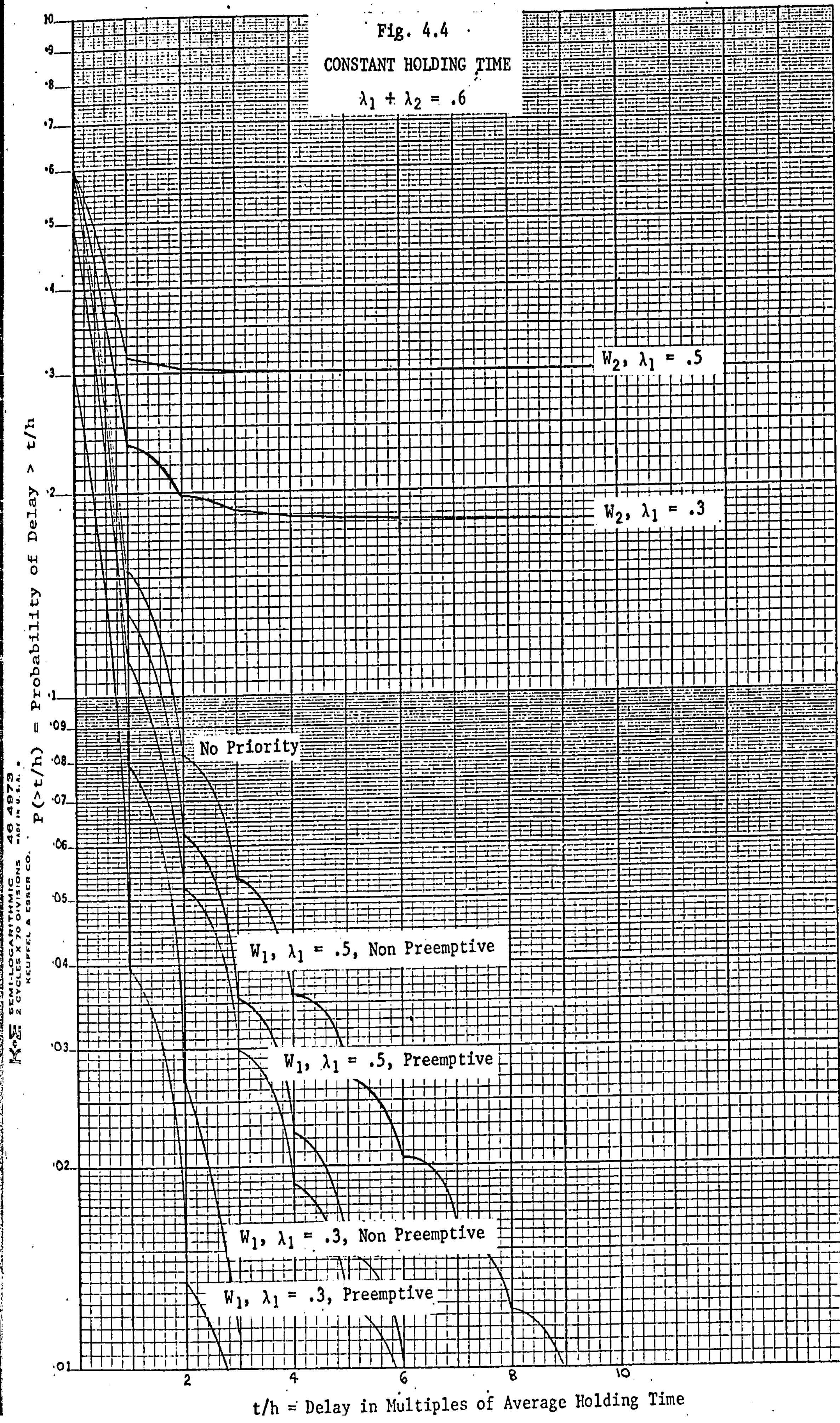


Fig. 4.4
CONSTANT HOLDING TIME
$\lambda_1 + \lambda_2 = .6$
W_2, λ_1 = .5
W_2, λ_1 = .3
No Priority
W_1, λ_1 = .5, Non Preemptive
W_1, λ_1 = .5, Preemptive
W_1, λ_1 = .3, Non Preemptive
W_1, λ_1 = .3, Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
.09
.08
.07
.06
.05
.04
.03
.02
.01
2
4
6
8
10
K+E SEMI-LOGARITHMIC 46 4973
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

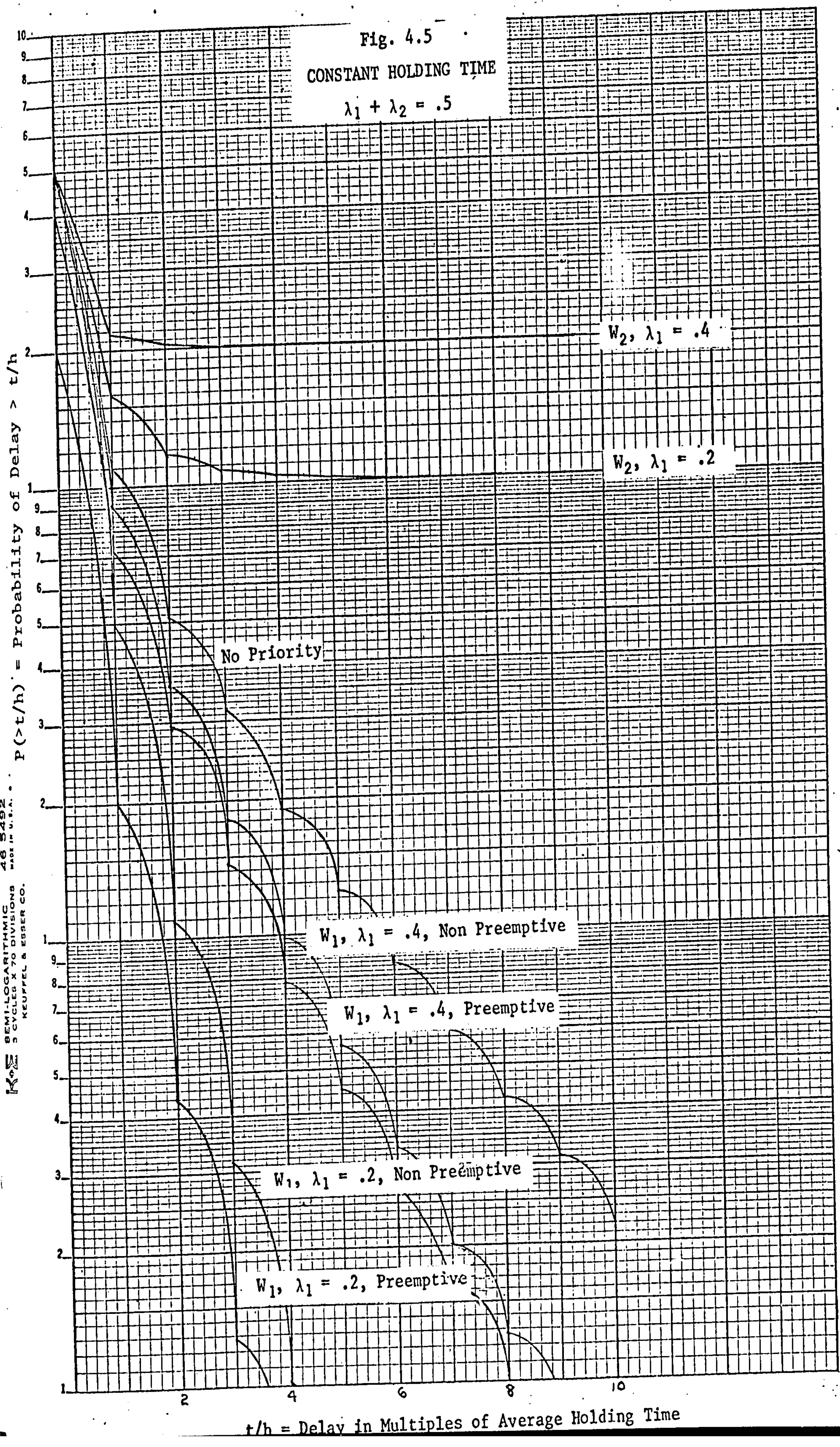


Fig. 4.5
CONSTANT HOLDING TIME
$\lambda_1 + \lambda_2 = .5$
W_2, λ_1 = .4
W_2, λ_1 = .2
No Priority
W_1, λ_1 = .4, Non Preemptive
W_1, λ_1 = .4, Preemptive
W_1, λ_1 = .2, Non Preemptive
W_1, λ_1 = .2, Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
SEMI-LOGARITHMIC 46 5492
3 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

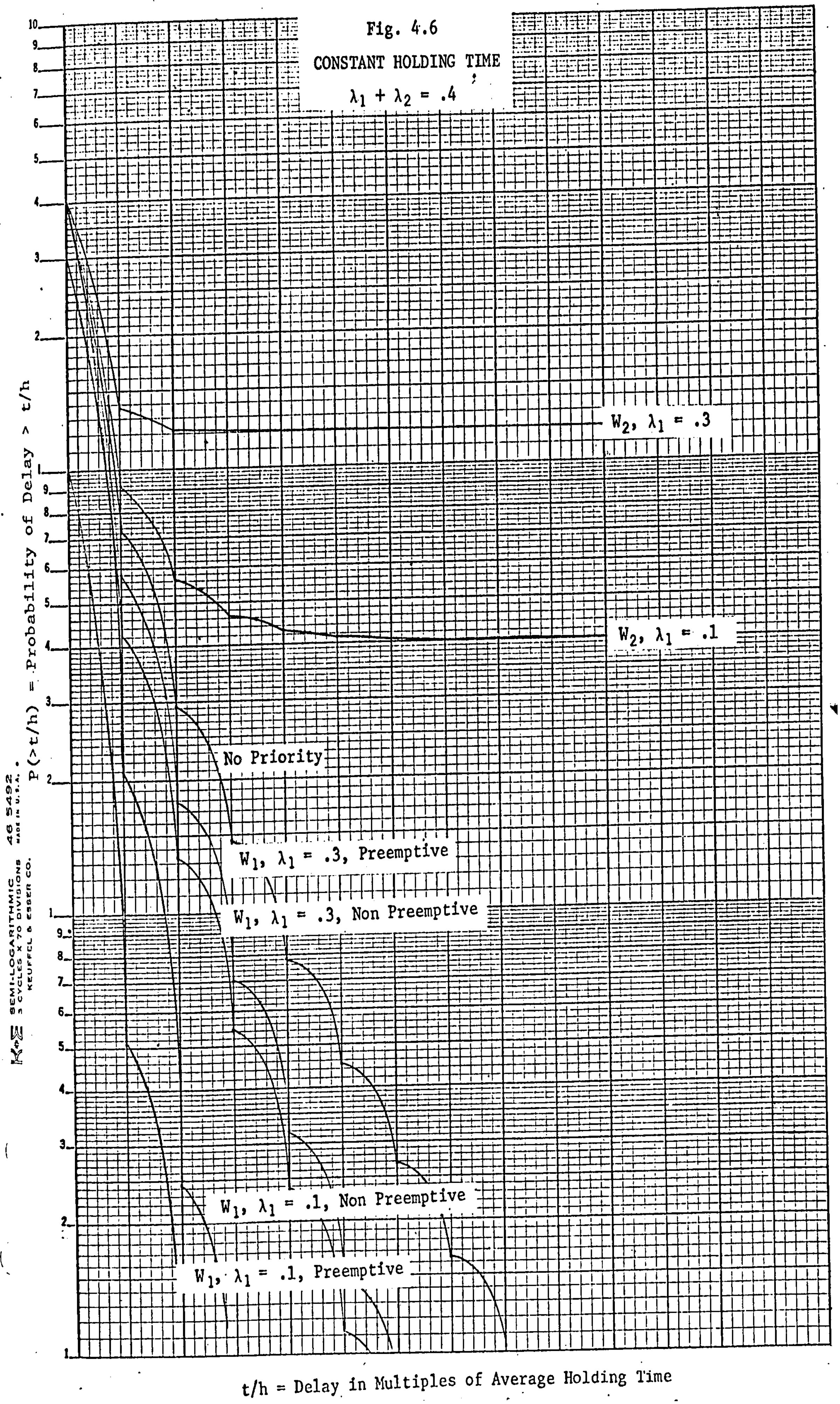
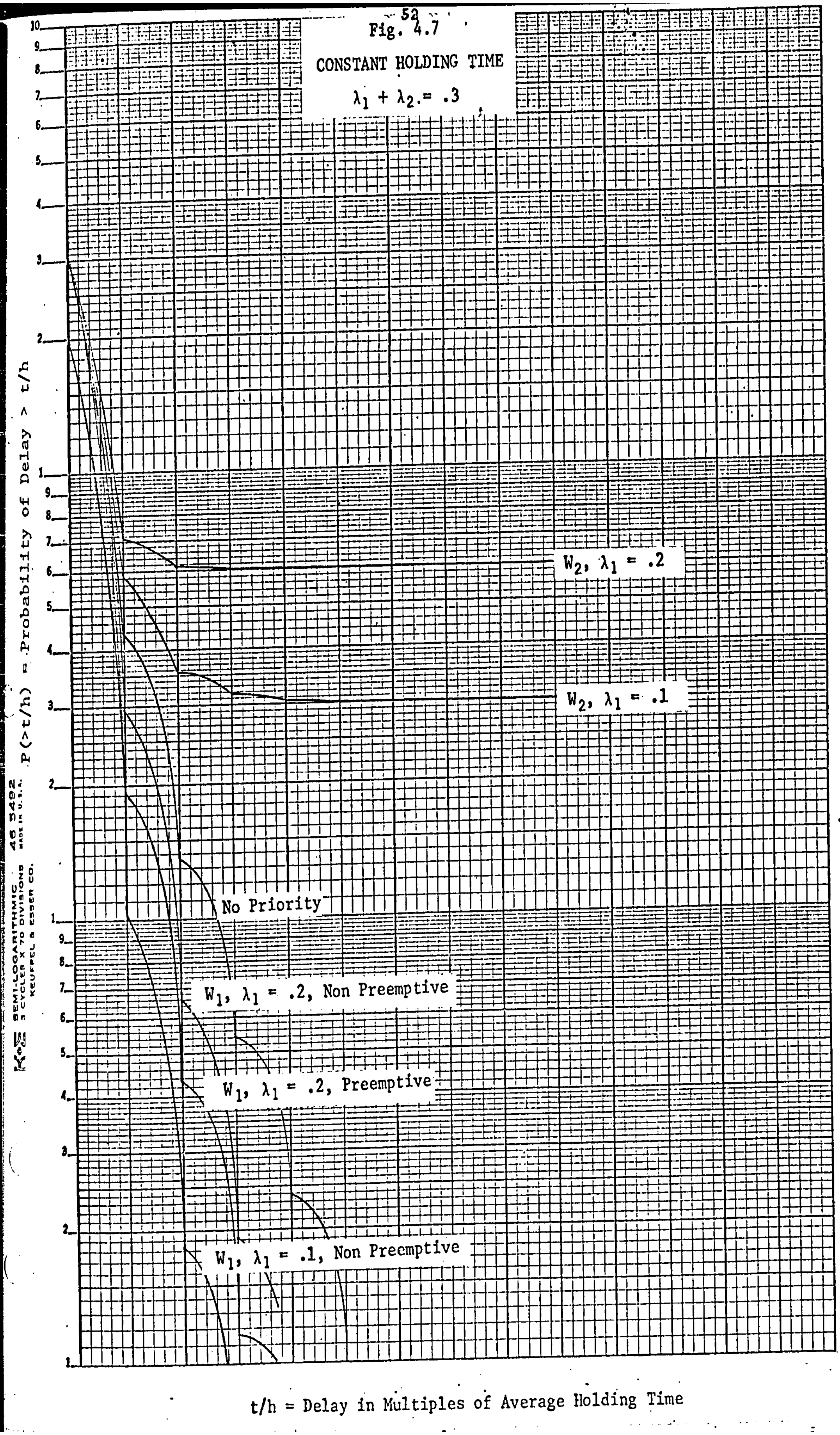
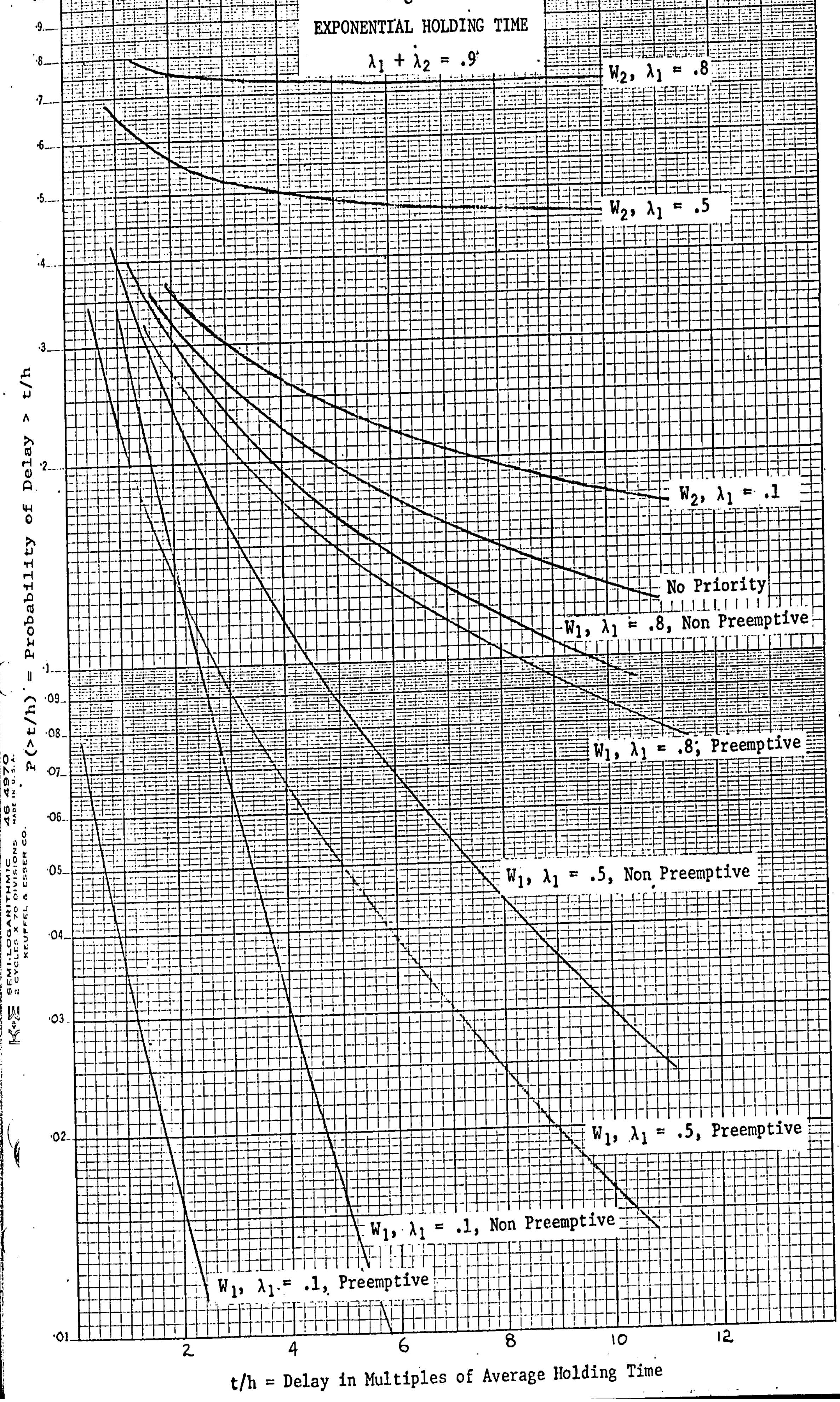


Fig. 4.6
CONSTANT HOLDING TIME
$\lambda_1 + \lambda_2 = .4$
W_2, λ_1 = .3
W_2, λ_1 = .1
No Priority
W_1, λ_1 = .3, Preemptive
W_1, λ_1 = .3, Non Preemptive
W_1, λ_1 = .1, Non Preemptive
W_1, λ_1 = .1, Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
K+E SEMI-LOGARITHMIC 46 5492
3 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.



t/h = Delay in Multiples of Average Holding Time

Fig. 4.8


EXPONENTIAL HOLDING TIME
$\lambda_1 + \lambda_2 = .9$
W_2, λ_1 = .8
W_2, λ_1 = .5
W_2, λ_1 = .1
No Priority
W_1, λ_1 = .8, Non Preemptive
W_1, λ_1 = .8, Preemptive
W_1, λ_1 = .5, Non Preemptive
W_1, λ_1 = .5, Preemptive
W_1, λ_1 = .1, Non Preemptive
W_1, λ_1 = .1, Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
10.
.9
.8
.7
.6
.5
.4
.3
.2
.1
.09
.08
.07
.06
.05
.04
.03
.02
.01
2
4
6
8
10
12
SEMI-LOGARITHMIC 46 4970
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

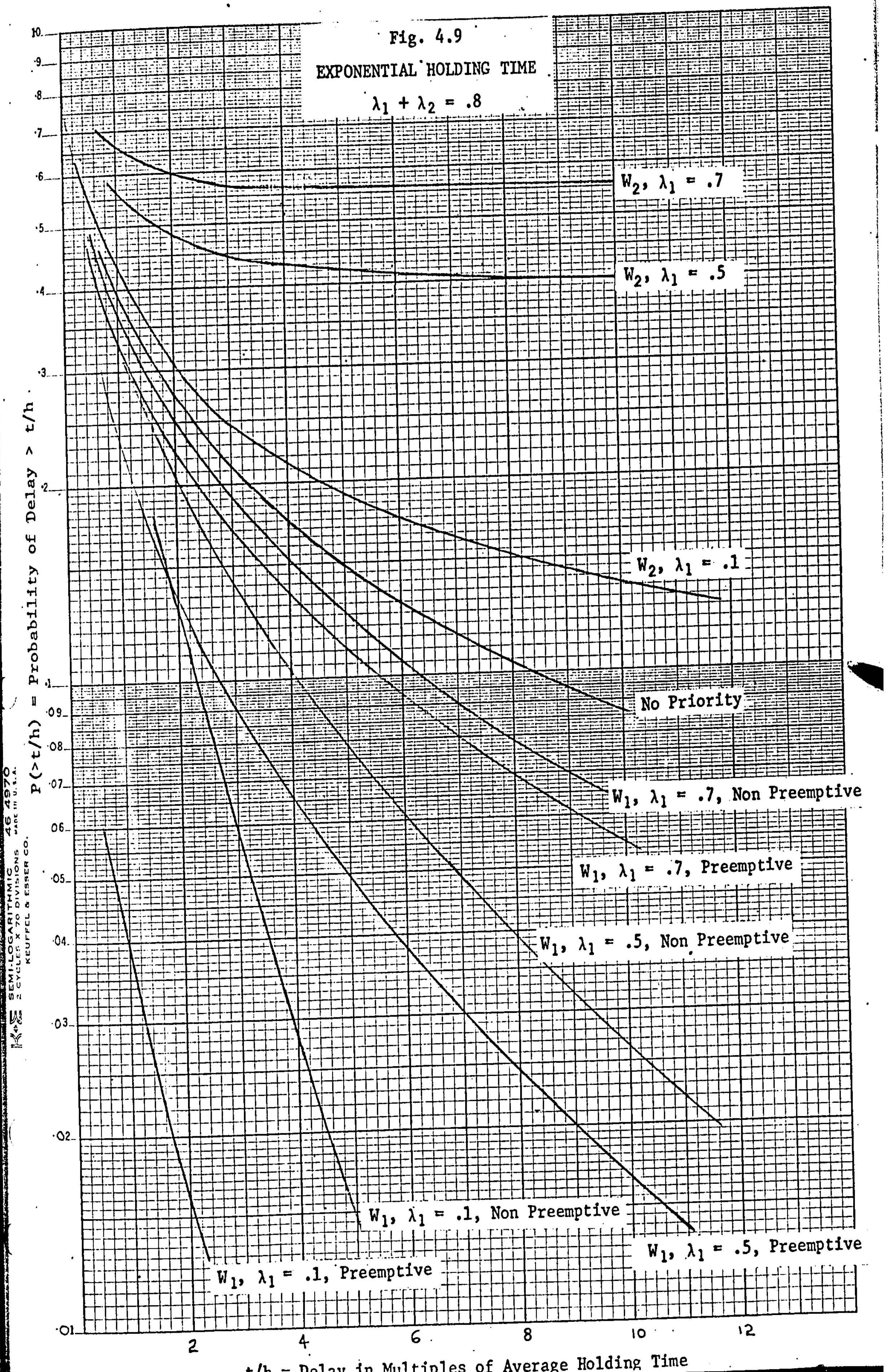


Fig. 4.9
EXPONENTIAL HOLDING TIME
$\lambda_1 + \lambda_2 = .8$
W_2, λ_1 = .7
W_2, λ_1 = .5
W_2, λ_1 = .1
No Priority
W_1, λ_1 = .7, Non Preemptive
W_1, λ_1 = .7, Preemptive
W_1, λ_1 = .5, Non Preemptive
W_1, λ_1 = .1, Non Preemptive
W_1, λ_1 = .5, Preemptive
W_1, λ_1 = .1, Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
SEMI-LOGARITHMIC 46 4970
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

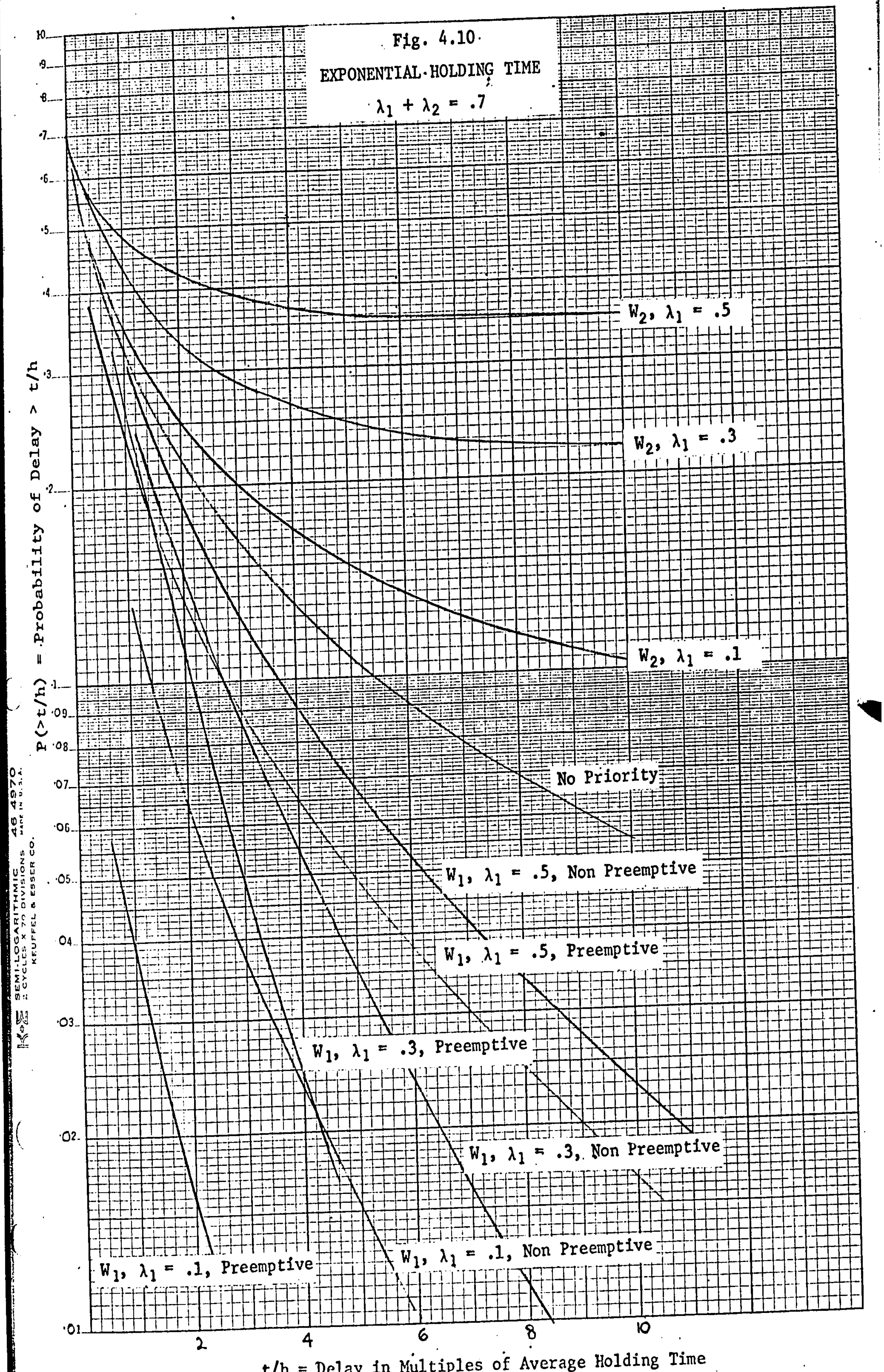


Fig. 4.10
EXPONENTIAL HOLDING TIME
$\lambda_1 + \lambda_2 = .7$
W_2, λ_1 = .5
W_2, λ_1 = .3
W_2, λ_1 = .1
No Priority
W_1, λ_1 = .5, Non Preemptive
W_1, λ_1 = .5, Preemptive
W_1, λ_1 = .3, Preemptive
W_1, λ_1 = .3, Non Preemptive
W_1, λ_1 = .1, Preemptive
W_1, λ_1 = .1, Non Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
1.
.9
.8
.7
.6
.5
.4
.3
.2
.1
.09
.08
.07
.06
.05
.04
.03
.02
.01
2
4
6
8
10
SEMI-LOGARITHMIC 46 4970
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

Fig. 4.11

EXPONENTIAL HOLDING TIME

$\lambda_1 + \lambda_2 = .6$

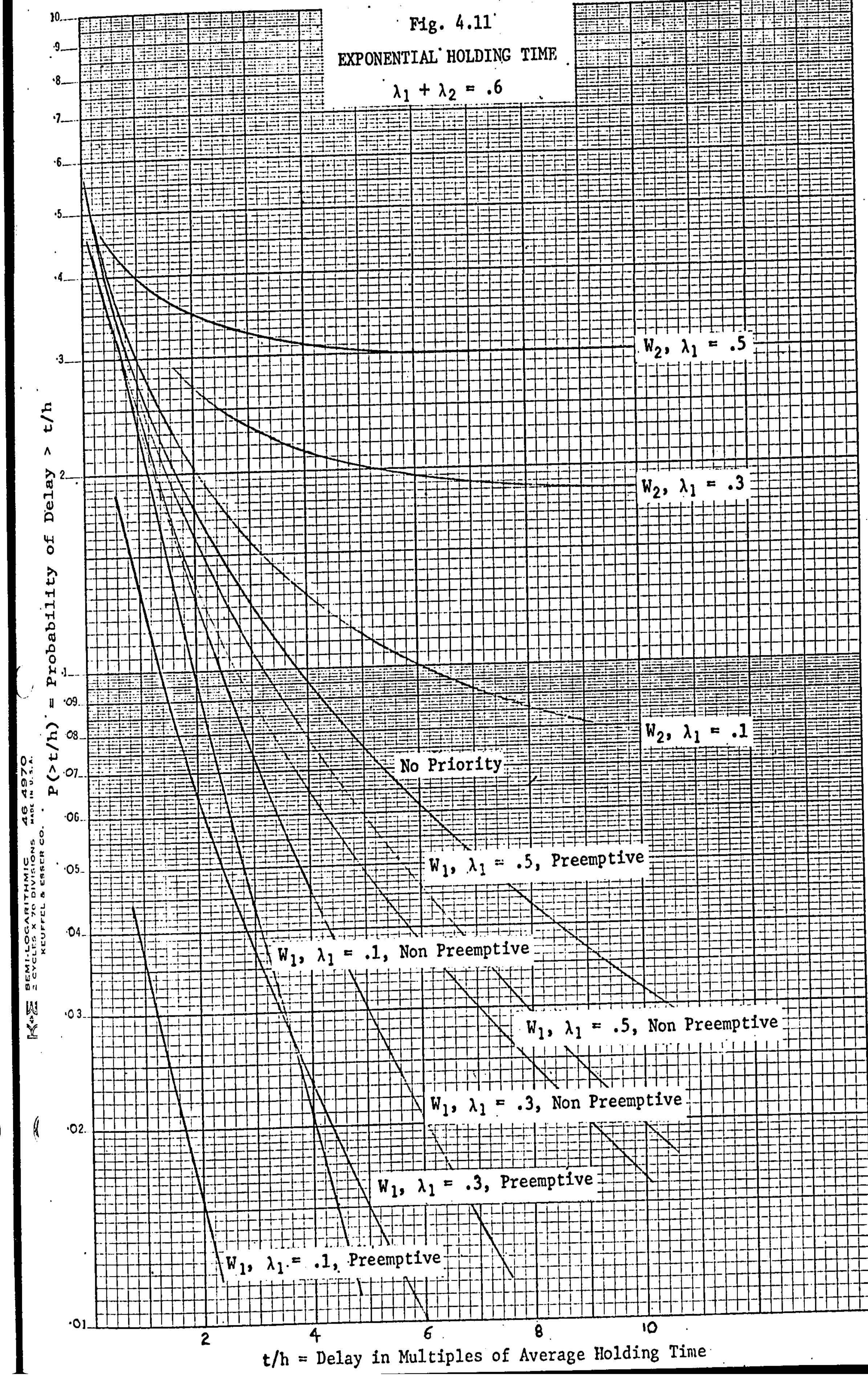
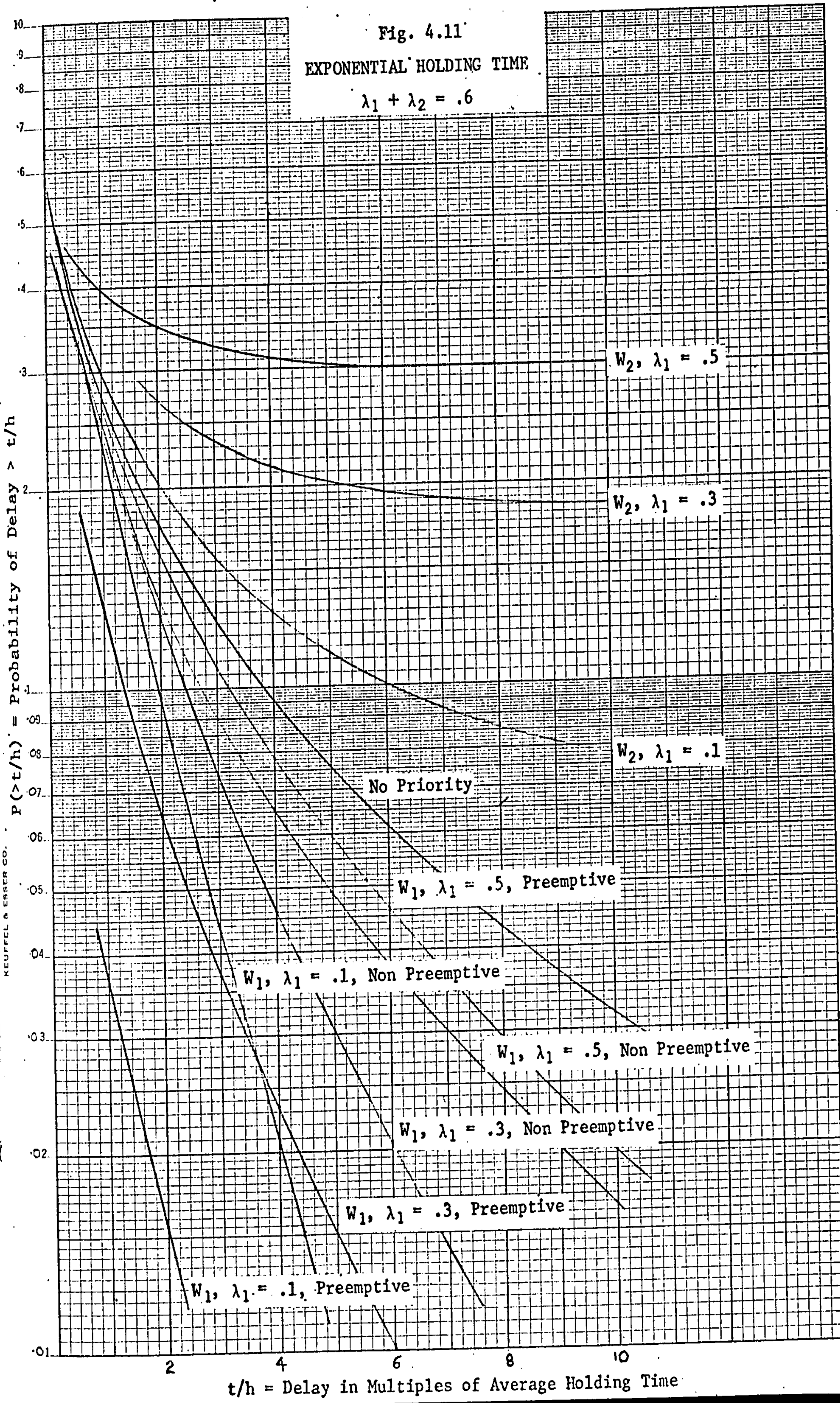


Fig. 4.11

EXPONENTIAL HOLDING TIME

$\lambda_1 + \lambda_2 = .6$



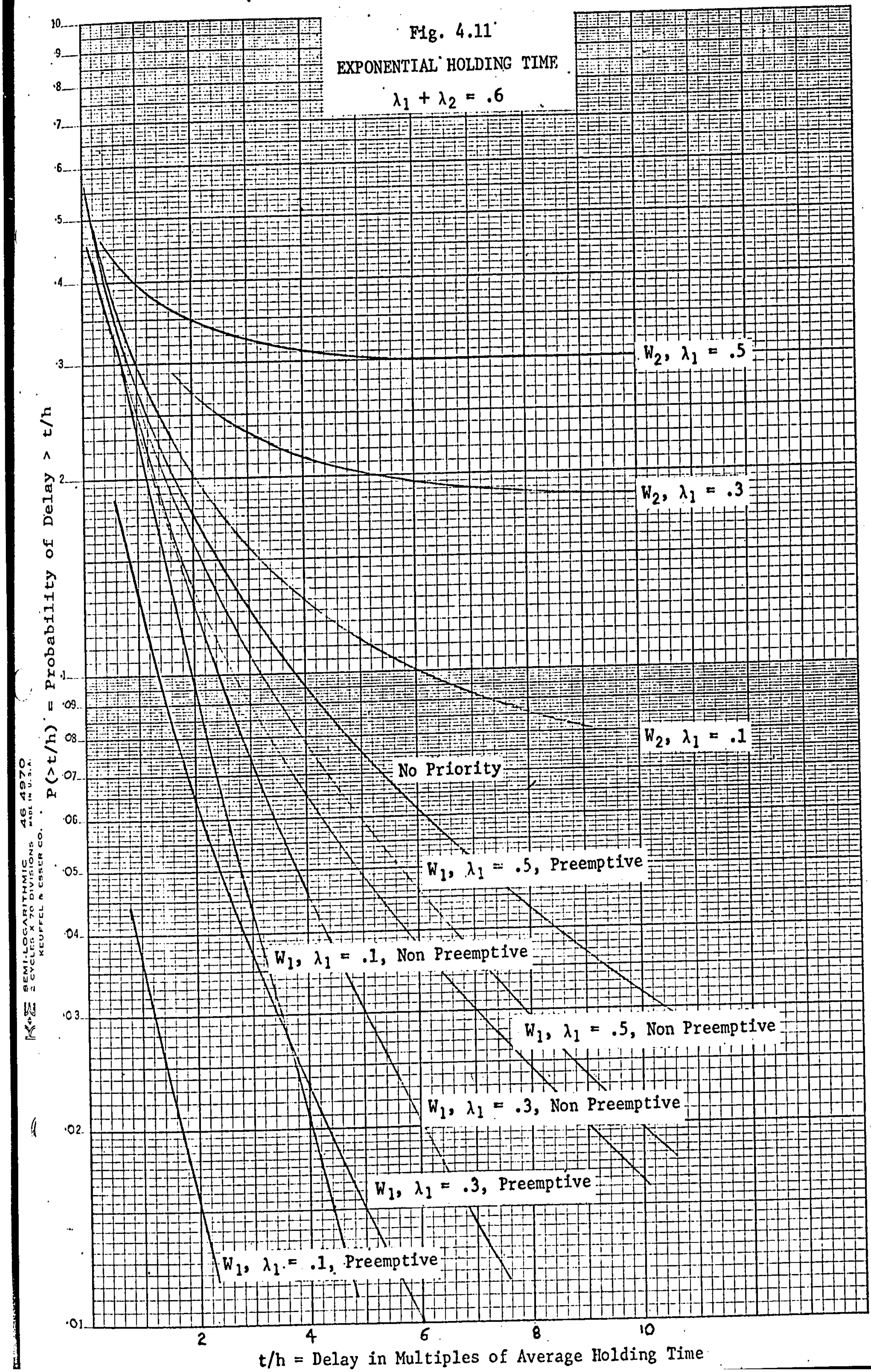


Fig. 4.11
EXPONENTIAL HOLDING TIME
$\lambda_1 + \lambda_2 = .6$
W_2, λ_1 = .5
W_2, λ_1 = .3
W_2, λ_1 = .1
No Priority
W_1, λ_1 = .5, Preemptive
W_1, λ_1 = .1, Non Preemptive
W_1, λ_1 = .5, Non Preemptive
W_1, λ_1 = .3, Non Preemptive
W_1, λ_1 = .3, Preemptive
W_1, λ_1 = .1, Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
SEMI-LOGARITHMIC 46 4970
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

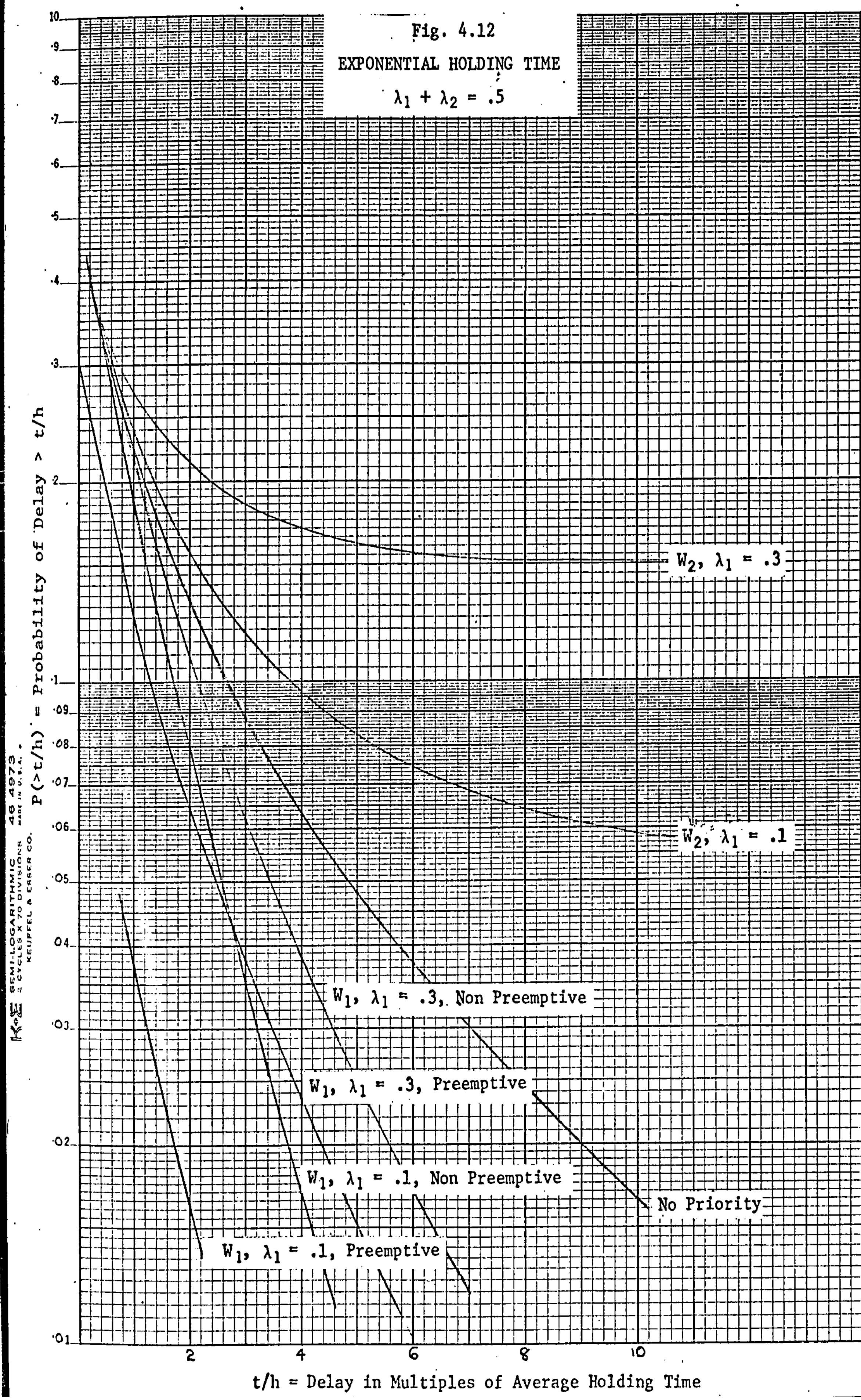


Fig. 4.12
EXPONENTIAL HOLDING TIME
λ1 + λ2 = .5
W2, λ1 = .3
W2, λ1 = .1
W1, λ1 = .3, Non Preemptive
W1, λ1 = .3, Preemptive
W1, λ1 = .1, Non Preemptive
No Priority
W1, λ1 = .1, Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
K&E SEMI-LOGARITHMIC 46 4973
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

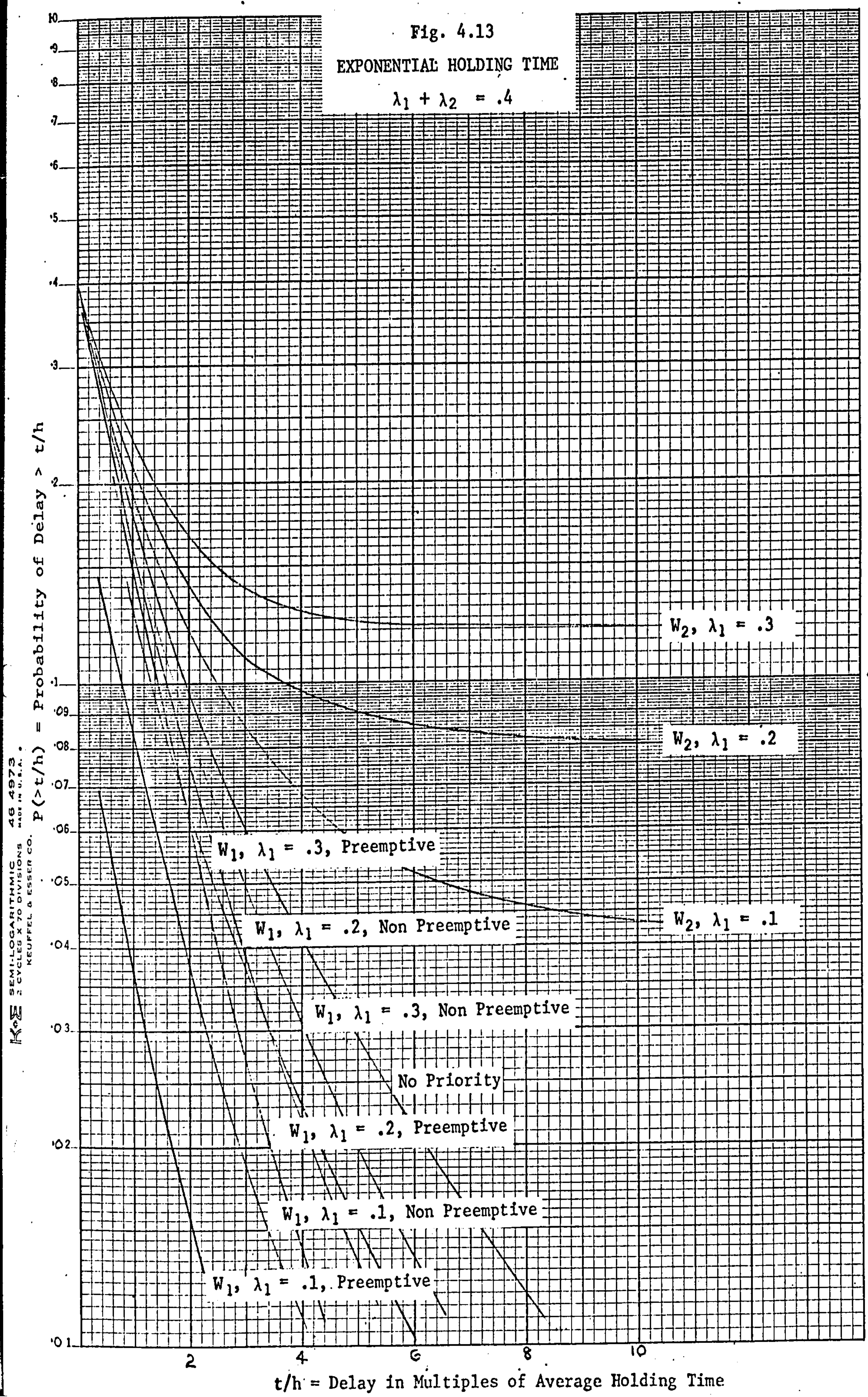


Fig. 4.13
EXPONENTIAL HOLDING TIME
$\lambda_1 + \lambda_2 = .4$
$W_2, \lambda_1 = .3$
$W_2, \lambda_1 = .2$
$W_2, \lambda_1 = .1$
$W_1, \lambda_1 = .3$, Preemptive
$W_1, \lambda_1 = .2$, Non Preemptive
$W_1, \lambda_1 = .3$, Non Preemptive
No Priority
$W_1, \lambda_1 = .2$, Preemptive
$W_1, \lambda_1 = .1$, Non Preemptive
$W_1, \lambda_1 = .1$, Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
2
4
6
8
10
SEMI-LOGARITHMIC 46 4973
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

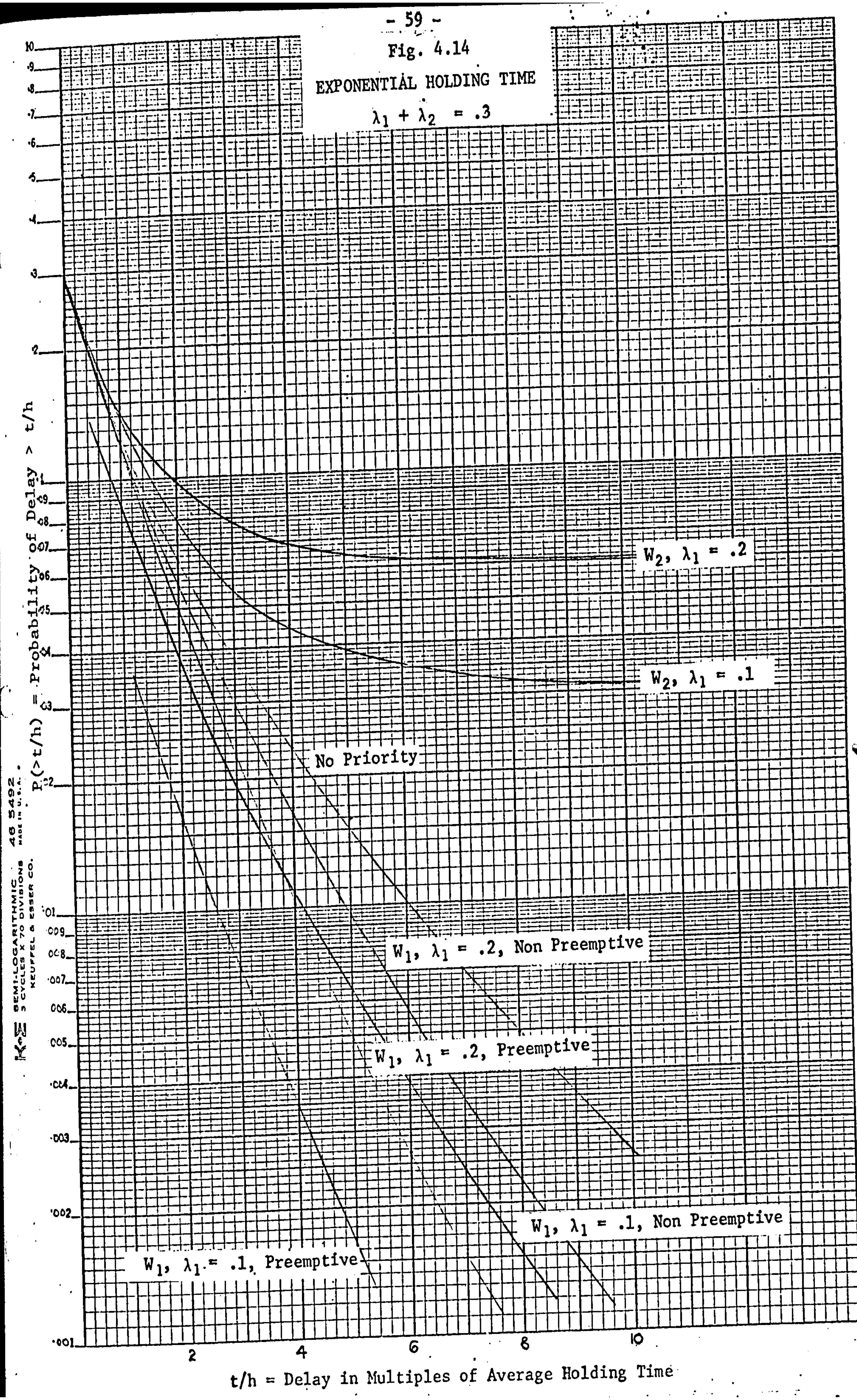


Fig. 4.14
EXPONENTIAL HOLDING TIME
$\lambda_1 + \lambda_2 = .3$
$W_2, \lambda_1 = .2$
$W_2, \lambda_1 = .1$
No Priority
$W_1, \lambda_1 = .2$, Non Preemptive
$W_1, \lambda_1 = .2$, Preemptive
$W_1, \lambda_1 = .1$, Non Preemptive
$W_1, \lambda_1 = .1$, Preemptive
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
2
4
6
8
10
K&E SEMI-LOGARITHMIC 46 5492
3 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

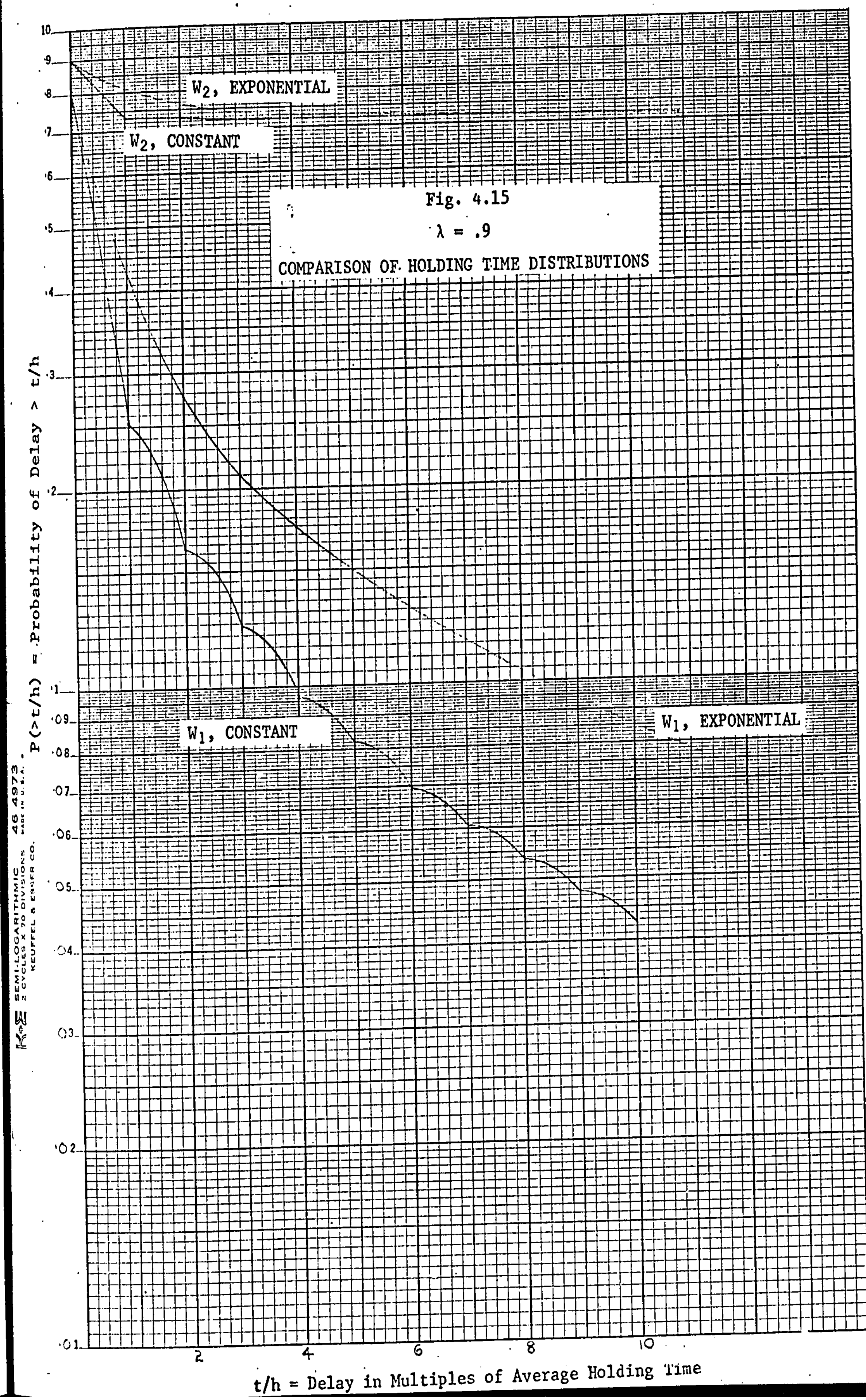


Fig. 4.15
λ = .9
COMPARISON OF HOLDING TIME DISTRIBUTIONS
W2, EXPONENTIAL
W2, CONSTANT
W1, CONSTANT
W1, EXPONENTIAL
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
10
.9
.8
.7
.6
.5
.4
.3
.2
.1
.09
.08
.07
.06
.05
.04
.03
.02
.01
2
4
6
8
10
K&E SEMI-LOGARITHMIC 46 4973
2 CYCLES X 70 DIVISIONS
KEUFFEL & ESSER CO.

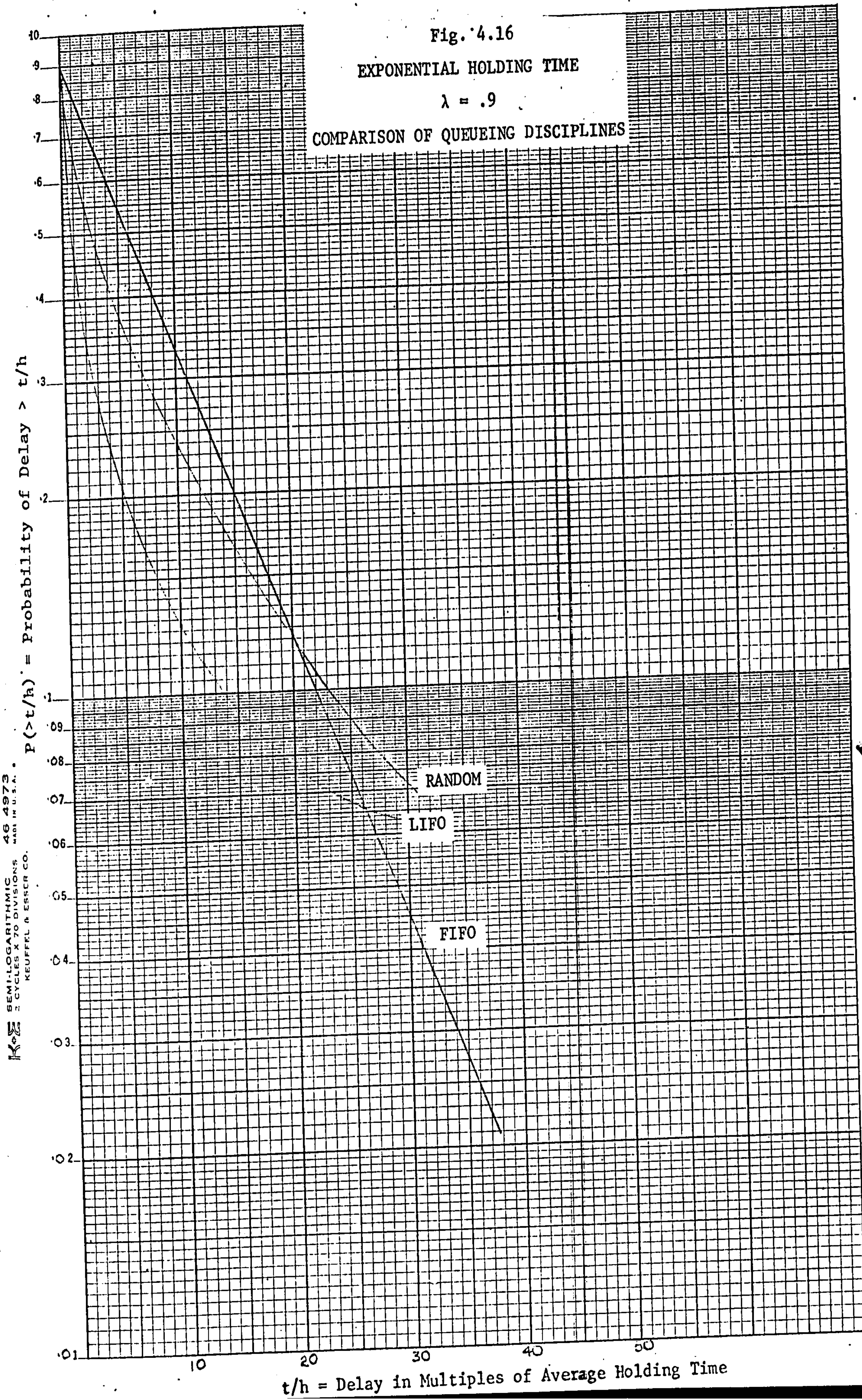


Fig. 4.16
EXPONENTIAL HOLDING TIME
λ = .9
COMPARISON OF QUEUEING DISCIPLINES
RANDOM
LIFO
FIFO
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
10
9
8
7
6
5
4
3
2
1
09
08
07
06
05
04
03
02
01
10
20
30
40
50
K+E SEMI-LOGARITHMIC 46 4973
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

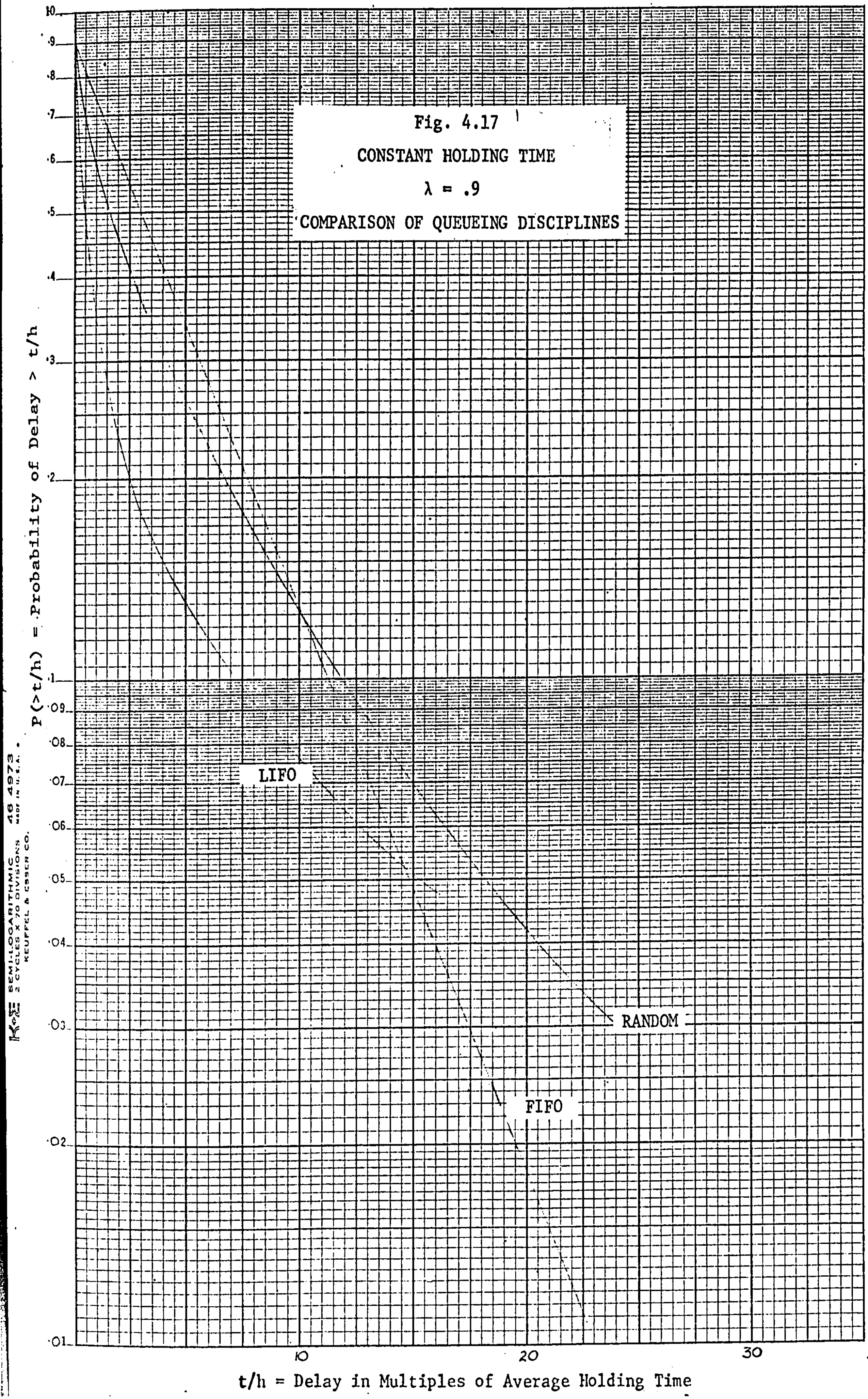


Fig. 4.17
CONSTANT HOLDING TIME
λ = .9
COMPARISON OF QUEUEING DISCIPLINES
P(>t/h) = Probability of Delay > t/h
t/h = Delay in Multiples of Average Holding Time
LIFO
RANDOM
FIFO
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
.09
.08
.07
.06
.05
.04
.03
.02
.01
10
20
30
SEMI-LOGARITHMIC 46 4973
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

CHAPTER V

Priority Systems with Random Order of Service

1. INTRODUCTION

In this chapter, we will analyse a model which is descriptive of the control unit when requests for service are taken from the queue at random within a priority group.

The results obtained here will also be useful for the cases when the queue discipline is somewhere between FIFO and Random. The results will be useful in the sense that the delay distribution, when the queue discipline is Random, can be used as a bound on the distributions to be expected in the cases where the queue discipline deviates from FIFO towards Random. This situation occurs in some electromechanical systems. The queue discipline is determined by a wired preference circuit. In order to provide completely random service the preference circuit would need to be very complicated and correspondingly expensive. So generally a compromise is made between expense and the service given. This compromise tends to make the queue discipline somewhere between FIFO and Random.

The term, bound, used in the previous context does not mean a bounding function in the strictest sense. Actually the delay distribution for models which only differ in queue discipline will cross each other and hence no individual distribution can be a true bound for a family of such distributions.

The basic model of interest in this chapter is described in section 2. The analysis is carried out in section 4 under the two priority disciplines and the assumptions of a Poisson Input Process and a holding time distribution which is negative exponential. The mean holding time is different for each priority. The chapter concludes with a discussion of the results.

2. THE MODEL

Calls of two priority classes are considered to be arriving at a single server in accordance with a Poisson process of total intensity λ. The calls are served in order of priority class, the call with the lowest priority class number being served first. Calls of the same priority class are served in a random order. Two types of service are considered:

Case (i): Non-preemptive type

The current service is never interrupted. This is true even if the arriving call has priority over the call in service at that time.

Case (ii): Preemptive Resume Type

The current service can be interrupted. If the arriving call has a lower priority number than the call being served, then the server interrupts the current service and immediately starts serving the call of higher priority. The interrupted service is resumed when there are no more calls of lower numbered priority in the system. An alternative discipline, which is not considered, would be to place the interrupted call back in the queue and let it complete for service with calls of the same priority class.

3. NOTATION

The service times of the calls are mutually independent random variables and are independent of the inter-arrival times. Denote by $H_p(t)$, $(p=1,2)$ the probability that the service time of a call with priority p is $\leqslant$ t.

Let
$$H_1(t) = 1-e^{-\mu_1 t}$$
$$H_2(t) = 1-e^{-\mu_2 t}$$

where μ_1 and μ_2 are the mean service rates of the priority 1 and priority 2 type calls, respectively. The arrivals with priority p follow a Poisson process with intensity λp.

Let

$W_p(t)$ = the queueing-time distribution for a priority p type call, ie. the probability that a priority p type call will wait at the most t.

$Q_p(t)$ = the distribution of the time to serve the first priority p type call, given that the server is busy when the test call arrives.

$W_{k,p}(t)$ = the probability that the service of a given priority p type call among k priority p type calls starts within time t, under the condition that k (k=1,2,...) calls of priority p are waiting in the system when a priority p service is about to begin, and that the time is measured from that instant.

$P(p,m;n)$ = the state probability that a priority p (p=1,2) call is in service, and that there are m priority 1 calls and n priority 2 calls in the system.

P_0 = the probability that the system is idle.

$M_2(t)$ = the distribution of the time to serve all the priority 1 type calls waiting when the priority 2 type call under consideration arrives in the system (including the remaining part of the current service time).

$B_2(t)$ = the distribution of the completion time for priority 2 type calls.

If U(t) is the distribution of a non-negative random variable, and

$$U(n) = \int_0^\infty t^n \, dU(t)$$

let

$$U^*(s) = \int_0^\infty e^{-st} dU(t)$$

$P(p)$ = the probability that an arriving call has priority $\leq p$.

Under the condition that an arriving call has a priority $\leq p$, the probability that its holding time is $\leq t$ is given by:

$$A_p(t) = \frac{1}{P(p)} \int_0^p H_y(t)\, dP(y)$$

In order that a stationary distribution of the waiting times may exist, we have assumed that $(\lambda_1/\mu_1 + \lambda_2/\mu_2) < 1$.

4. ANALYSIS

BUSY PERIODS AND COMPLETION TIMES

$D(t)$ is the distribution for the busy period in a single server system with Poisson input of intensity λ_1, and with identically distributed, mutually independent service times having the distribution $H_1(x)$.

$D^*(s)$ has been shown[24] to be the root with the smallest absolute value in z of the equation:

$$z = A_1^*(s+\lambda_1-\lambda_1 z)$$

Therefore:

$$D^{(1)} = 1/(\mu_1-\lambda_1) \tag{5.1}$$

$$D^{(2)} = 2\mu_1/(\mu_1-\lambda_1)^3 \tag{5.2}$$

Gaver[21] defines completion times as the duration of the period that elapses between the instant at which the service of a call of priority p (the n^{th} to arrive) begins and the instant at which service of the next may begin. This period is simply the n^{th} priority p type calls service time if there are no interruptions.

In our system, the priority 1 completion time is just the holding time of a priority 1 type call. The completion time for priority 2 type calls has been described by the distribution $B_2(t)$. From Gaver[21] therefore,

$$B_2{*}(s) = H_2{*}(s+\lambda_1-\lambda_1 D{*}(s))$$

and

$$B_2^{(1)} = H_2^{(1)}(1+\lambda_1 D^{(1)}) = \mu_1/\mu_2(\mu_1-\lambda_1) \tag{5.3}$$

$$B_2^{(2)} = H_2^{(2)}(1+\lambda_1 D^{(1)})^2 + H_2^{(1)}\lambda_1 D^{(2)}$$

$$= 2\mu_1\left[\mu_1(\mu_1-\lambda_1)+\lambda_1\mu_2\right]/\mu_2{}^2(\mu_1-\lambda_1)^3 \tag{5.4}$$

Although in case (ii) the order of servicing may be different the distribution of the completion time interval is the same as that in case (i). This is true for the queue discipline assumed but would not be true for the case mentioned earlier where the interrupted call competes for service with calls of the same priority class.

DISTRIBUTION OF $W_{k,p}(t)$

In case (i), the waiting time of a call can be split into two intervals. The first interval is the time elapsed from the instant the call arrived to the instant when the first call of the same priority class is served. This interval has the distribution $Q_p(t)$. If k calls of the same

Gaver[21] defines completion times as the duration of the period that elapses between the instant at which the service of a call of priority p (the n^{th} to arrive) begins and the instant at which service of the next may begin. This period is simply the n^{th} priority p type calls service time if there are no interruptions.

In our system, the priority 1 completion time is just the holding time of a priority 1 type call. The completion time for priority 2 type calls has been described by the distribution $B_2(t)$. From Gaver[21] therefore,

$$B_2^*(s) = H_2^*(s+\lambda_1-\lambda_1 D^*(s))$$

and

$$B_2^{(1)} = H_2^{(1)}(1+\lambda_1 D^{(1)}) = \mu_1/\mu_2(\mu_1-\lambda_1) \qquad (5.3)$$

$$\begin{aligned} B_2^{(2)} &= H_2^{(2)}(1+\lambda_1 D^{(1)})^2 + H_2^{(1)}\lambda_1 D^{(2)} \\ &= 2\mu_1\left[\mu_1(\mu_1-\lambda_1)+\lambda_1\mu_2\right]/\mu_2^{\,2}(\mu_1-\lambda_1)^3 \end{aligned} \qquad (5.4)$$

Although in case (ii) the order of servicing may be different the distribution of the completion time interval is the same as that in case (i). This is true for the queue discipline assumed but would not be true for the case mentioned earlier where the interrupted call competes for service with calls of the same priority class.

DISTRIBUTION OF $W_{k,p}(t)$

In case (i), the waiting time of a call can be split into two intervals. The first interval is the time elapsed from the instant the call arrived to the instant when the first call of the same priority class is served. This interval has the distribution $Q_p(t)$. If k calls of the same priority class as the test call are waiting at the end of this first interval then the second interval consists of the sum of a number of completion times and has the distribution $W_{k,p}(t)$. This distribution can be found from Takács[28] by substituting the completion time distribution for Takács holding time distribution.

ie.
$$W_{k,1}^{*}(s) = 1-(k-1)\int_{\alpha_1(s)}^{1} \exp\left(-\int_{u}^{1} \frac{dv}{v-\Psi_1(s+\lambda_1-\lambda_1 v)}\right) u^{k-2}du$$

and
$$W_{k,1}^{(1)} = (k-1)/(2\mu_1-\lambda_1) \tag{5.5}$$

$$W_{k,1}^{(2)} = 2(k-1)\{(k-2)+(6\mu_1-\lambda_1)/(2\mu_1-\lambda_1)\}/(2\mu_1-\lambda_1)(3\mu_1-2\lambda_1) \tag{5.6}$$

where $\alpha_1(s)$ = the Laplace Stieltjes Transform of the busy period, with arrival intensity λ_1, and holding time distribution $H_1(t)$.

$$\Psi_1(s) = \int_0^\infty e^{-sx}\, dH_1(x)$$

Similarly,
$$W_{k,2}^{*}(s) = 1-(k-1)\int_{\alpha_2(s)}^{1} \exp\left(-\int_{u}^{1} \frac{dv}{v-\Psi_2(s+\lambda_1-\lambda_1 v)}\right) u^{k-2}du$$

and
$$W_{k,2}^{(1)} = (k-1)B_2^{(1)}/(2-\lambda_2 B_2^{(1)})$$

$$W_{k,2}^{(2)} = \frac{2(B_2^{(1)})^2(k-1)(k-2)}{(2-\lambda_2 B_2^{(1)})(3-2\lambda_2 B_2^{(1)})} + \frac{(6-\lambda_2 B_2^{(1)})(k-1)B_2^{(2)}}{(2-\lambda_2 B_2^{(1)})^2(3-2\lambda_2 B_2^{(1)})}$$

where $\alpha_2(s)$ = the Laplace Stieltjes Transform of the busy period with arrival intensity λ_2 and holding time distribution $B_2(t)$.

$$\Psi_2(s) = \int_0^\infty e^{-sx} dH_2(x)$$

Therefore from (5.3) and (5.4)

$$W_{k,2}^{(1)} = (k-1)\mu_1/(2\mu_2\mu_1-2\mu_2\lambda_1-\lambda_2\mu_1) \tag{5.7}$$

and $W_{k,2}^{(2)} = 2(k-1)(k-2)\mu_1^2/(2\mu_1\mu_2-2\mu_2\lambda_1-\lambda_2\mu_1)(3\mu_1\mu_2-3\mu_2\lambda_1-2\lambda_2\mu_1)$

$$+\frac{(k-1)(6\mu_1\mu_2-6\mu_2\lambda_1-\lambda_2\mu_1)\left[\mu_1(\mu_1-\lambda_1)+\lambda_1\mu_2\right]2\mu_1}{(\mu_1-\lambda_1)(2\mu_1\mu_2-2\mu_2\lambda_1-\lambda_2\mu_1)^2(3\mu_1\mu_2-3\mu_2\lambda_1-2\lambda_2\mu_1)} \tag{5.8}$$

In case (ii), although the order of servicing is different the results are the same as calculated for case (i).

THE DETERMINATION OF $W_1(t)$ FOR CASE (ii)

In case (ii), $W_1(t)$ is the same distribution as one would encounter in considering priority 1 type calls in isolation, since priority 1 calls preempt priority 2 calls in service.

Thus, from Takacs[28], if $\lambda_1/\mu_1 < 1$.

$$W_1^*(s) = (1-\lambda_1/\mu_1)\left[1+\frac{\lambda_1}{s}\int_{\gamma(s)}^{1} \exp\left\{-\int_u^1 \frac{dv}{u-\Psi(s+\lambda_1(1-v))}\right\}.\left\{1+\frac{dv}{u-\Psi(\lambda_1-\lambda_1 u)}\right\}du\right]$$

where,

$$\Psi(x) = \int_0^\infty e^{-xy} dH_1(y) = \mu_1/(\mu_1+x)$$

and

$$\gamma(s) = (\lambda_1+\mu_1+s - \sqrt{(\lambda_1+\mu_1+s)^2-4\lambda_1\mu_1})/2\lambda_1$$

The first two moments are:

$$W_1^{(1)} = \lambda_1/\mu_1(\mu_1-\lambda_1) \tag{5.9}$$

$$W_1^{(2)} = 4\lambda_1/(2\mu_1-\lambda_1)(\mu_1-\lambda_1)^2 \tag{5.10}$$

Equation (5.9) was found by Miller[20] for the first-in first-out queue discipline.

DISTRIBUTION OF $W_1(t)$ FOR CASE (i)

The priority 1 test call arrives in the system at t_0. Its queueing time can be considered as consisting of two parts. The first is the time elapsed from t_0 to the instant when the first priority 1 call is served. This interval will be described by the residual holding time distribution of the call in service at t_0. Since we are considering exponentially distributed holding times, the residual holding time distribution is the same as the normal holding time distribution. If j calls (j=0, 1, ...) arrive in this first interval, and there are m calls of priority 1 in the system at t_0 when the test call arrived, then the second interval has the distribution $W_{j+m+1,1}(t)$ or $W_{j+m,1}(t)$ depending upon whether a priority 2 or a priority 1 call was in service at t_0.

ie.

$$W_1(t) = P_0 + \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} P(1,m,n)\left(\int_0^t \frac{e^{-\lambda_1 x}(\lambda_1 x)^j}{j!}\, dH_1(x)\right)**W_{j+m,1}(t)$$

$$+ \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} \sum_{j=0}^{\infty} P(2,m,n)\left(\int_0^t \frac{e^{-\lambda_1 x}(\lambda_1 x)^j}{j!}\, dH_2(x)\right)**W_{j+m+1,1}(t)$$

where ** denotes convolution.

Therefore taking the Laplace Stieltjes Transform:

$$W_1{*}(s) = P_0 + \sum_{m=1}^{\infty}\sum_{n=0}^{\infty}\sum_{j=0}^{\infty} P(1,m,n)\left(\int_0^{\infty} \frac{e^{-(\lambda_1+s)t}(\lambda_1 t)^j}{j!}\,dH_1(t)\right)W^{*}_{j+m,1}(s)$$

$$+ \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{j=0}^{\infty} P(2,m,n)\left(\int_0^{\infty} \frac{e^{-(\lambda_1+s)t}(\lambda_1 t)^j}{j!}\,dH_2(t)\right)W^{*}_{j+m+1,1}(s)$$

and

$$W_1^{(1)} = \sum_{m=1}^{\infty}\sum_{n=0}^{\infty}\sum_{j=0}^{\infty} P(1,m,n)\left(\int_0^{\infty} \frac{e^{-\lambda_1 t}(\lambda_1 t)^j}{j!}\,dH_1(t)\right)W_{j+m,1}^{(1)}$$

$$+ \sum_{m=0}^{\infty}\sum_{n=1}^{\infty}\sum_{j=0}^{\infty} P(2,m,n)\left(\int_0^{\infty} \frac{e^{-\lambda_1 t}(\lambda_1 t)^j}{j!}\,dH_2(t)\right)W_{j+m+1,1}^{(1)}$$

$$+ H_1^{(1)}\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} P(1,m,n) + H_2^{(1)}\sum_{m=0}^{\infty}\sum_{n=1}^{\infty} P(2,m,n) \tag{5.11}$$

$$W_1^{(2)} = \sum_{m=1}^{\infty}\sum_{n=0}^{\infty}\sum_{j=0}^{\infty} P(1,m,n)\int_0^{\infty} \frac{e^{-\lambda_1 t}(\lambda_1 t)^j}{j!}\,\{W_{j+m,1}^{(2)} + 2tW_{j+m,1}^{(1)}\}dH_1(t)$$

$$+ \sum_{m=0}^{\infty}\sum_{n=1}^{\infty}\sum_{j=0}^{\infty} P(2,m,n)\int_0^{\infty} \frac{e^{-\lambda t}(\lambda_1 t)^j}{j!}\,\{W_{j+m+1,1}^{(2)}+2tW_{j+m+1,1}^{(1)}\}dH_2(t)$$

$$+ H_1^{(2)}\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} P(1,m,n) + H_2^{(2)}\sum_{m=0}^{\infty}\sum_{n=1}^{\infty} P(2,m,n) \tag{5.12}$$

THE DETERMINATION OF $W_2(t)$

The difference between case (i) and case (ii) is that for case (ii) calls are allowed to preempt calls of higher numbered priority. In the case of priority 2 type calls there are no calls of higher numbered priority so that the waiting time distribution $W_2(t)$ will be the same for both case (i) and case (ii).

Let the priority 2 test call arrive in the system at t_0. If the system is idle at t_0 then it goes straight into service. If the server is occupied then the calls queueing time can be considered as consisting of two parts. The first part is the time elapsed from t_0 to the instant when the first priority 2 call is served. This interval has the distribution function $Q_2(t)$. If j calls (j=0,1,...) of priority 2 arrive in that interval, and there were n calls of priority 2 in the system at t_0, when the test call arrives then the second interval has a distribution $W_{j+n+1,2}(t)$ or $W_{j+n,2}(t)$ depending upon whether a priority 1 or a priority 2 call was in service at t_0, ie.

$$W_2(t) = P_0 + \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} P(1,m,n) \left(\int_0^t \frac{e^{-\lambda_2 y}(\lambda_2 y)^j}{j!} dQ_2(y)\right) ** W_{j+n+1,2}(t)$$

$$+ \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} \sum_{j=0}^{\infty} P(2,m,n) \left(\int_0^t \frac{e^{-\lambda_2 y}(\lambda_2 y)^j}{j!} dQ_2(y)\right) ** W_{j+n,2}(t)$$

where ** denotes confolution.

Taking the Laplace Stieltjes Transform:

$$W_2{*}(s) = P_0 + \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} P(1,m,n)\left(\int_0^{\infty} \frac{e^{-(s+\lambda_2)t}(\lambda_2 t)^j}{j!} dQ_2(t)\right) W^*_{j+n+1,2}(s)$$

$$+ \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} \sum_{j=0}^{\infty} P(2,m,n)\left(\int_0^{\infty} \frac{e^{-(s+\lambda_2)t}(\lambda_2 t)^j}{j!} dQ_2(t)\right) W^*_{j+n,2}(s)$$

The moments can be found using:

$$W_2^{(n)} = \left[\frac{d^n}{ds^n} \{(-1)^n W_2^*(s)\}\right]_{s=0}$$

and in particular

$$W_2^{(1)} = \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} P(1,m,n) \left(\int_0^{\infty} \frac{e^{-\lambda_2 t}(\lambda_2 t)^j}{j!} dQ_2(t)\right) W_{j+n+1,2}^{(1)}$$

$$+ \sum_{n=0}^{\infty} \sum_{n=1}^{\infty} \sum_{j=0}^{\infty} P(2,m,n) \left(\int_0^{\infty} \frac{e^{-\lambda_2 t}(\lambda_2 t)^j}{j!} dQ_2(t)\right) W_{j+n,2}^{(1)}$$

$$+ Q_2^{(1)} \left(\sum_{m=1}^{\infty} \sum_{n=0}^{\infty} P(1,m,n) + \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} P(2,m,n)\right) \tag{5.13}$$

and

$$W_2^{(2)} = \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} P(1,m,n) \int_0^{\infty} \frac{e^{-\lambda_2 t}(\lambda_2 t)^j}{j!} \{W_{j+n+1,2}^{(2)} + 2tW_{j+n+1,2}^{(1)}\} dQ_2(t)$$

$$+ \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} \sum_{j=0}^{\infty} P(2,m,n) \int_0^{\infty} \frac{e^{-\lambda_2 t}(\lambda_2 t)^j}{j!} \{W_{j+n,2}^{(2)} + 2tW_{j+n,2}^{(1)}\} dQ_2(t)$$

$$+ Q_2^{(2)} \left\{\sum_{m=0}^{\infty} \sum_{n=1}^{\infty} P(2,m,n) + \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} P(1,m,n)\right\} \tag{5.14}$$

All that is required now is the distribution function $Q_2(t)$ and the state probabilities.

DERIVATION OF $Q_2(t)$

The waiting time defined by the distribution $Q_2(t)$ can be considered as being composed of two intervals. The first is the time needed to complete the service of all calls with priority 1 which are in the system at t_0 (including the remaining part of the current service). The distribution of the first interval has been denoted by $M_2(t)$. The second interval is the time to serve the priority 1 calls which arrive in the two intervals. If j priority 1 calls $(j=0,1,\ldots)$ arrive in the first interval, then the distribution of the second interval is that of the total length of j independent busy periods, each with the distribution D(t), i.e.,

$$Q_2(t) = \Big(\sum_{j=0}^{\infty} \Big(\int_0^t \frac{e^{-\lambda_1 y}(\lambda_1 y)^j}{j!}\, dM_2(y)\Big)**D^j(t) - P_0\Big)/(1-P_0)$$

where

$D^j(t)$ is the jth iterated convolution of D(t) with itself.

$$1-P_0 = \lambda_1/\mu_1 + \lambda_2/\mu_2,$$

** denotes convolution.

Therefore:

$$Q_2{*}(s) = (M_2(s+\lambda_1-\lambda_1 D{*}(s)) - P_0)/(1-P_0)$$

and $$Q_2^{(1)} = M_2^{(1)}(1+\lambda_1 D^{(1)})/(1-P_0) \tag{5.15}$$

$$Q_2^{(2)} = \left[M_2^{(2)}(1+\lambda_1 D^{(1)})^2 + M_2^{(1)}\lambda_1 D^{(2)}\right]/(1-P_0) \tag{5.16}$$

The distribution $M_2(t)$ can be found from Takács[23]. In Takács study, the distribution of the time needed to complete service of all calls of class p or less present in the system at t_0 is derived while in the present study the time to service all units with priority less than p only is needed. Although the queue disciplines in each class in Takács' study and the present study are different, the queue length distributions in the two cases are the same. The waiting time to service calls of priority less than p present at t_0 in this study can be obtained from Takacs results by replacing p by (p-1).

Therefore,

$$M_2^*(s) = ((1-\lambda_1/\mu_1-\lambda_2/\mu_2)s+\lambda(1-A_2^*(s))-\lambda_1(1-A_1^*(s)))/(s-\lambda_1+\lambda_1 A_1^*(s))$$

$$M_2^{(1)} = (\lambda_1\mu_2^2 + \lambda_2\mu_1^2)/\mu_2^2\mu_1(\mu_1-\lambda_1) \tag{5.17}$$

$$M_2^{(2)} = \frac{2(\lambda_1\mu_2^3 + \lambda_2\lambda_1^3 - \lambda_1\lambda_2\mu_1^2 + \lambda_1\lambda_2\mu_1\mu_2)}{(\mu_1-\lambda_1)^2\ \mu_1\mu_2^3} \tag{5.18}$$

Therefore from (5.1), (5.15) and (5.17)

$$Q_2^{(1)} = (\lambda_1\mu_2^2 + \lambda_2\mu_1^2)\mu_1/\mu_2(\mu_1-\lambda_1)^2(\lambda_1\mu_2 + \lambda_2\mu_1) \tag{5.19}$$

and from (5.1), (5.2), (5.16), (5.17) and (5.18)

$$Q_2^{(2)} = \frac{2\{\lambda_1\mu_1\mu_2^3 + \lambda_2\mu_1^4 - \lambda_1\lambda_2\mu_1^3 + 2\lambda_1\lambda_2\mu_1^2\mu_2 + \lambda_1^2\mu_2^3\}\mu_1}{(\mu_1-\lambda_1)^4\mu_2^2(\lambda_1\mu_2+\lambda_2\mu_1)} \tag{5.20}$$

STATE PROBABILITIES

Case (i):

Morse[16] has derived the generating function of the state probabilities.

Let:

$$G(s,t) = \sum_{m,n} P(m,n)s^m t^n$$

then:

$$G(s,t) = \frac{\mu_1(s-1)(1-\lambda_1/\mu_1-\lambda_2/\mu_2)}{\lambda_1 s(1-s)+\lambda_2 s(1-t)+\mu_1(s-1)}$$

$$+ \frac{P_{20}(t)\{\lambda_1+\lambda_2(1-t)+\mu_2\}\{\mu_1(s-1)t-\mu_2 s(t-1)\}}{t\{\lambda_1(1-s)+\lambda_2(1-t)+\mu_2\}\{\lambda_1(1-s)s+\lambda_2(1-t)s+\mu_1(s-1)\}}$$

Where

$$P_{20}(t) = \frac{t\{1-\lambda_1/\mu_1-\lambda_2/\mu_2\}\{\lambda_1+\lambda_2(1-t)-\mu_1\alpha(t)\}\{(\mu_1-\mu_2)(\lambda_1+\lambda_2(1-t)+\mu_2\}-\lambda_1\mu_1}{\{\lambda_1+\lambda_2(1-t)+\mu_2\}\{\lambda_1\mu_1 t-(\mu_1-\mu_2)t(\lambda_1+\lambda_2(1-t)+\mu_2)+\mu_2(\mu_1(1-\alpha(t))-\mu_2)\}}$$

and

$$\alpha(t) = \frac{\lambda_1+\lambda_2(1-t)+\mu_1-\sqrt{(\lambda_1+\lambda_2(1-t)+\mu_1)^2-4\lambda_1\mu_1}}{2\mu_1} \tag{5.21}$$

From this Morse shows:

$$\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} P(1,m,n) = \lambda_1/\mu_1 \tag{5.22}$$

$$\sum_{m=0}^{\infty}\sum_{n=1}^{\infty} P(2,m,n) = \lambda_2/\mu_2 \tag{5.23}$$

$$\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} nP(1,m,n) + \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} nP(2,m,n) = \frac{\lambda_2}{\mu_2} + \frac{\lambda_2(\lambda_1\mu_2{}^2+\lambda_2\mu_1{}^2)}{(\mu_1-\lambda_1)(\mu_1\mu_2-\lambda_1\mu_2-\lambda_2\mu_1)\mu_2} \tag{5.24}$$

$$\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} mP(1,m,n) + \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} mP(2,m,n) = \frac{\lambda_1}{(\mu_1-\lambda_1)}\left(\frac{\mu_2{}^2+\lambda_2\mu_1}{\mu_2{}^2}\right) \tag{5.25}$$

<u>Case (ii)</u>:

White and Christie[17] have obtained the generating function:

$$G(s,t) = \frac{(1-\lambda_1/\mu_1-\lambda_2/\mu_2)\mu_2(1-t)}{(\mu_1 t\alpha(t) - \lambda_1 t - \lambda_2 t(1-t) - \mu_2(t-1))(1-s\alpha(t))}$$

where $\alpha(t)$ is given by (5.21)

and:
$$\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} P(1,m,n) + \sum_{n=1}^{\infty} P(2,0,n) = \lambda_1/\mu_1 + \lambda_2/\mu_2 \tag{5.26}$$

$$\sum_{n=1}^{\infty} P(2,0,n) = \lambda_2\mu_1/\mu_2(\mu_1-\lambda_1) \tag{5.27}$$

$$\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} nP(1,m,n) + \sum_{n=1}^{\infty} nP(2,0,n) = \frac{\lambda_2(\mu_2\lambda_1+\mu_1{}^2-\mu_1\lambda_1)}{(\mu_2\mu_1-\lambda_2\mu_1-\lambda_1\mu_2)(\mu_1-\lambda_1)} \tag{5.28}$$

$$\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} n^2P(1,m,n) + \sum_{n=1}^{\infty} n^2P(2,0,n) = \frac{2\lambda_1\lambda_2{}^2\mu_1\mu_2}{(\mu_1-\lambda_1)^3(\mu_1\mu_2-\mu_1\lambda_2-\mu_2\lambda_1)}$$

$$+\frac{\lambda_1\lambda_2\mu_2(\mu_1-\lambda_1)(\mu_1\mu_2-\lambda_1\mu_2-\lambda_2\mu_1)+\lambda_2{}^2\left[\mu_1{}^2-\mu_1\lambda_1+\mu_2\lambda_1\right]^2+\mu_1\mu_2\lambda_2(\mu_1-\lambda_1)^3+\lambda_1{}^2\lambda_2{}^2\mu_2{}^2}{(\mu_1-\lambda_1)^2(\mu_1\mu_2-\mu_1\lambda_2-\mu_2\lambda_1)^2} \tag{5.29}$$

MOMENTS OF THE WAITING TIME DISTRIBUTION

In case (ii),

$W_1{}^{(1)}$ and $W_1{}^{(2)}$ are given in equations (5.9) and (5.10)

From (5.7), (5.13), (5.19), (5.26), (5.27) and (5.28)

$$W_2{}^{(1)} = (\lambda_1\mu_2{}^2+\lambda_2\mu_1{}^2)/(\mu_1-\lambda_1)\mu_2(\mu_1\mu_2-\lambda_1\mu_2-\lambda_2\mu_1) \tag{5.30}$$

From (5.8), (5.14), (5.20), (5.26), (5.27), (5.28) and (5.29)

$$W_2{}^{(2)} = \frac{4\left[\lambda_1\mu_2{}^3 + \lambda_2\mu_1{}^3\right]}{\mu_2\left[\mu_1-\lambda_1\right]\left[2\mu_1\mu_2-2\lambda_1\mu_2-\lambda_2\mu_1\right]\left[\mu_1\mu_2-\lambda_1\mu_2-\lambda_2\mu_1\right]}$$

$$+\frac{4\left[\lambda_1\mu_2{}^2 + \lambda_2\mu_1{}^2\right]^2}{\mu_2\left[\mu_1-\lambda_1\right]\left[2\mu_1\mu_2-2\lambda_1\mu_2-\lambda_2\mu_1\right]\left[\mu_1\mu_2-\lambda_1\mu_2-\lambda_2\mu_1\right]^2}$$

$$+\frac{4\lambda_1\left[\lambda_1\mu_2{}^2+\lambda_2\mu_1{}^2\right]}{(\mu_1-\lambda_1)^2\left[2\mu_1\mu_2-2\lambda_1\mu_2-\lambda_2\mu_1\right]\left[\mu_1\mu_2-\lambda_1\mu_2-\lambda_2\mu_1\right]} \tag{5.31}$$

In case (i),

From (5.5), (5.11), (5.22), (5.23), and (5.25)

$$W_1^{(1)} = (\lambda_1\mu_2^2 + \lambda_2\mu_1^2)/\mu_1\mu_2^2(\mu_1-\lambda_1) \qquad (5.32)$$

and from (5.6), (5.12), (5.27), (5.23) and (5.25)

$$W_1^{(2)} = 4(\lambda_1\mu_2^3 + \lambda_2\mu_1\ (\mu_1^2-\lambda_1\mu_1+\lambda_1\mu_2))/(2\mu_1-\lambda_1)\mu_2^3(\mu_1-\lambda_1)^2 \qquad (5.33)$$

$W_2^{(1)}$ and $W_2^{(2)}$ are the same as those shown in (5.30) and (5.31).

For the first-in first-out discipline (5.30) was found by Miller[20] and (5.32) was found by Cobham[11] and Kesten and Runnenburg[15].

For the last-in first-out discipline (5.9), (5.30) and (5.32) were found by Durr[29].

5. DISCUSSION

Table 1 shows the relationship between the moments of the delay in queue for the various queue disciplines.

	First-in First-out (FIFO)	Random	Last-in First-out (LIFO)
Case (i)			
$W_1^{(1)}$	X_1	X_1	X_1
$W_1^{(2)}$	X_2	$2\mu_1X_2/(2\mu_1-\lambda_1)$	$\mu_1X_2/(\mu_1-\lambda_1)$
$W_2^{(1)}$	X_3	X_3	X_3
$W_2^{(2)}$	X_4	$(2\mu_1\mu_2-2\lambda_1\mu_2)X_4/(2\mu_1\mu_2-2\lambda_1\mu_2-\lambda_2\mu_1)$	$\mu_2(\mu_1-\lambda_1)X_4/(\mu_1\mu_2-\lambda_1\mu_2-\lambda_2\mu_1)$
Case (ii)			
$W_1^{(1)}$	Y_1	Y_1	Y_1
$W_1^{(2)}$	Y_2	$2\mu_1Y_2/(2\mu_1-\lambda_1)$	$\mu_1Y_2/(\mu_1-\lambda_1)$
$W_2^{(1)}$	X_3	X_3	X_3
$W_2^{(2)}$	X_4	$(2\mu_1\mu_2-2\lambda_1\mu_2)X_4/(2\mu_1\mu_2-2\lambda_1\mu_2-\lambda_2\mu_1)$	$\mu_2(\mu_1-\lambda_1)X_4/(\mu_1\mu_2-\lambda_1\mu_2-\lambda_2\mu_1)$

TABLE 1

In each case the average delay is the same. This independence of the order of service may be explained roughly by saying that the average rate at which queues are removed depends only the average rate of arrival of calls and the rate at which they are served.

Let us now look at the second moments. In both cases (i) and (ii),

$$W_1^{(2)}{}_{(\mathrm{LIFO})} = \frac{2\mu_1-\lambda_1}{2(\mu_1-\lambda_1)}\, W_1^{(2)}{}_{(\mathrm{RANDOM})}$$

and $$W_1^{(2)}{}_{(\mathrm{RANDOM})} = \frac{2\mu_1}{2\mu_1-\lambda_1}\, W_1^{(2)}{}_{(\mathrm{FIFO})}$$

Now $\quad 2\mu_1-\lambda_1 > 2\mu_1-2\lambda_1$ and $2\mu_1 > 2\mu_1 - \lambda_1$

Therefore,

$$W_1^{(2)}{}_{(\mathrm{LIFO})} > W_1^{(2)}{}_{(\mathrm{RANDOM})} > W_1^{(2)}{}_{(\mathrm{FIFO})}$$

Similarly

$$W_2^{(2)}{}_{(\mathrm{LIFO})} = \frac{(2\mu_1\mu_2-2\lambda_1\mu_2-\lambda_2\mu_1)}{2(\mu_1\mu_2-\lambda_1\mu_2-\lambda_2\mu_1)}\, W_2^{(2)}{}_{(\mathrm{RANDOM})}$$

and

$$W_2^{(2)}{}_{(\mathrm{RANDOM})} = \frac{2\mu_2\mu_1-2\lambda_1\mu_2}{2\mu_1\mu_2-2\lambda_1\mu_2-\lambda_2\mu_1}\, W_2^{(2)}{}_{(\mathrm{FIFO})}$$

$$2\mu_1\mu_2-2\lambda_1\mu_2-\lambda_2\mu_1 > 2\mu_1\mu_2-2\lambda_1\mu_2-2\lambda_2\mu_1$$

and $$2\mu_2\mu_1-2\lambda_1\mu_2 > 2\mu_1\mu_2-2\lambda_1\mu_2-\lambda_2\mu_1$$

$$\therefore \quad W_2^{(2)}{}_{(\mathrm{LIFO})} > W_2^{(2)}{}_{(\mathrm{RANDOM})} > W_2^{(2)}{}_{(\mathrm{FIFO})}$$

ie. The variance of the waiting time distribution for random order of service lies between the corresponding variances for 'last-in first-out' and 'first-in first-out' order of service.

This means that the delays become more variable as the queue discipline goes from FIFO to RANDOM to LIFO. Thus the proportion of waiting calls receiving quick service is increased but this is achieved at the cost of making other calls wait much longer.

CHAPTER VI

Multiple Server Systems

1. Introduction

The capacity of a system with a single control unit can of course be improved by increasing the service rate of the control. This is not always feasible so that another method that is used is to add similar control units which operate in parallel with the original control unit forming a multiple server system.

In many ways this latter procedure has advantages. The mean delay in queue is reduced because the queue size is reduced in comparison to the number in the system. The system can also be run at a higher occupancy level. As with the single control unit case, there are situations in which the incoming calls have to join a queue, so that it is advisable to examine the multiple server case.

The problems arising when a multiple server system is considered are far more difficult than the corresponding problems for the single server case. It is understandable therefore that when the additional complication of priorities is added the mathematical analysis becomes very complicated.

Cobham[11] considered a multiple non-preemptive priority system with exponential holding times and found the average waiting times for FIFO queue discipline. Davis[30] studied the same model with the exception that the priority classes all had the same holding time distributions and found an explicit formula for the waiting time distribution. The work in this chapter is an extension of Davis's work to the case where the priority system considered is of the preemptive resume type. The basic model is described in section 2, and the analysis is carried out in section 4. The results are analysed in section 5.

2. The Model

Calls with different priorities are arriving at a facility consisting of a number of identical control units operating in parallel. The calls arrive as a Poisson process with inter-arrival - time distribution $\lambda e^{-\lambda x}dx,\ x>0$. The calls are served in order of priority and for each priority with the first-in first-out discipline.

Calls of lower numbered priority are served first and the priority discipline is of the preemptive type. This discipline has been designated as case (ii) in previous chapters.

The holding times of calls are considered to be mutually independent random variables and are independent of the inter-arrival times.

3. Notation

Denote by $H_p(t)$ $(p=1,2,\ldots,k)$ the probability that the holding time of a call with priority p is $\leqslant t$.

Let
$$H_p(t) = 1-e^{-\mu t} \tag{6.1}$$

The arrivals with priority p follow a Poisson process with intensity λ_p.

$$\sigma_p = \sum_{m=1}^{p} \lambda_m$$

Let

$W_p(t)$ = the queueing time distribution for a priority p type call ie. the probability that a priority p type call will wait at the most t.

c = the number of identical parallel servers.

$P(n,p,t)$ = the probability that all the c servers are busy and there are n calls with priority number p or less on queue when a call of priority p arrives at time t.

$P_s(t)$ = the probability that there are s calls of any priority in the system at time t. $s \leqslant c$.

$P_{s,p}(t)$ = the probability that there are s calls of priority number p or less in the system. $s \leqslant c$.

$B(c,p)(t)$ = the busy period distribution for a c server system.

4. Analysis

In order to find the waiting time distribution we first need to find the state probabilities of the size of the queue.

Let us make the following assumptions:

1) The probability of a new call originating in the interval $(t,t+dt)$ $\simeq \lambda dt$. The probability of two or more calls originating in $(t,t+dt)$ is negligibly small.

2) The probability of a call terminating in $(t,t+dt)$ is $c\mu dt$. The probability of two or more terminating is negligibly small.

The state equations for $n \geq 1$ are fairly straightforward.

<u>For $n \geq 2$</u>

$$P(n,p,t+dt) = c\mu dt\ P(n+1,p,t) + \sigma_p dt\ P(n-1,p,t) + P(n,p,t)(1-c\mu dt-\sigma_p dt) \tag{6.2}$$

<u>For $n = 1$</u>

$$P(1,p,t+dt) = c\mu dt\ P(2,p,t) + \sigma_p dt\ P_{c,p}(t) + P(1,p,t)(1-c\mu dt-\sigma_p dt) \tag{6.3}$$

<u>For $n = 0$</u>

The system will be in state $(0,p)$ at time $t=t_0$ if and only if at least one of the following mutually exclusive events has occurred.

(i) State at time $t=t_0-dt$ was $(0,p)$; no call terminated and no call of priority $\leq p$ arrived since t_0-dt.

4. Analysis

In order to find the waiting time distribution we first need to find the state probabilities of the size of the queue.

Let us make the following assumptions:

1) The probability of a new call originating in the interval $(t,t+dt)$ $\simeq \lambda dt$. The probability of two or more calls originating in $(t,t+dt)$ is negligibly small.

2) The probability of a call terminating in $(t,t+dt)$ is $c\mu dt$. The probability of two or more terminating is negligibly small.

The state equations for $n \geq 1$ are fairly straightforward.

For $n \geq 2$

$$P(n,p,t+dt) = c\mu dt\ P(n+1,p,t) + \sigma_p dt\ P(n-1,p,t) + P(n,p,t)(1-c\mu dt-\sigma_p dt) \tag{6.2}$$

For $n = 1$

$$P(1,p,t+dt) = c\mu dt\ P(2,p,t) + \sigma_p dt\ P_{c,p}(t) + P(1,p,t)(1-c\mu dt-\sigma_p dt) \tag{6.3}$$

For $n = 0$

The system will be in state $(0,p)$ at time $t=t_0$ if and only if at least one of the following mutually exclusive events has occurred.

(i) State at time $t=t_0-dt$ was $(0,p)$; no call terminated and no call of priority $\leq p$ arrived since t_0-dt.

(ii) State at time $t=t_0-dt$ was $(1,p)$; a call terminated but no call of priority $\leq p$ arrived since t_0-dt.

(iii) There were $(c-1)$ calls of any priority in the system at time $t=t_0-dt$; a call (of any priority) arrived.

(iv) There were only calls with priority $> p$ in the queue and the c servers were busy at time $t=t_0-dt$; a call terminated.

(v) There were no calls of priority $\leq p$ in the queue and there was at least one call of priority $> p$ amongst the c calls being served at time $t=t_0-dt$; a call with priority $\leq p$ arrived.

Thus

$$\begin{aligned} P(0,p,t+dt) = {} & P(0,p,t)(1-c\mu dt-\sigma_p dt) + P(1,p,t)c\mu dt \\ & + P_{c-1}(t)\ \lambda dt + c\mu dt\ (P(0,p,t)-P_c(t)) \\ & + \sigma_p dt\ (P(0,p,t) - P_{c,p}(t)) \end{aligned} \tag{6.4}$$

Dividing both sides of each equation by dt and passing to the limit when dt tends to zero the Kolmorgorov difference differential equations are obtained.

<u>For $n \geq 2$</u>

$$\begin{aligned} \frac{dP}{dt}(n,p,t) = {} & c\mu\ P(n+1,p,t) + \sigma_p\ P(n-1,p,t) \\ & - (c\mu + \sigma_p)P(n,p,t) \end{aligned} \tag{6.5}$$

<u>For $n = 1$</u>

$$\begin{aligned} \frac{dP}{dt}(1,p,t) = {} & c\mu\ P(2,p,t) + \sigma_p\ P_{c,p}(t) \\ & - (c\mu + \sigma_p)P(1,p,t) \end{aligned} \tag{6.6}$$

For n = 0

$$\frac{dP}{dt}(0,p,t) = c\mu\, P(1,p,t) + \lambda P_{c-1}(t) - c\mu\, P_c(t) - \sigma_p P_{c,p}(t) \tag{6.7}$$

The steady state (or stationary equilibrium) equations are then:

For n ≥ 2

$$(c\mu + \sigma_p)P(n,p) = c\mu\, P(n+1,p) + \sigma_p\, P(n-1,p) \tag{6.8}$$

For n = 1

$$(c\mu + \sigma_p)P(1,p) = c\mu\, P(2,p) + \sigma_p\, P_{c,p} \tag{6.9}$$

For n = 0

$$c\mu\, P(1,p) + \lambda P_{c-1} = c\mu\, P_c + \sigma_p\, P_{c,p} \tag{6.10}$$

From [1]

$$\lambda P_{c-1} = c\mu\, P_c \tag{6.11}$$

$\therefore$ substituting in (6.10)

$$c\mu\, P(1,p) = \sigma_p\, P_{c,p} \tag{6.12}$$

Now the solution to (6.8) can be put in the form:

$$P(n,p) = R\left(\frac{\sigma_p}{c\mu}\right)^n \tag{6.13}$$

where R is a constant.

Substituting in (6.8) when n = 2 gives:

$$P(1,p) = R\,\sigma_p/c\mu \tag{6.14}$$

ie.

$$P(n,p) = R\left(\frac{\sigma_p}{c\mu}\right)^n \qquad \text{for } n \geq 1 \tag{6.15}$$

and from (6.12)

$$R = P_{c,p} \tag{6.16}$$

Now $P_{c,p} + \sum_{n=1}^{\infty} P(n,p)$ = the probability of delay in a non priority c server system with arrival rate σ_p

$$= PD_p$$

$$= \left(1+\left((1-\sigma_p/c\mu)c!/(\sigma_p/\mu)^c\right)\sum_{j=0}^{c-1} (\sigma_p/\mu)^j/j!\right)^{-1} \tag{6.17}$$

ie. from (6.15), (6.16) and (6.17)

$$P_{c,p} \frac{c\mu}{(c\mu-\sigma_p)} = PD_p \tag{6.18}$$

and

$$R = \frac{(c\mu-\sigma_p)PD_p}{c\mu} \tag{6.19}$$

Waiting Time Distribution

The method used in this section is similar to that used by Davis[30].

If there are less than c calls with priority number $\leq$ p in the system when the test call of priority p type arrives then the call will go straight into service and the waiting time is zero. If there are exactly c calls of priority number $\leq p$ in the system then the test call will be delayed. Its waiting time is distributed as the busy period of a c server system serving only calls with priority number $< p$. Riordan[9] showed that this busy period is distributed the same as the busy period of a single server system, serving only calls with priority number $< p$ and with servicing rate $c\mu$.

If there are n calls with priority number $\leq p$ on queue when the test call arrives then[14] the distribution of the waiting time is the (n+1) fold convolution of the busy period of a c server system servicing only the calls with priority $\leq p$.

The waiting time distribution can thus be described by:

$$W_p(t) = 1-PD_p + P_{c,p}B(c,p)(t) + \sum_{n=1}^{\infty} P(n,p)\,{}_{(n+1)}B(c,p)(t) \qquad (6.20)$$

where ${}_{(n+1)}B(c,p)(t)$ is the (n+1) fold convolution of $B(c,p)(t)$ with itself.

$\therefore$ Taking the Laplace-Stieltjes Transform

$$W_p^*(s) = 1-PD_p + R\sum_{n=0}^{\infty}\left(\frac{\sigma_p}{c}\right)^n (B(c,p)(s))^{n+1} \qquad (6.21)$$

$$\text{where } B(c,p)^*(s) = s + \sigma_{p-1} + c\mu - \sqrt{\frac{(s+\sigma_{p-1}+c\mu)^2-4c\mu\sigma_{p-1}}{2\sigma_{p-1}}} \qquad (6.22)$$

$\therefore$ From (6.15), (6.16), (6.18) and (6.21)

$$W_p^*(s) = 1-PD_p + PD_p\,\frac{(c\mu-\sigma_p)B(c,p)^*(s)}{(c\mu-\sigma_p B(c,p)^*(s))} \qquad (6.23)$$

i.e.

$$W_p^*(s) = 1-PD_p + \frac{2PD_p(c\mu-\sigma_p)}{Q_{c,p}(s)} \qquad (6.24)$$

where

$$Q_{c,p}(s) = s+\sigma_{p-1}+c\mu-2\sigma_p + \sqrt{(s+\sigma_{p-1}+c\mu)^2-4c\mu\sigma_{p-1}}$$

The inversion of equations of the type (6.24) has been carried out by a number of people [15,31].

The result is as follows:

$$W_p(t) = 1 - \frac{PD_p(\rho_p^2-\rho_{p-1})e^{-\alpha t}}{\rho_p(\rho_p-\rho_{p-1})} - \frac{PD_p(1-\rho_p)}{2\Pi\rho_p}\int_{\alpha_1}^{\alpha_2} f(r)dr \qquad (6.25)$$

where $\rho_p^2 \geq \rho_{p-1}$

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

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

$$\alpha = \left((\rho-\rho_1)(1-\rho)/\rho\right)c\mu$$

and

$$f(r) = e^{-rt}\sqrt{(\alpha_2-r)(r-\alpha_1)/r(r-\alpha)}$$

5. RESULTS

Consider the case where there are only two priority classes. The average time spent in the queue for each priority class can be found from (6.24) by differentiating w.r.t. s and putting s=0.

i.e.

$$W_1^{(1)} = \frac{PD_1\ c\mu}{(c\mu-\lambda_1)} \tag{6.26}$$

$$W_2^{(2)} = \frac{PD_2\ c\mu}{(c\mu-\lambda)(c\mu-\lambda_1)} \tag{6.27}$$

$\therefore$ The average overall queueing time W_a is given by:

$$W_a = \frac{\lambda_1 W_1^{(1)} + \lambda_2 W_2^{(1)}}{\lambda} \tag{6.28}$$

i.e.

$$W_a = \frac{c\mu(\lambda_2 PD_2 + \lambda_1 PD_1(1-\lambda/c\mu)}{\lambda(c\mu-\lambda_1)(c\mu-\lambda)} \tag{6.29}$$

The equivalent value W_b for the case when the discipline is non preemptive can be found in a similar manner from[30].

$$W_b = \frac{PD_2}{(c\mu-\lambda)} \tag{6.30}$$

Fig. 6.1 shows a comparison of (6.29) and (6.30) for various values of c. W_b is given by $\lambda_1/\lambda = 1$. The curves with $0 \leq \lambda_1/\lambda < .5$ and $.5 < \lambda_1/\lambda < 1$

all fall between the two curves shown.

It can be seen from Fig. 6.1 that for a given value of $(\lambda/\mu c)$, increasing c decreases the average delay. The extra reduction due to changing the discipline from non-preemptive to preemptive is very small however. Also as the average queue size is proportional to the average delay there is a corresponding small reduction in queue size.

Another way to look at Fig. 6.1 is that if it is assumed that the cost of service per server is proportional to its rate of service, increasing the rate of a single server by a factor c will cost the same amount as increasing the number of servers to c, keeping the rate of each unchanged.

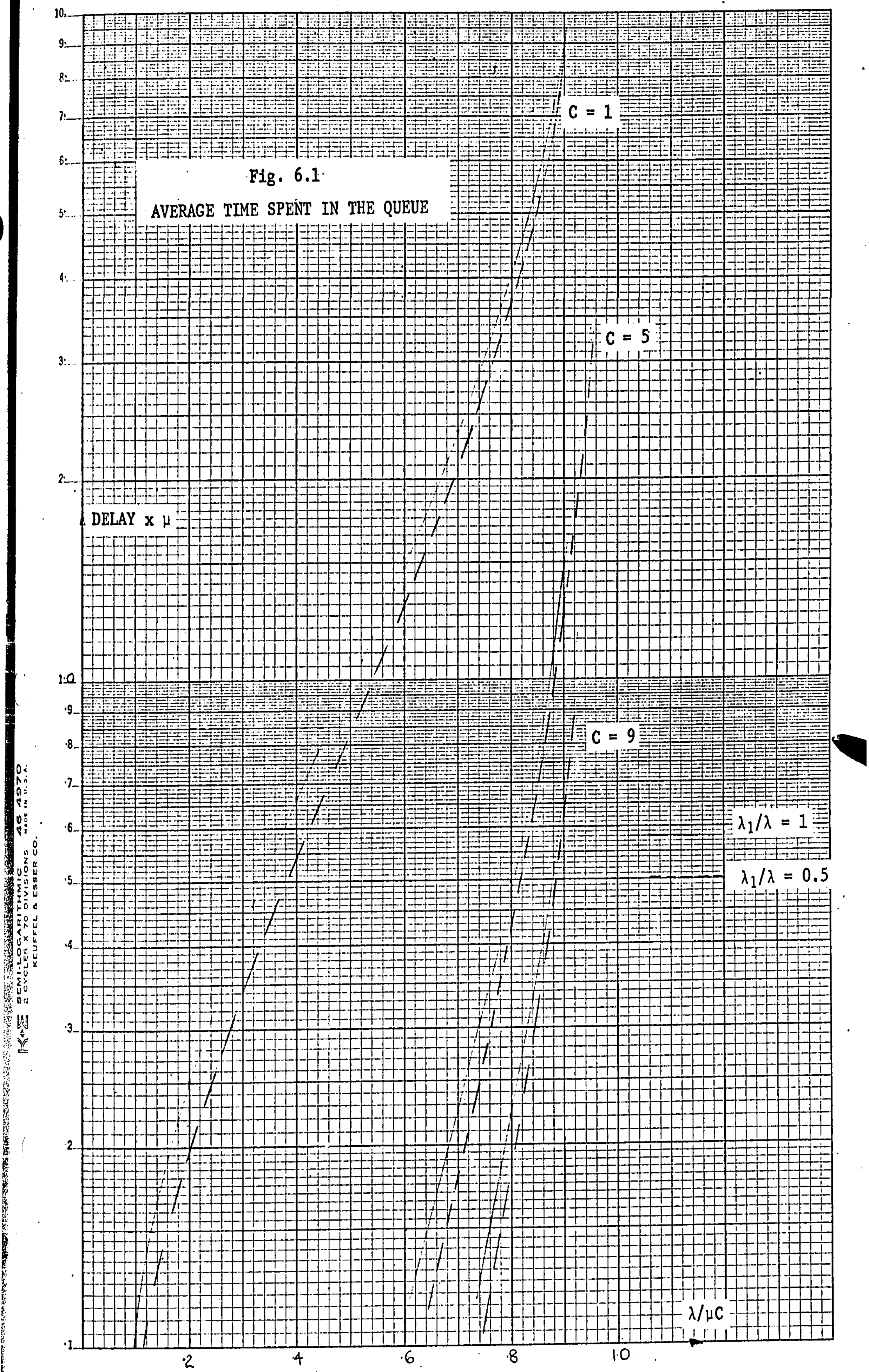


Fig. 6.1
AVERAGE TIME SPENT IN THE QUEUE
C = 1
C = 5
C = 9
DELAY x μ
λ1/λ = 1
λ1/λ = 0.5
λ/μC
10.
9.
8.
7.
6.
5.
4.
3.
2.
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
.2
.4
.6
.8
1.0
SEMI-LOGARITHMIC 46 4970
2 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.

CHAPTER VII

CONCLUSION

1. Summary

In studying the type of systems that have been modeled the view has been taken that substantially more can be learned through the analysis of mathematical models than is presently the case with priority service systems. In this respect a number of observations have been made regarding the performance of priority service disciplines and the general nature of parallel processing within the framework of these disciplines.

An efficient assignment of priorities to input streams to stochastic control systems can be accomplished by means of the analytical processes developed in this work. The service disciplines discussed, offer a number of techniques by which the waiting times of the different classes of inputs can be manipulated or adjusted to meet a set of operational requirements. The models presented are distinguished by the fact that the priority assignments are known beforehand and are thus independent of the dynamic operating characteristics of the system.

A number of examples have been given to show practical applications of the general solutions.

2. Future Investigations

There are a number of areas of work which suggest themselves for future investigations. Relaxation of some of the assumptions is perhaps the most obvious area. Under this category would fall the following:

(a) Introducing, into the models studied, the effects of limited sources.

(b) Relaxing the assumption of a Poisson input.

This latter area would introduce a number of problems. If the Poisson assumption is relaxed, then one finds a non-Markovian stochastic process requiring two or more supplementary variables to make it Markovian. This difficulty is essentially similar to the problem encountered when considering an N server system with a general holding time distribution. One is confronted with the problem of solving a differential difference equation in N variables. In the past the effect of a general input in a priority system has been studied by considering one type of arrival process as Erlangian or a weighted mixture of Erlangian distributions. The mathematics involved however, became exceedingly complicated so that possibly research should be carried out to find simpler methods for studying non-Markovian processes which require more than one supplementary variable before looking at a priority system with a general input.

Other areas which suggest themselves as extension of this work are:

(c) Extending the work on the LIFO queue discipline to a multi-server system.

(d) Same as (c) but with random order of service.

All of the discussions involved in this thesis have concerned the general question of the appropriate level of service to provide and how best to provide that service. The economics of the situation have not been considered unless one considers that the waiting time is proportional to the cost of service. Some alternative cost models have been produced but very few in the field of priority systems.

More analytical effort is obviously still needed to understand the complications of the systems discussed. It is hoped that the present work will help to stimulate effort in this direction.

REFERENCES

1. Syski, R.;
"Introduction to Congestion Theory in Telephone Systems", Oliver and Boyd, London (1960).

2. Jacobaeus, C.;
"A Study on Congestion in Link Systems", Ericsson Technics 51 (1950), 1-68.

3. Lee, C. Y.;
"Analysis of Switching Networks", Bell System Tech. J. 34 (1955) 1287-1315.

4. Berkeley, G. S.;
"Traffic and Trunking Principles in Automatic Telephony", Ernest Benn, Ltd., London 1934 (reprinted, 1949).

5. Lee, L. and Brzozowski, J. A.;
"An Approximate Method for Computing Blocking Probability in Switching Networks", IEEE, Com-14, No. 2 (1966) 85-93.

6. Saaty, T. L.;
"Elements of Queueing Theory with Applications", McGraw-Hill Book Company, Inc. (1961).

7. Takacs, L.;

"Introduction to the Theory of Queues", New York, Oxford Univ. Press (1962).

8. Cox, D. R., and Smith, W. L.;

"Queues", London, Methuen, New York, Wiley (1961).

9. Riordan, J.;

"Stochastic Service Systems", Wiley, New York (1962).

10. Delbrouck, L. E. N., and Lee, L.;

"The Anatomy of Waiting Time Distributions in Single Server Systems", IEEE Com-15, No. 1 (1967) 11-17.

11. Cobham, A.;

"Priority Assignment in Waiting Line Problems", Operations Research 2, (1954) 70-76.

12. ——————

"Priority Assignment - A Correction", Operations Research 3, (1955), 375-586.

13. Holley, J. L.;

"Waiting Line Subject to Priorities", Operations Research 3, (1954) 341-343.

14. Dressin, S. A., and Reich, E.;

"Priority Assignment on a Waiting Line", Quarterly of Applied Math. Vol. 15 (1957) 208-211.

15. Kesten, H., and Runnenberg, J. Th.;

"Priority in Waiting Line Problems", I and II, Indag. Math. 19, (1957) 312-336.

16. Morse, P. M.;

"Queues, Inventories and Maintenance", John Wiley and Sons, New York, 1958.

17. White, H., and Christie, L. S.;

"Queueing with Preemptive Priorities or with Breakdown", Operations Research 6, (1958) 79-95.

18. Stephan, F.;

"Two Queues Under Preemptive Priority with Poisson Arrival and Service Rates", Operations Research 6, (1958) 399-418.

19. Kendall, D. G.;

"Some Problems in the Theory of Queues", J. Roy. Stat. Soc., Series B, 2 (1951) 151-173.

20. Miller, R. G.;

"Priority Queues", Ann. Math. Stat. 31, (1960) 86-103.

21. Gaver, D. P.;

"A Waiting Line with Interrupted Service, Including Priorities", J. Roy. Stat. Soc. B, 24 (1962) 73-90.

22. Avi-Itzhak, B., and Naor, P.;

"On a Problem of Preemptive Priority Queueing", Operations Research 9, (1961) 664-672.

23. Takacs, L.;

"Priority Queues", Operations Research 12, (1964) 63-74.

24. ―――――

"The Stochastic Law of the Busy Period for a Single Server Queue with Poisson Input", J. Math. Anal. and Appl. 6, (1963) 33-42.

25. Jordan, C. H.;

"Calculus of Finite Differences", Budapest (1939) Second ed., Chelsea, New York (1947).

26. Whittaker, E. T., and Watson, G. N.;

"A Course of Modern Analysis", Cambridge University Press, Cambridge 1952.

27. Wishart, D. M. G.;

"Queueing Systems in Which the Discipline is Last-Come, First-Served", Operations Research 8 (1960) 591-599.

28. Takacs, L.;

"Delay Distributions for One Line, With Poisson Input, General Holding Times, and Various Orders of Service", B.S.T.J. 42, (1963) 487-503.

29. Durr, L.;

"A Single Server Priority Queueing System With General Holding Times, Poisson Input, and Reverse-Order-of-Arrival Queueing Discipline", Operations Research 17 (1969), 351-358.

30. Davis, R. H.;

"Waiting Time Distribution of a Multi-Server Priority Queueing System", Operations Research, 14 (1966), 133-136.

31. Nee, D.;

"Application of Queueing Theory to Information Systems Design", Stanford Research Institute Report.

VITA

Name: Laurence Durr

Born: April 14, 1939, England

Education:

Primary: England

Secondary: England

University: Cambridge University, England

(a) 1961 B.A. (Mechanical Sciences Tripos)

(b) 1965 M.A.