76-27,226

SHIMSHAK, Daniel G., 1949-
A STUDY OF QUEUES IN SERIES.

City University of New York, Ph.D., 1976
Operations Research

Xerox University Microfilms, Ann Arbor, Michigan 48106

A STUDY OF QUEUES IN SERIES

by

DANIEL G. SHIMSHAK

A dissertation submitted to the Graduate Faculty in Business in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1976

This manuscript has been read and accepted for the Graduate Faculty in Business in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

7/15/76
date

Georghios C. Sphicas
Chairman of Examining Committee

July 15 1976
date

Arthur L. Litman
Executive Officer

Lloyd Rosenberg

William Amadio
Supervisory Committee

The City University of New York

Abstract

A STUDY OF QUEUES IN SERIES

by

Daniel G. Shimshak

Adviser: Professor Georghios P. Sphicas

Two aspects are considered in this study of series queueing systems. The first deals with determining an optimal sequence of service stations in a series system. Optimality is defined in terms of the total time spent waiting for service. Sequences are compared on the basis of the moments of their steady state total waiting time. In addition, the rules for first and second degree stochastic dominance are applied which allow comparison of sequences on the basis of their waiting time distributions. The second aspect under study is the determination of single station queueing systems which are isomorphic with the series queueing system. Here the distribution functions of certain output characteristics are equivalent for both systems. Isomorphic queues of the original system allow for further study to be made that would otherwise not be possible on the complex series system.

Analytic results in the sequencing of service stations in tandem queues had been limited to stations

with constant or exponential service times. This study extended the investigation to service distributions with varying degrees of statistical regularity given by the family of Erlang distributions with parameter k. The model employed consists of two single channel service stations with Poisson arrivals to the first station. Only a first-in, first-out service discipline is considered and both possible ordered arrangements of the service stations in the system are feasible.

A series of exploratory simulation experiments, programmed in GPSS, is used to isolate the critical factors in determining optimality of sequences. These are used in the analytical derivation of an indifference equation for the mean waiting time between the two feasible sequences of the service stations. This gives a range of parameters where the mean waiting times in the two sequences are equivalent. By using some statistical techniques, this relationship is extended so that it can predict which sequence is optimal on the basis of first or second degree stochastic dominance. Validation is accomplished by simulating a number of systems and comparing the waiting time distribution functions for each sequence. The relationship is shown to be a good predictor and useful in the study and design of systems of servers in series.

The two server model is also used in the study of isomorphism. This work attempts to find single server

queues whose waiting time distribution function is statistically equivalent to the total waiting time distribution function of the series system. Characteristics of the series queueing system are gathered through simulation. Two methods of determining isomorphs are used. One is based on fitting the waiting time distribution of the series system in order to estimate the parameters of the isomorph. The second is a method of fitting moments of the total waiting time in the series system. Waiting time characteristics of the isomorphs found by both methods are compared to the actual series system. Results are limited and suggest that the isomorphs may be useful in estimating only portions of the waiting time distribution function of the series queueing system. However, the findings imply a vast potential and usefulness for isomorphism in the study of series queueing systems.

Further research in isomorphism will allow the n server queueing system to be reduced through the determination of single server isomorphic queues. Then methods for evaluating sequences of stations can be applied to study the system. Together, research in sequencing and isomorphism will enable the analysis of large systems of queues in series.

ACKNOWLEDGEMENTS

At the completion of this thesis, it is most appropriate to acknowledge those who contributed directly to providing the guidance and stimulation which made the research possible.

I would like to thank the members of my dissertation committee, each of whom displayed great interest in my work. These include Professors A. Ghosal, Lloyd Rosenberg, and Georghios Sphicas, Chairman, from Bernard M. Baruch College, and Professor William Amadio of Rider College. Special thanks to Professor Sphicas for all his time and assistance.

Finally I would like to thank my wife, Marcia, for her encouragement and understanding.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The Problem

Queueing theory involves the mathematical study of queues or waiting lines. Whenever the demand for service exceeds the capacity to provide that service, a waiting line forms. Since it is often impossible to predict when units will arrive for service and how much time will be needed to provide that service, decisions concerning the amount of service to provide are often difficult to make. Too much service involves excessive costs; not enough results in long waiting lines which is also costly. The objective is to attain a balance between the cost of service and the cost of waiting for that service. Queueing theory, by gathering data on the behavior of the waiting line, provides the vital information required for making this decision.

Queueing theory has been a very popular research subject for many years. One reason for the amount of interest in waiting line systems has been the awareness of a wide range of practical applications. Little of that effort has been spent on the area of queueing systems that deal with service facilities in series, in which the

departure process from one service station forms the arrival process at the next service station. Each customer, on arrival goes to the first station. Upon completion of service, he enters the second station. This continues until completion of service at the last station when he leaves the system. Such systems have been referred to as tandem queues or queues in series. Because departures from one station form arrivals into the next station, the analysis of these systems are much more complicated than traditional analysis for ordinary service stations.

Queueing problems concern the arrival of units, the operation of a service, and the departure of a completed unit. The elements of all queueing systems are the same, no matter how complex. That is, the specific features needed to define a particular system are statistical descriptions of the arrival and service processes, the number of servers, the size of the allowable queues, the arrangement of service facilities, and the service discipline. The possible combinations of these features allows for an unlimited number of specific systems for study.

Most of the work in the area of series queueing systems has consisted of mathematical analysis. However, results have been obtained only for problems that have severe assumptions associated with them. A relaxation of these assumptions results in problems that are not tractable through analytical techniques. There is no indication of major breakthroughs in the theory using mathematical analysis

alone, and significant extensions can only come from the use of simulation techniques. In this way, study of more than just the most simply structured problems can be made.

This work is not intended to be a contribution to the art of simulation. Instead an important and novel investigation of complex queueing systems is being conducted with the use of simulation techniques. This is the only method feasible for solving new and significant problems involving tandem queues and expanding the current knowledge of theoretical results.

The primary objective of this thesis is the development of mathematical techniques which can be used by the systems analyst and operations researcher for steady state analysis of tandem queueing systems of unlimited capacity. The problem is approached through simulation and analysis of the behavior of a queueing system consisting of two service facilities in series. Two aspects of this system will be dealt with. First, the sequence or order of these service stations will be studied for the purpose of prescribing an optimal order. A particular sequence is optimal if the time spent waiting for service is smaller than for any other sequence. It is most desirable to consider optimality in terms of waiting time. Decisions in problems involving queueing systems are usually based on the balance between the cost of service and cost of waiting. Since service costs are generally the same for any arrangement of the same service stations, waiting time is the

dominant criterion. Even if operating costs are different, valid comparisons can be made between sequences of service stations.

This study will look at a given arrival process and vary the nature of the service process. There are several factors which can affect the total waiting time distribution in any series queueing system. The intention is to find a range of service distribution parameters within which one sequence is optimal. Sequences of queueing systems will be compared for optimality of the basis of their steady state total waiting time moments. In addition, the rules of stochastic dominance of Hadar and Russell[1] will be considered which allow comparison of systems on the basis of their waiting time distributions. An application of stochastic dominance rules in evaluating single server queues on the basis of waiting times can be found in Rolski and Stoyan.[2]

Secondly, a means of mathematically determining a single station queueing system which is isomorphic with the series queueing system will be developed. Here the distribution functions of certain output characteristics are equivalent for both systems. Isomorphic queues of the

---

[1]Josef Hadar and William R. Russell, "Rules for Ordering Uncertain Prospects," American Economic Review 59 (March 1969):25-34.

[2]Tomasz Rolski and Dietrich Stoyan, "On the Comparison of Waiting Times in GI/G/1 Queues," Operations Research 24 (January-February 1976):197-200.

original system will allow for further study to be made that would otherwise not be possible on the complex system. This work is based upon previous research presented by Ghosal[3,4] on isomorphism in queueing processes.

The findings of this study can prove significant in extending the current knowledge in the theory of tandem queues as well as having practical application to important problems. A systems manager must always be conscious of the design and operation of the system. Clearly, determining the optimal sequence will allow the designer to rearrange service stations to reduce over-all customer delay. The potential exists to make better use of existing facilities, alter server busy patterns, control the buildup of items in intermediate queues, and other advantages. The net result is a reduction in time spent waiting and greater efficiency throughout the system.

Further research into the behavior of the output in the complex system of queues in series can be conducted on the isomorphic single station queues. Mathematical analysis of series queueing systems has been extremely limited. To the system manager, it is necessary to study the system at hand in order to evaluate proposed changes. If such a system is a series of queues, a very convenient manner of

---

[3]A. Ghosal, "Some Problems in Applied Cybernetics," SCIMA 2 (1973):35-50.

[4]A. Ghosal at the International Conference on Stochastic Processes, "Isomorphic Queueing Systems," University of Maryland, 1975. (Mimeographed.)

study is on a simpler isomorphic system. Isomorphism will allow the n server queueing system to be reduced through the determination of single server isomorphic queues. Then methods for evaluating sequences of stations can be applied. The analysis of complex systems now becomes a possibility and the use of sequencing and isomorphism in queueing processes becomes an important part of the work on large systems.

Series queueing systems are rather common in practice. R. R. P. Jackson[5] noted some industries in which activity takes place in several successive but distinct phases. An overhaul procedure involves five stages of activity including stripping, detailed examination, repair, assembly, and testing. Similar situations are found in stores and offices where parts of the service are performed at different counters, in telephone networks, cafeteria serving lines, urban automobile traffic, and school registration processes. Avi-Itzhak and Yadin[6] were motivated by the inspection system of a vehicle's mechanical condition. The vehicle undergoes inspection through three stations with no intermediate queues. One station checks front wheels and the steering system, a second the mechanical condition

---

[5]R. R. P. Jackson, Queueing Systems with Phase Type Service," Operational Research Quarterly 5 (December 1954):109-20.

[6]B. Avi-Itzhak and M. Yadin, "A Sequence of Two Servers with No Intermediate Queue," Management Science 11 (March 1965):565-71.

of the transmission and light system, and the third station inspects the braking system. There are many quality control systems with a sequence of inspection stations of this type. Cycling arose in Koenigsberg's[7] study of conventional mechanized coal mining operations and he approached the design of these operations using queueing theory.

In the context of production, systems of queues in series are models of many production line systems. This idea was first suggested by Richman and Elmaghraby[8] and Koenigsberg.[9] An entire area of research is devoted to the topic of assembly line balancing. Here all stations operate continuously when the line is running. The movement of parts through the line may be "paced" in that the flow of parts from station to station is geared to the slowest operation station in the line. Although the line may be designed for an inventory of parts between stations, this inventory is not expected to fluctuate to any large degree. This situation often results in an under-utilization of servers on the line.

It is possible to study "unpaced" production lines in which the operation time at each station is a random

---

[7]Ernest Koenigsberg, "Cyclic Queues," Operational Research Quarterly 9 (March 1958):22-35.

[8]Eugene Richman and Salah Elmaghraby, "The Design of In-Process Storage Facilities," Journal of Industrial Engineering 8 (January-February 1957):7-9.

[9]Ernest Koenigsberg, "Production Lines and Internal Storage--A Review," Management Science 5 (July 1959):410-33.

variable. The most useful approach to this problem is to treat the line as though it were a system of queues with service stations in series. The inventory between stations would be the length of the queue that forms ahead of each station.

Queueing theory is useful in the complete design of production lines, including the following decisions:

1) The number of stations to employ in the line,
2) The order in which to place the stations, and
3) The amount of interstation storage capacity to provide for inventory buildup.

Various aspects of the design problem have been studied by Barten,[10] Goode and Saltzman,[11] and Freeman.[12]

Dam reservoirs and inventory systems have often been thought of as analogues of queueing problems. In these models both the input and release of output are random variables. The major probability problem in such storage systems is to determine the distribution of the storage level. In dams this is the dam level, in inventory it is the stock level, and in queueing systems it is the waiting

---

[10]Kenneth Barten, "A Queueing Simulator for Determining Optimum Inventory Levels in a Sequential Process," Journal of Industrial Engineering 13 (July-August 1962):245-52.

[11]Henry P. Goode and S. Saltzman, "Estimating Inventory Limits in a Station Grouped Production Line," Journal of Industrial Engineering 13 (November-December 1962):484-90.

[12]Michael C. Freeman, "The Effects of Freakdowns and Interstage Storage on Production Line Capacity," Journal of Industrial Engineering 15 (July-August 1964):194-200.

time or queue size. Ghosal[13] considered two infinite dams in series where the release from the first dam goes as input into the second dam. Finding the distribution of the dam level is quite a lengthy task, and queueing theory serves as a means of viewing the problem. Simulation methods are suggested for solving the more complicated problems.

Review of the Literature

The earliest research with tandem or series queueing systems was performed by R. R. P. Jackson.[14] For a system with Poisson arrivals and exponential service time, he found that the queue length of the service stations are independent random variables in the steady state. He later extended this work to include service stations composed of a number of identical servers in parallel.[15] J. R. Jackson[16] showed that, for this same Poisson-exponential system, the steady state joint probability distribution of customers waiting in the system is equal to the product of the probabilities for each individual Poisson-exponential service station.

---

[13] A. Ghosal, Some Aspects of Queueing and Storage Systems, Lecture Notes in Operations Research and Mathematical Systems, vol. 23 (Heidelberg and New York: Springer-Verlag, 1970), pp. 72-3.

[14] R. R. P. Jackson, "Queueing Systems."

[15] R. R. P. Jackson, "Random Queueing Processes with Phase Type Service," Journal of the Royal Statistical Society, ser. B, 18 (1956):129-32.

[16] James R. Jackson, "Networks of Waiting Lines," Operations Research 5 (August 1957):518-21.

Nelson[17] went on to derive the joint waiting time distribution for this series of service stations.

Some important results were reported by Burke, Finch, and Reich dealing with the steady state departure process of a Poisson input-exponential service system. Burke[18] showed that for each service station, the steady state output process, and therefore the input process to the next station, is also Poisson. This proof was supplemented by Finch[19] who found Burke's Poisson departure to hold only when infinite queue lengths are allowed between stations. In addition he proved that successive interdeparture intervals are independent in the steady state only in the case of exponential service time and unbounded queue lengths. For general service distributions, other considerations must be given to the input of each station in the series. Working on the same system, Reich[20] proved that the durations of time spent by a customer in successive stations are independent

---

[17]Rosser T. Nelson, "Waiting-Time Distributions for Application to a Series of Service Centers," Operations Research 6 (November-December 1958):856-62.

[18]Paul J. Burke, "The Output of a Queueing System," Operations Research 6 (December 1956):699-704.

[19]P. D. Finch, "The Output Process of the Queueing System M/G/1," Journal of the Royal Statistical Society, ser. B, 21 (1959):375-80.

[20]Edgar Reich, "Waiting Times When Queues Are in Tandem," Annals of Mathematical Statistics 28 (September 1957):768-73.

in single server tandem queues. Burke[21,22] obtained some additional results to show that when multiple servers in parallel are permitted at each service station, some of the waiting times are dependent.

Reich stated that if waiting times are defined as only the time spent in queue and not including the service times, then the question of independence of these quantities is an open problem. Burke[23] studied a system with two exponential service stations in series and a Poisson arrival pattern, and found, rather remarkably, that the steady state waiting times in each phase are dependent. These results hold for all queue disciplines that do not allow defections or pre-emptions.

Most of the analytical work done with series queueing systems has been limited to Poisson-exponential networks. However, Ghosal[24] found results under approximation for a system of one exponential service station and a second station with Erlang distribution of parameter 2. The lack of theory for any but the simplest classical queueing systems

---

[21]Paul J. Burke, "The Input Process of a Stationary M/M/s Queueing System" Annals of Mathematical Statistics 39 (August 1968):1144-52.

[22]Paul J. Burke, "The Dependence of Sojourn Times in Tandem M/M/s Queues," Operations Research 17 (July-August 1969):754-55.

[23]Paul J. Burke, "The Dependence of Delays in Tandem Queues," Annals of Mathematical Statistics 35 (June 1964): 874-75.

[24]A. Ghosal, "Queues in Series," Journal of the Royal Statistical Society, ser. B, 24 (1962):359-63.

suggests a simulation approach toward further study. Some work was performed by Nelson[25] who simulated a two server network model in order to estimate steady state queue statistics. He considered the exponential, Erlang with parameter 2, and constant distributions as arrival and service processes and conducted experiments for possible combinations of these. Nelson examined both a tandem flow system and job-shop type network system.

A useful contribution was made by Fraker[26] who developed an approximate formula for the steady state mean waiting time in a system of single server, infinite capacity queues with Erlang service. This was developed through observation and analysis of simulations of tandem queueing systems. Rosenshine and Chandra[27] extended this in considering multiple servers at each stage of the system.

The results of Finch[28] indicate that departure intervals are statistically dependent random variables for any system other than Poisson arrival and exponential service

---

[25]Rosser T. Nelson, "A Simulation Study and Analysis of a Two Station, Waiting-Line Network Model," (Ph.D. dissertation, UCLA, 1965).

[26]John R. Fraker, "Approximate Techniques for the Analysis of Tandem Queueing Systems," (Ph.D. dissertation, Clemson University, 1971).

[27]Matthew Rosenshine and M. Jeya Chandra, "Approximate Solutions for Some Two-Stage Tandem Queues, Part 1: Individual Arrivals at the Second Stage," Operations Research 23 (November-December 1975):1155-66.

[28]Finch, "Output of M/G/1."

times. Since departures from one service station are arrivals at the next station, arrival intervals are also statistically dependent. The presence of statistical dependence in the service station arrival and departure processes will influence the queueing statistics of the individual service stations in the system. In addition, this feature violates one fundamental assumption of elementary queueing theory and complicates further analytic work on series queueing systems. In his research, Nelson[29] studied the effect of statistical dependence on equilibrium properties of the system using simulation.

The sequence of service stations in a series of queues was first investigated by Avi-Itzhak.[30] He derived some characteristics of a queueing system with finite queues in series and constant service times. It was proven that the time spent in the system, for any specified process of arrivals, is independent of the order of the servers and is independent of the allowable intermediate queue size. A problem called blocking was encountered whenever a server is occupied by a customer whose service at that server is already completed. This condition occurs when the customer cannot move to the next server in the sequence due to the presence of the preceding customer still being there. By

---

[29]Nelson, "Simulation of Two Stations."

[30]B. Avi-Itzhak, "A Sequence of Service Stations with Arbitrary Input and Regular Service Times," Management Science 11 (March 1965):553-64.

changing the order of the servers such that the one with the longest service time is first in the sequence will eliminate the blocking situation. Friedman,[31] in an independent investigation, found similar results.

As implied by the works of Reich[32] and J. R. Jackson,[33] the expected total time in a system with exponential servers is independent of the order of the stations. Avi-Itzhak and Yadin[34] studied the effect of intermediate queues between service stations in this system and found that by introducing these queues, not only will blocking be reduced, but in fact the time spent in the system will reduce.

Much of the previous work on the sequence of servers drops out as special cases of the results found by Tembe and Wolff.[35] Their main concern was to compare the total time in the system for customers under different orderings. Tembe and Wolff ordered two service stations to form two different queueing systems. They found a reduction in the time in the system by having the longer service time station perform first. If the shorter service is a constant, then

---

[31]Henry D. Friedman, "Reduction Methods for Tandem Queueing Systems," Operations Research 13 (January-February 1965):121-31.

[32]Reich, "Waiting Times in Tandem."

[33]James R. Jackson, "Networks of Lines."

[34]Avi-Itzhak and Yadin, "Sequence of Two Servers."

[35]Shantanu V. Tembe and Ronald W. Wolff, "The Optimal Order of Service in Tandem Queues," Operations Research 22 (July-August 1974):824-32.

the time in the two systems are equal and independent of order. This is what Friedman[36] found. Another finding showed a reduction in time in the system when one station, being of constant service time, is placed first. Though based upon some limiting assumptions, this significant study demonstrated the ability of sequencing queues in series to reduce the time spent in the system, a result which can prove extremely helpful in the design of systems.

In discussing areas of further research, Tembe and Wolff stated that their results were based upon stringent assumptions about the service distributions and are not dependent on the utilization factor at each station. In general this will not be true. Thus the area is still open for useful investigation of series queueing problems that will provide an insight into the more realistic situations that exist in system operations. This thesis intends to seek results and methods of analysis for predicting system behavior which may enhance the ability to attack the more complex systems. Both simulation and analytical techniques will be used to present new findings and extend current theoretical knowledge in this very important area of study.

The queueing model to be used is described in Chapter II, along with a presentation of the theoretical foundations and a statement of the hypotheses of this study. Chapter III discusses some problems in the simulation of

---

[36]Friedman, "Reduction Methods for Queues."

stochastic processes, presents the simulation program, and outlines the experimental design. Results of the experimental and analytical studies in the sequence of servers are dealt with in Chapter IV, and the work on isomorphic systems in Chapter V, including a review of previous research. The conclusions of this study are given in Chapter VI.

CHAPTER II

MODEL DEVELOPMENT AND THEORETICAL FOUNDATIONS

The Queueing Model

All waiting line systems have a large number of features which are unique to the specific system being modeled. However, it is most practical if a model is selected which includes certain fundamental properties descriptive of a broad class of systems. In this way, studying the model may produce results of direct application to a wide class of problems. The model selected is intended to represent a blend of both simplicity and detail and use current knowledge and research techniques as a foundation to expand into new areas.

The following assumptions apply to the model employed in this study:

1) Each station operates as a single server queue.
2) Each arrival requires service at all service stations in the system.
3) All possible ordered arrangements of service stations in the queueing system are feasible.
4) Only a first-in, first-out service discipline is considered.

5) Service and arrival rates are independent of the state of the system.

6) Service rates are independent of the arrival process.

7) Infinite queues are allowable throughout the system.

The model deals only with pre-specified series queueing systems with each arriving unit having to be serviced at each station in order. This study avoids the job-shop production system where an arrival from outside the system requires its first service at any of the stations with certain probabilities, then conditional probabilities describe the movement of the job through the system, station by station. While considering all possible arrangements of the service stations, any zoning constraints or precedence relationships are ignored. However, the results that follow from this study can apply to situations where such restrictions do exist. The assumption of infinite queues eliminates the problem of blocking at queue i when the waiting line at queue i+1 is full. One can study the system under the condition of infinite capacity queues in order to determine the maximum queue length that develops. Finite queues, though of interest, should be reserved for later study.

In working with a series queueing system, there are many descriptive features that could be included in the model. Some of these include labor and capital restrictions, service repair, set-up costs, inventory procedures, and

cost effectiveness measures for the system. However, these characteristics, most useful in evaluating the operation of a specific firm, would greatly complicate the model. The main purpose of this simulation is to investigate the sequence of service stations in series, according to a particular optimizing criterion, on a model which approximates the operations of a queueing system. All other factors regarding labor, set-up costs and quality of performance are considered the same for any arrangement of the same service stations. Waiting time has been chosen as the optimizing criterion. As compared to other system characteristics that are related to measures of effectiveness, waiting time and the costs involved in waiting are usually the bases for decisions in queueing problems. Clearly a reduction in total waiting time can only lead to a reduction in total costs in the long run. Since it is the operation of the system that is of most interest, the scope of the experimentation on the queueing model will be limited to studies of the system and its waiting time behavior for a variety of system parameters.

For practical reasons, in order to limit the dimension of the problem to a manageable degree, the model used in this study is limited to two stations in series. Each arrival is serviced by one station, then the second before leaving the system. Two arrangements exist; in sequence A, service 1 is performed first followed by service 2, and in sequence B, service 2 first then service 1. (See Fig. 1.)

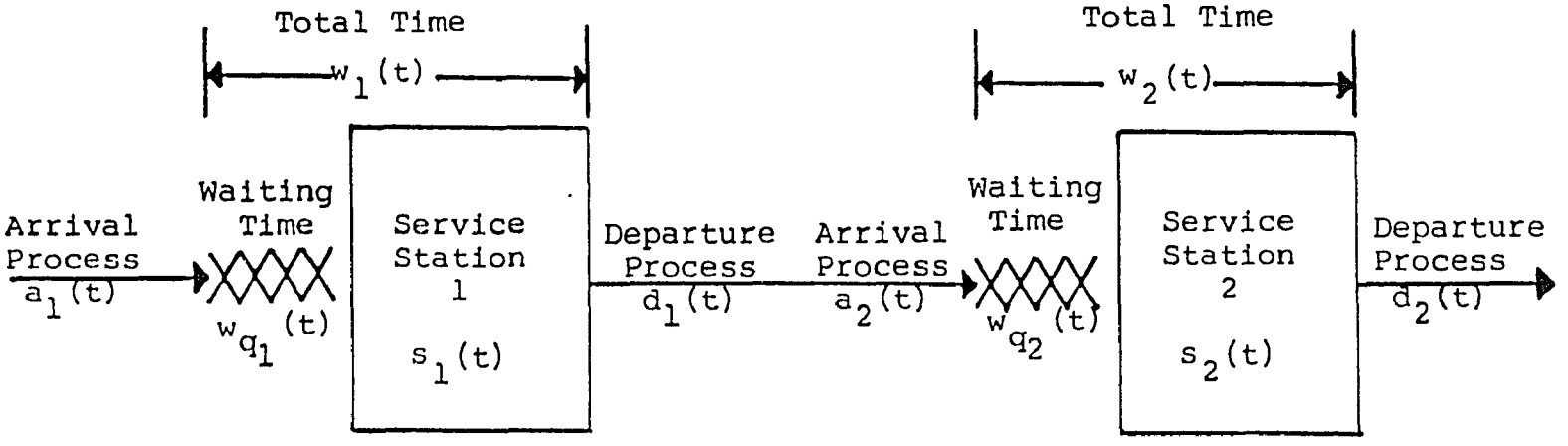


Fig. 1. Diagram of Queueing System under Sequence A

Definitions

The following is a list of symbols used in the analysis with their operational meanings. All refer to steady state characteristics.

$a_1(t)$, $a_2(t)$..... the probability density functions of time between arrivals to service station 1 and to service station 2

λ_1, λ_2..... the mean arrival rate to service station 1 and to service station 2

$\bar{a}_1$, $\bar{a}_2$..... the mean arrival intervals to service station 1 and service station 2

$\sigma^2_{a_1}$, $\sigma^2_{a_2}$..... the variance of the arrival intervals to service station 1 and service station 2

$s_1(t)$, $s_2(t)$..... The probability density functions of service times at service station 1 and at service station 2

μ_1, μ_2..... the mean service rate for service station 1 and for service station 2

$\bar{s}_1$, $\bar{s}_2$..... the mean service time for service station 1 and for service station 2

$\sigma^2_{s_1}$, $\sigma^2_{s_2}$..... the variance of service time for service station 1 and for service station 2

$c^2_{s_1}$, $c^2_{s_2}$..... The square of the coefficient of variation (ratio of the variance to the square of the mean service time) for service station 1 and for service station 2

$d_1(t)$, $d_2(t)$..... the probability density functions for times between departures from service station 1 and service station 2

$\bar{d}_1$, $\bar{d}_2$..... the mean departure intervals from service station 1 and from service station 2

$\sigma^2_{d_1}$, $\sigma^2_{d_2}$..... the variance of the departure intervals from service station 1 and from service station 2

$P(n_1,n_2)$..... the joint probability of there being n_1 customers in service station 1 and n_2 customers in service station 2, both waiting and in service

$\bar{n}_1$, $\bar{n}_2$..... the mean number of customers waiting and in service in service station 1 and in service station 2

$\bar{n}_A$, $\bar{n}_B$..... the mean number of customers waiting and in service in both stations under sequence A and under sequence B

$\bar{n}_{q_1}$, $\bar{n}_{q_2}$..... the mean number of customers waiting at service station 1 and at service station 2

$\bar{n}_{q_A}$, $\bar{n}_{q_B}$..... the mean number of customers waiting at both stations under sequence A and under sequence B

$w_1(t)$, $w_2(t)$..... the probability density functions for the time waiting and in service at service station 1 and at service station 2

$w_A(t)$, $w_B(t)$..... the probability density functions for the time waiting and in service at both stations under sequence A and under sequence B

$W_1(t)$, $W_2(t)$..... the cumulative distribution functions for the time waiting and in service at service station 1 and at service station 2

$W_A(t)$, $W_B(t)$..... the cumulative distribution functions for the time waiting and in service at both stations under sequence A and under sequence B

$\bar{w}_1$, $\bar{w}_2$..... the mean time waiting and in service at service station 1 and at service station 2

$\bar{w}_A$, $\bar{w}_B$..... the mean time waiting and in service at both stations under sequence A and under sequence B

$\sigma^2_{w_1}$, $\sigma^2_{w_2}$..... the variance of time waiting and in service at service station 1 and at service station 2

$\sigma^2_{w_A}$, $\sigma^2_{w_B}$..... the variance of time waiting and in service at both stations under sequence A and under sequence B

$w_{q_1}(t)$, $w_{q_2}(t)$..... the probability density functions for the time waiting at service station 1 and at service station 2

$w_{q_A}(t)$, $w_{q_B}(t)$..... the probability density functions for the time waiting at both stations under sequence A and under sequence B

$W_{q_1}(t)$, $W_{q_2}(t)$..... the cumulative distribution functions for the time waiting at service station 1 and at service station 2

$W_{q_A}(t)$, $W_{q_B}(t)$..... the cumulative distribution functions for the time waiting in both stations under sequence A and under sequence B

$\bar{w}_{q_1}$, $\bar{w}_{q_2}$..... the mean time waiting at service station 1 and at service station 2

$\bar{w}_{q_A}$, $\bar{w}_{q_B}$..... the mean time waiting at both stations under sequence A and under sequence B

$\sigma^2_{w_{q1}}$, $\sigma^2_{w_{q2}}$..... the variance of the time waiting at service station 1 and at service station 2

$\sigma^2_{w_{qA}}$, $\sigma^2_{w_{qB}}$..... the variance of the time waiting at both stations under sequence A and under sequence B

$c^2_{w_{q1}}$, $c^2_{w_{q2}}$..... the square of the coefficient of variation (ratio of the variance to the square of the mean of waiting times) for service station 1 and for service station 2

$c^2_{w_{qA}}$, $c^2_{w_{qB}}$..... the square of the coefficient of variation (ratio of the variance to the square of the mean of waiting times) for both stations under sequence A and under sequence B

ρ_1, ρ_2..... the utilization factor for service station 1 and for service station 2, where $\rho = \lambda/\mu$

ρ_A, ρ_B..... the utilization factor for both stations under sequence A and under sequence B where ρ = 1 - probability of no waiting in both stations

α_1, α_2..... the ratio of the variance of the service time to the utilization factor for service station 1 and for service station 2 ($\sigma^2_{s_1}/\rho_1$ and $\sigma^2_{s_2}/\rho_2$)

k_1, k_2..... the number of phases in the Erlang service distribution for service station 1 and for service station 2.

In referring to queueing systems, Kendall's notation will be used. $A/S_1/m_1 \rightarrow ./S_2/m_2$ is a system of two servers in series where A denotes the distribution of interarrival times, S_1 and S_2 the distribution of service times at station 1 and station 2 respectively, and m_1 and m_2 the number of servers at each station. The symbols usually used for these distributions include:

M for Markovian or exponentially distributed interarrival or service times,

D for deterministic or constant interarrival or service times,

G for general distribution of interarrival or service times,

GI for general independent distribution of interarrival or service times,

N for normal distribution of interarrival or service times,

H for hyperexponential distribution of interarrival or service times, and

E_k for Erlangian distribution with parameter k of interarrival or service times.

In the search for an optimal sequence of servers, several comparison measures were used. The primary measures were the first and second moments of the waiting time. This allowed direct comparison of sequences on the basis of the most commonly recognized performance standards. The sequence with the smaller mean and variance of waiting time was considered to be optimal. In addition, sequences were evaluated in terms of their waiting time distributions. In comparing sequences A and B of service stations in series, $\bar{w}_{q_A} \leq \bar{w}_{q_B}$ if and only if $\int_0^\infty [1-W_{q_A}(t)]\,dt \leq \int_0^\infty [1-W_{q_B}(t)]\,dt$. However, the relationship between distributions took on two forms that could be described according to the rules of stochastic dominance, defined as follows:

(1) Sequence A dominates sequence B by first degree stochastic dominance ($A \overset{(1)}{\leq} B$) if $W_{q_A}(t) \geq W_{q_B}(t)$ for all t. Here the distribution for waiting times in sequence A lies entirely above the distribution for waiting times in sequence B.

(2) Sequence A dominates sequence B by second degree stochastic dominance ($A \overset{(2)}{\leq} B$) if $\int_{-\infty}^{t}\left[W_{q_A}(x)-W_{q_B}(x)\right]dx \geq 0$ for all t. This is a situation in which the distributions intersect and $W_{q_A}(t) > W_{q_B}(t)$ for some values of t and $W_{q_A}(t) < W_{q_B}(t)$ for other values of t. However, the area under the distribution in sequence A is equal to or larger than that under the distribution in sequence B.

The dominance rules add depth into the comparison and evaluation of sequences of servers in series. A third degree stochastic dominance rule exists but has no application in this analysis of queueing systems. A derivation of the third degree rule was given by Whitmore.[1]

Theoretical Foundation

Analytic results in the study of tandem queueing systems have been quite limited. The major constraint on this work has been the need to assume certain well-known distribution forms of the arrival and service patterns. For an infinite input with Poisson distribution and rate λ to the first of two stations in series, with exponential service times having rates μ_1 and μ_2 respectively, R. R. P. Jackson[2] found the steady state solution for the number of

---

[1] G. A. Whitmore, "Third-Degree Stochastic Dominance," American Economic Review 60 (June 1970):457-59.

[2] R. R. P. Jackson, "Queueing Systems with Phase Type Service," Operational Research Quarterly 5 (December 1954): 109-20.

customers in the system to be

$$P(n_1, n_2) = (\rho_1)^{n_1}(\rho_2)^{n_2}P(0,0)$$

where $$P(0,0) = (1-\rho_1)(1-\rho_2).$$

The mean number in the system is

$$\bar{n}_s = \frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2}$$

which is simply the sum of the mean number of customers at each individual station. Jackson noted that although the number in any station would seem to be dependent upon the output from the previous station, each station behaves independently in the steady state. This was later proven true only for the case of Poisson input and exponential service in the stations.

These ideas were extended to i service stations in series with queues developing before each station and unlimited Poisson input to the first station. The j^{th} station consists of r_j parallel channels all with exponential service rates μ_j. R. R. P. Jackson[3] again found the steady state solution for the number of customers and total time spent in the system.

This same system was used by Nelson[4] in determining the cumulative distribution function of waiting times at each service station. This is applicable to any problem

---

[3]R. R. P. Jackson, "Random Queueing Processes with Phase Type Service," Journal of the Royal Statistical Society, ser. B, 18 (1956):129-32.

[4]Rosser T. Nelson, "Waiting-Time Distributions for Application to a Series of Service Centers," Operations Research 6 (November-December 1958):856-62.

for which the waiting times at individual stations are independent with distributions of the form

$$W_{q_j}(t) = 1-K_j\exp(C_j t)$$

for the probability of waiting less than time t at the j^{th} station, which consists of r_j identical servers in parallel. The constants are

$$C_j = -r_j\mu_j(1-\rho_j)$$

and

K_j = probability of a customer waiting at j^{th} station.

The cumulative distribution function of the total waiting time at all i service stations is

$$W_q(t) = 1-\sum_{j=1}^{j=i} A_{ji}\exp(C_j t)$$

where

$$A_{ji} = K_j\prod_{k=1}^{k=i}\frac{1-K_kC_j}{C_j-C_k}$$

for $j = 1,2,\ldots,i$ and $k \neq j$

and where $C_k \neq C_j$ for any $k \neq j$.

DeBaum and Katz[5] presented an approximation of the sum of the exponentials in the distribution of the waiting times that is readily computable and considerably useful.

Extending the theory beyond the assumptions of exponential interarrival and service distributions was a formidable task. Fraker[6] was able to develop an approximate

---

[5]R. M. DeBaum and S. Katz, "An Approximation to Distributions of Summed Waiting Times," Operations Research 7 (November-December 1959):811-13.

[6]John R. Fraker, "Approximate Techniques for the Analysis of Tandem Queueing Systems," (Ph.D dissertation, Clemson University, 1971).

formula for the mean waiting time in a system of queues in series. He found an approximation for the variance of the departure intervals to be

$$\sigma_d^2 \approx 1/n\lambda^2 + (n-1)/n\mu^2 + (1-\rho)(n-1)/mn\mu^2 - (m-1)/m\mu^2 + 0.5(1-\rho)(m-1)(n-1)/m^2n\mu^2 + 2(1-\rho)(m-1)(n-1)/mn^2\mu^2 \quad (2.1)$$

where

n = (coefficient of variation of interarrival times)$^{-2}$

and

m = (coefficient of variation of service times)$^{-2}$.

With this approximation and the relation

$$\bar{w}_q = \lambda(\sigma_a^2 + 2\sigma_s^2 - \sigma_d^2)/2(1-\rho)$$

developed by Marshall,[7] Fraker developed a technique for analyzing tandem queueing systems.

Consider the system $M/E_{k_1}/1 \rightarrow ./E_{k_2}/1$. In the first service station, $n_1=1$, $m_1=k_1$, therefore

$$\sigma_{d_1}^2 = 1/\lambda_1^2 - (k_1-1)/k_1\mu_1^2 = \left[1-(k_1-1)\rho_1^2/k_1\right]/\lambda_1^2$$

so that

$$\bar{w}_{q_1} = \left[(k_1+1)/2k_1\right]\left[\rho_1/(1-\rho_1)\mu\right] \quad (2.2)$$

as known from theory. Since the output from station 1 is the input for station 2, $\sigma_{a_2}^2 = \sigma_{d_1}^2$. Also under steady state conditions, $\lambda_1 = \lambda_2$, therefore

$$n_2 = (1/\lambda_2)^2/\sigma_{a_2}^2 = (1/\lambda_2)^2/\sigma_{d_1}^2 = \left[1-(k_1-1)\rho_1^2/k_1\right]^{-1}.$$

---

[7] K. T. Marshall, "Some Inequalities in Queueing," Operations Research 16 (May-June 1968):651-65.

Since $m_2=k_2$, $\sigma_{s_2}^2=1/k_2\mu_2^2$, and $\sigma_{d_2}^2$ can be found using formula (2.1), the mean waiting time in station 2 is

$$\begin{aligned}\bar{w}_{q_2} = [\lambda_2/2(1-\rho_2)]\{&[1-(k_1-1)\rho_1^2/k_1]/\lambda_1^2 + 2/k_2\mu_2^2 - 1/n_2\lambda_2^2 \\ &-(n_2-1)/n_2\mu_2^2 - (1-\rho_2)(n_2-1)/k_2n_2\mu_2^2 + (k_2-1)/k_2\mu_2^2 \\ &-0.5(1-\rho_2)(k_2-1)(n_2-1)/k_2^2n_2\mu_2^2 \\ &-2(1-\rho_2)(k_2-1)(n_2-1)/k_2n_2^2\mu_2^2\}.\end{aligned}$$

When $\lambda=\lambda_1=\lambda_2$ under steady state, the first and third terms in the expression cancel to give

$$\begin{aligned}\bar{w}_{q_2} = [\lambda/2(1-\rho_2)][&2/k_2\mu_2^2 - (n_2-1)/n_2\mu_2^2 - (1-\rho_2)(n_2-1)/k_2n_2\mu_2^2 \\ &+(k_2-1)/k_2\mu_2^2 - 0.5(1-\rho_2)(k_2-1)(n_2-1)/k_2^2n_2\mu_2^2 \\ &-2(1-\rho_2)(k_2-1)(n_2-1)/k_2n_2^2\mu_2^2].\end{aligned} \quad (2.3)$$

Total waiting time in this system of two servers is $\bar{w}_{q_1} + \bar{w}_{q_2}$ found from equations (2.2) and 2.3). Fraker's work was supported by simulations of tandem queueing systems. The theory has yet to be advanced beyond first moment approximations.

In the investigation of the sequence of servers in series, analytic results through queueing theory were again restricted by the severe limitations imposed on the problems. Considering a system of i stations in series with an arbitrary arrival process, constant service times at each station, and an unlimited queue before the first station and queues of arbitrary sizes allowed before other stations,

Avi-Itzhak[8] proved the following theorems:

(1) The time spent in the system by any customer is independent of the order of the servers and of the allowable intermediate queue sizes.

(2) Letting s_m = maximum (s_j) for $j = 1,2,\ldots,i$ where s_j is the service time of the j^{th} server, then the time spent in the system by any customer is equal to $\sum_{j=1}^{j=i} s_j$ plus the time that the same customer would have been waiting in the queue of a single server system (i=1) with constant service time of s_m.

Studying the case of two stations in a sequence with no queue permitted to accumulate between the stations, Avi-Itzhak and Yadin[9] found that for Poisson arrivals with rate λ and constant service times s_1 and s_2 where s_m = maximum (s_1,s_2), then the moment generating function of the steady state time in the system is

$$M_t(z) = E(e^{zt}) = \frac{(1-\rho_m)z\exp(\left[s_1+s_2\right]z)}{z+\lambda-\lambda\exp(s_m z)}$$

and the mean time in the system is

$$\bar{w} = s_1 + s_2 + \frac{\lambda s_m^2}{2(1-\rho_m)} .$$

If the service times at both stations are exponentially distributed with rates μ_1 and μ_2, then the moment

---

[8]B. Avi-Itzhak, "A Sequence of Service Stations with Arbitrary Input and Regular Service Times," Management Science 11 (March 1965):553-64.

[9]B. Avi-Itzhak and M. Yadin, "A Sequence of Two Servers with No Intermediate Queue," Management Science 11 (March 1965):565-71.

generating function of the customers time in the system is

$$M_t(z) = \frac{\rho_1+\rho_2-\rho_1^2-\rho_2^2-\rho_1\rho_2}{\rho_1+\rho_2+\rho_1\rho_2} \frac{\mu_1\mu_2(\lambda+\mu_1+\mu_2-z)}{(\mu_1-z)(\mu_2-z)(\mu_1+\mu_2-z)-\lambda\mu_1(\mu_1-z)-\lambda\mu_2(\mu_2-z)-\lambda(\mu_1-z)(\mu_2-z)} .$$

In both cases, the time spent in the system is independent of the order of the stations.

Tembe and Wolff[10] did work on a system of two servers in series with the assumptions of unlimited queues, identically distributed and mutually independent service times, and neither arrivals nor service times dependent on the order of service. Results were derived for the sequence of two stations in terms of the total time in the system. They considered $s_1(t)$ and $s_2(t)$ where s_1 and s_2 were service times at the respective stations. If $P(s_1 \geq s_2) = 1$, then the waiting time distribution is stochastically smaller when the station with the longer service (station 1) is first in the order. This is equivalent to first degree stochastic dominance. If one station has a constant service time and the other station has a service time which is a random variable, the stochastically smaller waiting time distribution will occur when the constant service station is first in order. This same conclusion holds when the first service station is made up of many servers in parallel, all with constant service times. All

---

[10]Shantanu V. Tembe and Ronald W. Wolff, "The Optimal Order of Service in Tandem Queues," Operations Research 22 (July-August 1974):824-32.

of these results were found after imposing rather stringent assumptions about the service distributions.

Various areas of operations research were considered as possible sources for suggesting rules for ordering service stations in the tandem queueing problem. A popular topic in the current literature deals with job shop scheduling. The majority of academic research in shop scheduling has centered on the sequencing problem. This is concerned with determining the sequence in which a set of jobs is to be performed on each of a number of machines in order to optimize a particular performance measure. However, there are several differences between the sequencing of jobs and the ordering of servers in queueing systems. In jog sequencing, it is the job which is of concern and all dispatching rules relate to it. The machine ordering is usually described by some precedence relations so that the operations on each product are to be performed in a specified order. Machines operate independently and machine processing times are known and finite. In series queueing systems, customers, analogous to jobs in the scheduling problem, are served on a first-come, first-serve basis. It is the service stations, analogous to machines, whose ordering is of concern. In these stations, service times may not be independent and are represented by probability distributions.

Job sequencing and shop scheduling problems have been formulated and analyzed using combinatorial analysis,

heuristics, integer programming, and network analysis approaches.[11] Systems of servers in series must rely on queueing theory and simulation techniques. Because of the extensive differences in the nature of the problems, there is little application of solution rules developed in job sequencing to the problem of tandem queues.

Assembly line balancing involves finding a sequential arrangement of work stations. Each station is composed of several tasks to be performed on a product. Since the assembly line is balanced, the product spends the same time at each work station, called the cycle time. Precedence relations and zoning constraints again have an important part in determining feasible sequences of tasks and solution techniques generally rely on combinatorial analysis. Heuristics used for the solution of complex problems suggest rules for ordering tasks based on two ideas:[12]

(1) Order a task on the basis of its performance time. Almost all analysis assumes constant performance times although recent work is beginning to relax this assumption.[13]

---

[11] S. Ashour, Sequencing Theory, Lecture Notes in Economics and Mathematical Systems, vol. 69 (Heidelberg and New York: Springer-Verlag, 1972), p.3.

[12] Fred M. Tonge, "Assembly Line Balancing Using Probabilistic Combinations of Heuristics," Management Science 11 (May 1965):727-35.

[13] Fred N. Silverman, "The Effects of Stochastic Work Times on the Assembly Line Balancing Problem," (Ph.D. dissertation, Columbia University, 1974).

(2) Order a task based on precedence relations and the number of tasks that must follow it.

In the queueing model under study, neither constant performance times nor precedence relations are relevant assumptions. Thus the concept of line balancing contributes little to the study of series queueing systems.

The first applications of queueing theory to line balancing were made by Richman and Elmaghraby[14] and Koenigsberg.[15] They considered work cycle times as random variables with an exponential distribution. In addition, raw material enters the production line at a rate which has a Poisson distribution. With these assumptions they were able to study the production line on the basis of known results in tandem queueing theory.

The use of queueing theory in the study of production lines led to new developments. New designs suggested an operator's performance time be represented by a probability distribution and not a constant. With queues of parts allowed to build up, an operator could balance his long performance cycles with his shorter cycles and the result would be a more efficient use of the worker. This was referred to as an unpaced assembly line.

---

[14]Eugene Richman and Salah Elmaghraby, "The Design of In-Process Storage Facilities," Journal of Industrial Engineering 8 (January-February 1957):7-9.

[15]Ernest Koenigsberg, "Production Lines and Internal Storage--A Review," Management Science 5 (July 1959): 410-33.

As compared with the series queueing problem, the design of unpaced lines has a different objective. For the production lines, the problem has been reduced to determining the optimum inventory levels where operator service times are random variables. The order of production stations has been treated as a fixed variable.[16,17] Thus, whereas queueing theory can offer guidelines for handling storage problems in unpaced production lines, solutions to the production line problem offer little help in determining the optimal sequence of service stations in series.

Hypotheses

There are many voids that remain in the theory of queues in series that have significant and practical importance. A natural extension of the literature is to consider many types of stochastic service processes and determine rules for ordering the service stations to result in optimal performance standards. A review of current findings in the sequencing of servers shows limited results; these refer to exponential and constant service processes. However, they do suggest the following factors to be studied:

(1) The effect of the variance of the service distributions on the sequence of station in series. This was

---

[16]Kenneth Barten, "A Queueing Simulator for Determining Optimum Inventory Levels in a Sequential Process," Journal of Industrial Engineering 13 (July-August 1962):245-52.

[17]Henry P. Goode and S. Saltzman, "Estimating Inventory Limits in a Station Grouped Production Line," Journal of Industrial Engineering 13 (November-December 1962):484-90.

implied by observing the results of positioning the constant service stations in the order of servers.

(2) The effect of service station utilization rates on the sequence of stations in series. Results of ordering on the basis of service times had suggested this factor for study.

(3) Some relationship between the variance and utilization rates of the servers.

(4) The effect of the coefficient of variation of the service distribution on the sequence of stations in series. The indifference relationship between stations of exponential service brought out this factor. However, experimentation not reported here as well as analytical studies had shown this not to be a major factor in the analysis of the sequencing of servers.

After a thorough investigation of the current literature dealing with queues in series, the following hypotheses are proposed:

(1) For a range of system parameters a sequence can be found which is optimal on the basis of total waiting time moments and stochastic dominance rules.

(2) A single station queueing system can be found which is isomorphic with the series queueing system with respect to waiting time.

The first hypotheses is tested in Chapter IV where results of the analysis of the sequencing problem are presented. In Chapter V, concerning the study of isomorphic

systems, the second hypothesis is examined. Conclusions based upon these investigations can be found in Chapter VI.

The research in ordering service stations goes beyond the current literature to extend the realm of known results. Also, the study of isomorphic queueing systems opens up an area which has not been greatly researched. Its significance can be quite meaningful. This thesis, in dealing with these problems, will result in contributions to the theoretical framework of the study of series queueing systems.

CHAPTER III

SIMULATION TECHNIQUES, PROGRAM, AND EXPERIMENTAL DESIGN

Problems in the Simulation of Stochastic Systems

Digital simulation is a powerful tool for analyzing complex systems. However, there are many problems that deserve serious attention. The performance of a system is measured by one or more characteristics of system state variables. Generally this characteristic is a frequency distribution or a mean value. In queueing situations there are a number of simulation methods which are available for determining such variables as the number of customers in the queue and their waiting time.

Gafarian and Ancker[1] introduced two dichotomies in constructing a digital simulation model.

(1) Event-sequencing and time-slicing programming. In event-sequencing, the simulation program moves from one event to the next while recording the state of the system at all times and the time between events. All system changes are recorded. Time-slicing simulation will record

---

[1]A. V. Gafarian and C. J. Ancker, Jr., "Mean Value Estimation from Digital Computer Simulation," Operations Research 14 (January-February 1966):26.

the state of the system at regular, fixed intervals of time. However, some information may be lost within any interval of time. As a result, event-sequencing is preferred and outweighs the savings in computer-time costs using the time-slicing method. GPSS (General Program Simulation System) is an event-sequencing simulation language chosen to be used in this research.

(2) Terminating and nonterminating systems. Terminating processes are ones in which the simulation ends if a specified event occurs, such as an equipment failure model. In a nonterminating system, no special event stops the simulation run. The system can be stopped and started arbitrarily but not as a result of a critical occurrence. The problem of series queueing systems meets the conditions of a nonterminating process.

In this research the estimation of the steady state characteristics are of importance rather than transient results. It is necessary to determine the appropriate simulation procedure for this analysis. Three methods have been cited most frequently in the literature for estimating the steady state response in nonterminating systems. These are (1) independent blocks or tours, (2) replicated runs, and (3) continued runs.

(1) Independent blocks was first suggested by Cox and Smith[2] in a discussion of the concept of tours and

---

[2]D. R. Cox and Walter L. Smith, Queues (London: Methuen & Co., 1961), pp. 127-36.

developed by Kabak.[3] The state of the system is defined as the number of customers in the system either being served or waiting to be served. When an event, either an arrival or a departure, causes the system to be in state i, a tour is begun. A departure from state i and a subsequent return, caused by the same event that began the tour, completes it. The procedure defined by Crane and Iglehart[4] suggests that the simulation start in the empty state and a tour is completed when the system returns to its empty state. All observations, transient and steady state ones, are averaged within a tour. The blocks of observations created in this fashion are independent and identically distributed. No observations are wasted since all are recorded. The key requirement for obtaining these independent and identically distributed blocks is that the system being simulated returns to an empty state infinitely often and that the mean time between such returns is finite.

(2) Replicated runs is a classical method employed in simulation problems. The technique involves replicating the entire process by making a series of runs that are completely separate and independent, usually by using a new random number generator seed for each run. The analysis

---

[3]Irwin W. Kabak, "Stopping Rules for Queueing Simulations," Operations Research 16 (March-April 1968):431-37.

[4]Michael A. Crane and Donald L. Iglehart, "Simulating Stable Stochastic Systems, I: General Multiserver Queues," Journal of the Association of Computing Machinery 21 (January 1974):103-13.

of replicated runs is very simple since they are independent random variables. Each run, however, involves the same problems of starting conditions and steady state. It takes some time for the simulation to overcome the artificiality introduced by the abrupt beginning of operation of the system. The initial period or transient state is a period of time when the performance of the simulated system is distorted. Since this research studies only steady state, the solution is to exclude information from consideration from an initial stabilization period and run the system through steady state. Unfortunately transient state observations are thrown away for each replication resulting in a wast of computer time.

(3) Continued runs has been discussed by Conway, Johnson, and Maxwell[5] and Conway.[6] A long, continued run can be divided into n subruns with the transient observations excluded for the first subrun. Then the conditions that exist at the end of this subrun may provide information for reasonable starting conditions of the second subrun in order to accelerate steady state. The stabilization period for the second subrun should be shorter than for the first subrun and so on for the subsequent subruns. For each subrun, information from the previous subrun is used to

---

[5]R. W. Conway, B. M. Johnson, and W. L. Maxwell, "Some Problems of Digital Systems Simulation," Management Science 6 (October 1959):109.

[6]R. W. Conway, "Some Tactical Problems in Digital Simulation," Management Science 10 (October 1963):55-57.

specify a pre-load starting condition which will expedite steady state. This method is just a continuation of the initial run with intervals during which no measurements are obtained.

A problem of dependence among observations exists with this method. However, serial correlation between observations decreases as the distance between observations increases. Thus, only the first few observations of a subrun will be correlated with the last few observations of the previous subrun and output of each subrun show only small correlations. If the subruns are long enough, the correlations among subrun output can be neglected and the result is independent random variables from each subrun.

The choice of simulation technique for this research paper was guided by the comparison and application of the three methods in the current literature. In addition, consideration was given to the fact that the main purpose of this thesis is to study the behavior of series queueing systems and not to solve problems in the simulation of stochastic processes. Each method has its own advantages that must be weighed in light of the problem to be studied and the availability and cost of computer time.

Some of the recent literature dealing with the simulation of stochastic processes have praised the advantages of using independent blocks. Fine applications of this procedure to queueing simulations have been made by

Fishman[7] and Law[8]. In order to estimate the expected value of the total waiting time in the system for a customer, data must be gathered for each block. This data includes the total waiting time and the number of customers during each block. A point estimate of the total waiting time would be obtained by taking the ratio of the average waiting time per block to the average number of customers per block. This ratio estimate has a bias of order $1/n$ where n is the number of independent blocks. The standard error of the estimate is of order $1/\sqrt{n}$, and likewise the ratio of the bias to the standard error is also of order $1/\sqrt{n}$. This bias becomes negligible as n becomes large and, in practice, is usually found to be unimportant even in samples of moderate size.[9] Statistical formulas necessary to calculate a confidence interval around this estimate are given in Hillier and Lieberman.[10]

Serious thought was given to adopting independent blocks as the simulation procedure for this analysis. One

---

[7]George S. Fishman, "Statistical Analysis for Queueing Simulations," Management Science 20 (November 1973):363-69.

[8]Averill M. Law, "Efficient Estimators for Simulated Queueing Systems," Management Science 22 (September 1975): 30-41.

[9]William G. Cochran, Sampling Techniques, 2d ed. (New York:John Wiley & Sons, 1963), p. 160.

[10]Frederick S. Hillier and Gerald J. Lieberman, Operations Research, 2d ed. (San Francisco: Holden-Day, 1974), pp. 645-46.

problem considered was that of sample size. For the case of service stations in series, the probability that all servers are idle at the same time and the system is in an empty state, is small. The simulation would have to run considerably longer to collect data on n tours in the series queueing system than in the single server system for the same level of activity. Some pilot runs at various utilization rates showed that this was not a serious problem with two servers in series.

The most decisive factor was the particular information that had to be generated by simulation. Independent blocks are useful in estimating the first moment of some random variable of interest. However, the problem being studied calls for the variance as well as a probability distribution of the waiting times, information which can not easily be found by use of independent blocks. In addition, Crane and Iglehart[11] found no significant differences to exist between performance estimates using the methods of independent blocks and replicated runs. All these considerations led to a choice against the use of the independent blocks technique.

A number of authors have made a comparison between the techniques of continued runs and replicated runs and all agree that replication of runs is more efficient. Gafarian and Ancker[12] concluded that it is better to replicate

---

[11]Crane and Iglehart, "Stochastic Systems," pp.111-13.

[12]Gagarian and Ancker, "Mean Value Estimation," pp. 33-34.

than to extend the observation interval in the form of a continued run if individual observations have a positive serial correlation, a mild assumption which is common in many simulations. Mihram[13] claimed that the recommended procedures for determining the stabilization periods between subruns are somewhat heuristic in a continued run, and recommended the use of independent replicated runs. Kleijnen[14] summarizes the arguments in favor of replication of runs.

Considering the arguments for and against each method, and keeping in mind the particular problem to be investigated in this research, it was decided to use the replication of runs technique for generating simulation output. It then became necessary to deal with the problem of transient behavior in the system. Since only steady state results are of interest, the transient bias must be eliminated. Conway[15] recognized the solution to the problem as simply (a) to exclude data from some initial period from consideration, and (b) to choose starting conditions that make the necessary excluded interval as short as possible.

---

[13]G. A. Mihram, Simulation: Statistical Foundations and Methodology (New York: Academic Press, 1972), pp. 448-49.

[14]Jack P. C. Kleijnen, Statistical Techniques in Simulation, part 1 (New York: Marcel Dekker, 1974), pp. 85-87.

[15]Conway, "Tactical Problems," pp. 48-51

There are no statistical techniques known for determining when the transient state ends. Conway[16] suggested making a few pilot runs with periodic reporting over short intervals and plotting the performance of these runs against the number of customers processed through the system. This gives a rough idea of what might be a reasonable exclusion. Next, leave out the initial observations as long as they continue to increase or decrease. This decision of a stabilization period is then applicable to each of the independent runs.

Since transient observations are excluded, the computer time needed to reach steady state is wasted. The selection of adequate starting conditions can minimize this wasted time. The choices of initial conditions available for selection include (1) the system in its empty state, or (2) the system in a non-empty state determined by some a priori knowledge of the system. Whereas the empty starting condition is simplest to work with, it usually prolongs reaching steady state. Selecting the initial state on the basis of prior knowledge has an advantage in terms of saving computer time. However, if an unlikely state were chosen to start the simulation, it could bias the resulting output.

The search for good starting conditions is a practical question. There are certain circumstances in which

---

[16]Ibid., p. 49.

it would be difficult to determine reasonable starting conditions and the investigator was willing to consume the additional computer time to avoid this task. The main factors to consider are the extent of a priori knowledge of the system and the availability of computer time. Hillier and Lieberman[17] advised using the same starting conditions for all simulation runs once initial conditions are chosen. This is especially true when comparing alternative systems in order to avoid biasing the comparisons.

Pilot runs were made on several series queueing systems with various utilization rates at each service station. Due to the lack of prior knowledge, it was decided to start the system in the empty state for all experiments. While a comprehensive study of transient behavior under different starting conditions would certainly be interesting, it should be reserved for later work. In order to analyze the transient and steady state conditions, the behavior of the system was recorded at various intervals up to a total of 30,000 customers. Both the average waiting time and the standard deviation about this estimate were plotted against the number of customers. The suggestions of Conway[18] and Kleijnen[19] were followed in determining a

---

[17]Frederick S. Hillier and Gerald J. Lieberman, Introduction to Operations Research (San Francisco: Holden-Day, 1967), pp. 462-63.

[18]Conway, "Tactical Problems," p. 49.

[19]Kleijnen, Statistical Techniques, pp. 70-71.

stabilization period. It was found sufficient to limit attention to just 25,000 customers. The analysis suggested that the first 7,000 customers be excluded from each run, and the following 18,000 customers be analyzed.

The next step was to determine the sample size or number of replications of each experimental run. Using the pilot runs, confidence intervals were constructed about the mean waiting time estimates. For any increase in the sample size, the standard error about the mean is reduced making the confidence interval narrower. As a result of this relationship between the standard error of the mean and the sample size, the population mean may be estimated within any desired degree of precision, given a large enough sample size. A decision was made as to the size of the deviation about the mean and the number of replications was found based on 95% confidence. It should be obvious that as the utilization rates are increased at each service station in the system, the mean waiting time increases along with the standard error of the mean. In order to maintain the same degree of precision, more replications are needed. As a result, based upon pilot runs with different utilization rates, required sample sizes vary for each experiment. Anywhere from seven to thirteen replications were needed depending upon the utilization rates of the service stations in the series.

The final step was to validate the computer model. Several steps in the validation process are discussed by

Meier, Newell, and Pazer.[20] Generally, assurances of validity are provided when the simulator can produce results that are consistent with the known performance of the real system for some alternative versions of the simulated system under certain conditions. In this case, statistical tests were used to compare sample data generated by the simulation model with known analytic data for certain system performance outputs. The system used for study was the M/M/1-►./M/1 with various server utilization rates. Upon passing the tests, the simulation program was judged valid and ready to be used to conduct further experiments.

The Computer Program

The simulation program used in this analysis is of a more general nature than required for the systems studied in this research. It can be useful for a large variety of queueing networks. Because of the type of problems to be investigated, the choice was to use a special-purpose simulation language. These languages are designed to assist analysts in the design, programming, and analysis of simulation models. Although language developers often claim that their particular language is all-encompassing in application, sufficient differences exist so that proper selection may result in considerable time and cost savings in implementing a simulation model.

---

[20]Robert C. Meier, William T. Newell, and Harold L. Pazer, Simulation in Business and Economics (Englewood Cliffs, New Jersey: Prentice-Hall, 1969), pp. 294-96.

Kiviat[21] identified ten important features common to simulation programming languages. The way they are elaborated and implemented makes particular special-purpose languages difficult or easy to use, programmer- or analyst-oriented, and so on. These features include: modeling a system's static state, modeling system dynamics, statistical sampling, data collection, analysis and display, monitoring and debugging, initialization and language usability.

GPSS is the most applicable language for the problem under study. GPSS stands for "General Purpose Simulation System" and was developed by IBM. Historically, the language had its beginnings as early as the late 1950s but it has moved through several stages over the years. Actually GPSS is both a computer language and a computer program. As a language, it has a well-defined vocabulary and grammar to describe certain types of system models. As a computer program, it interprets a model described in the GPSS language. The normal procedure is to prepare a block diagram of the system to be simulated which is then punched on cards. This is then interpreted by the GPSS Processor and the system is simulated by the computer, allowing experiments to be conducted with the model.

---

[21]Philip J. Kiviat, "Simulation Languages," in Thomas H. Naylor, ed., Computer Simulation Experiments with Models of Economic Systems (New York: John Wiley & Sons, 1971), pp. 413-36.

The orientation of GPSS is one of transactions moving in time through a system composed essentially of facilities, storages, and queues. Simulated time moves forward in unequal time increments. Being event oriented, the "clock" is moved from one instant of time to the next instant where the system changes its state. The program can compile and print out certain statistics regarding facility utilization, storage utilization, queue contents, and the number of transactions flowing through the blocks in the system. A great deal of flexibility exists in the construction of a model which allows systems of considerable complexity to be simulated by the use of basic block types.

The computer program used in this research can be divided into three parts: input data for the simulator, the simulator which simulates and analyzes the behavior of the system, and the output data which supplies the desired statistical information.

All input data is supplied by the experimenter. These include:

(1) The number of service centers in the series queueing system.

(2) The probability density function of time between arrivals to the system. In this study the density function of the interarrival times was always exponential although other forms may have been specified.

(3) The parameters for the density function of the time between arrivals. This includes the mean interarrival

time and any other parameters necessary to be specified for a given form of the density function.

(4) The probability density function of the service times for each service station. Although an Erlang density with parameter k was used throughout this study, the computer program can accommodate many other forms. The density function may vary at each service station.

(5) The parameters for the density function of the service times. The mean service time at each station must be specified along with any other parameters which are needed to describe a particular form of the density function. In this study the service times were represented by Erlang density functions. Therefore, the Erlang parameter k, where k was an integer greater than or equal to one, was required.

(6) The number of customers to be processed before reaching steady state conditions and the total number of customers to be generated. Through some trial runs it was decided to eliminate the accumulated system statistics for the first 7,000 customers. This data was considered to represent the transient behavior of the system. An additional 18,000 customers were processed for each run through the system.

(7) Random number seeds. A random number generator is built into GPSS and is used each time a random number is needed to determine interarrival times or service times. A total of eight base numbers called seeds are included

from which uniform random numbers are calculated. A duplication of each run can be obtained by using the same seeds.

The major part of the computer program is involved with the simulation and analysis of the behavior of the system by generating and processing customers. This portion of the program is expressed in GPSS block format in Appendix A. Each run of the computer program simultaneously simulates the operation of two sequences of stations that are to be compared on the basis of their waiting times. The results gathered for each individual sequence include information on utilization rates, waiting times, size of waiting lines, and interarrival and interdeparture times. A final analysis is made by determining the difference in total waiting time for individual customers in the two sequences being studied. The reliability of this estimate is increased by use of a variance reduction technique described by Kleijnen.[22]

The general expression for the variance of the difference between w_{q_A}, the estimated total waiting time in sequence A, and w_{q_B}, the estimated total waiting time in sequence B, is

$$\sigma^2_{w_{q(A-B)}} = \sigma^2_{w_{qA}} + \sigma^2_{w_{qB}} - 2\mathrm{cov}(w_{q_A}, w_{q_B}).$$

The variance of the estimated difference is decreased if the covariance term can be made positive. By using the

---

[22]Kleijnen, Statistical Techniques, pp. 200-206.

same sequence of random numbers to determine the stochastic interarrival and service times, such a positive covariance is created. This is because both sequences react to the stochastic input variables in the same way. Low interarrival times and long service times tend to result in longer waiting times irregardless of the order of the service stations, the number of stations, the arrival and service rates, etc.

The computer program reports an extensive amount of useful output regarding the behavior of the systems being simulated. Although most of the analysis in this study was based upon an evaluation of system waiting times, the other information which is generated might prove useful in subsequent studies. The following output is obtained from the program for each individual system:

(1) The mean utilization rate of each service station.

(2) The mean number of customers waiting in the queue before each service station and the maximum contents of each queue.

(3) The mean, variance, histogram, and cumulative distribution of the waiting time in the system and at each service station queue.

(4) The mean, variance, histogram, and cumulative distribution of the total time (waiting and in service) in the system and in each service station.

(5) The mean, variance, histogram, and cumulative distribution of arrival intervals to each service station.

(6) The mean, variance, histogram, and cumulative distribution of departure intervals from each service station.

(7) The mean, variance, histogram, and cumulative distribution of service times at each station.

In addition the following information comparing the two sequences of stations which are simulated simultaneously is reported:

(8) The mean, variance, histogram, and cumulative distribution of the difference in total waiting times for individual customers in the two sequences.

The entire computer program can best be expressed in the following flow chart to show the step-by-step approach used in the simulation of series queueing systems. (See Fig. 2.) Each computer run represents the operation of two sequences of service stations. These two sequences differ only with respect to the order of their stations in the series. They are being simulated simultaneously so that a comparison can be made as each individual customer is being processed.

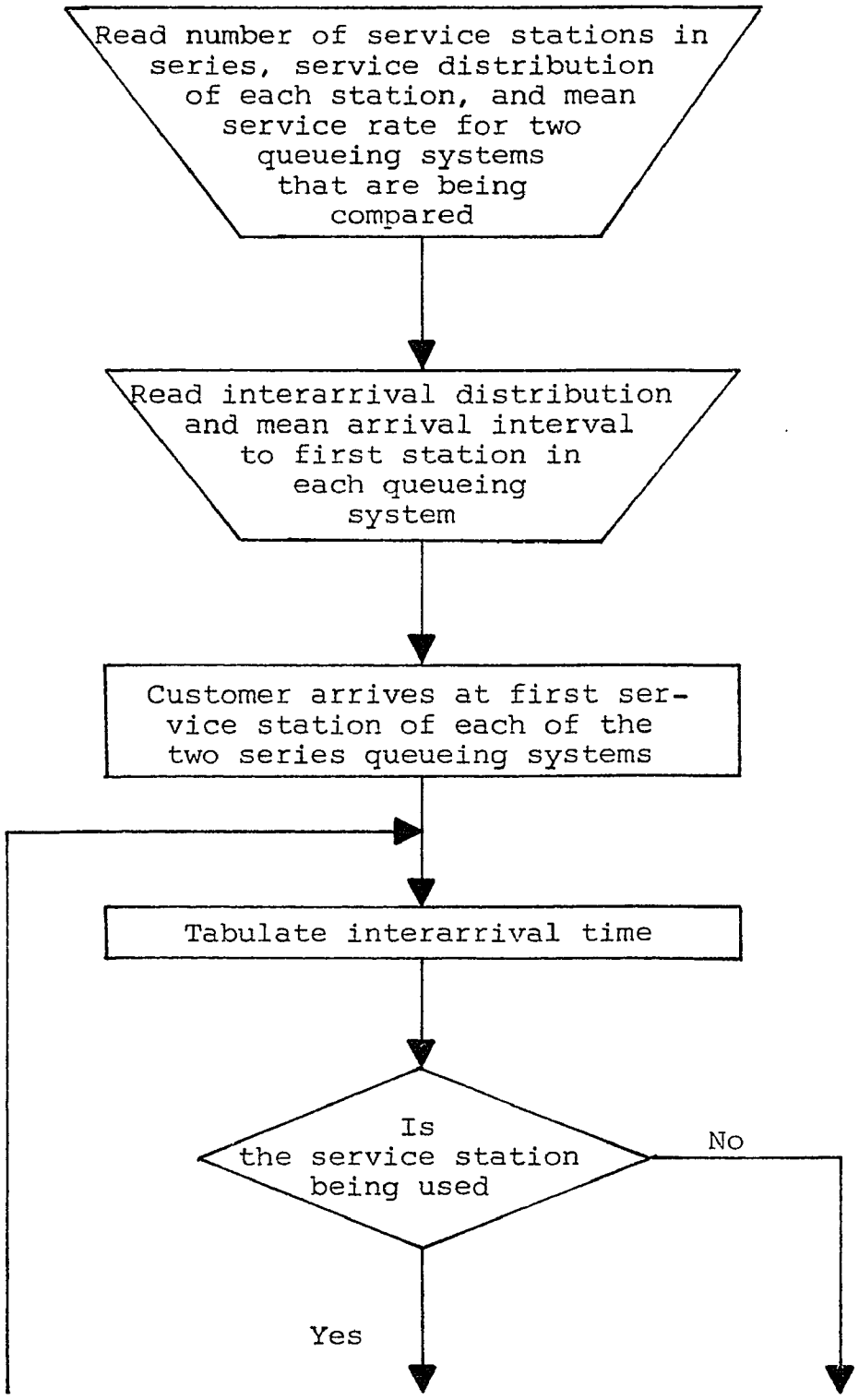


Fig. 2. Flow Chart of Computer Program

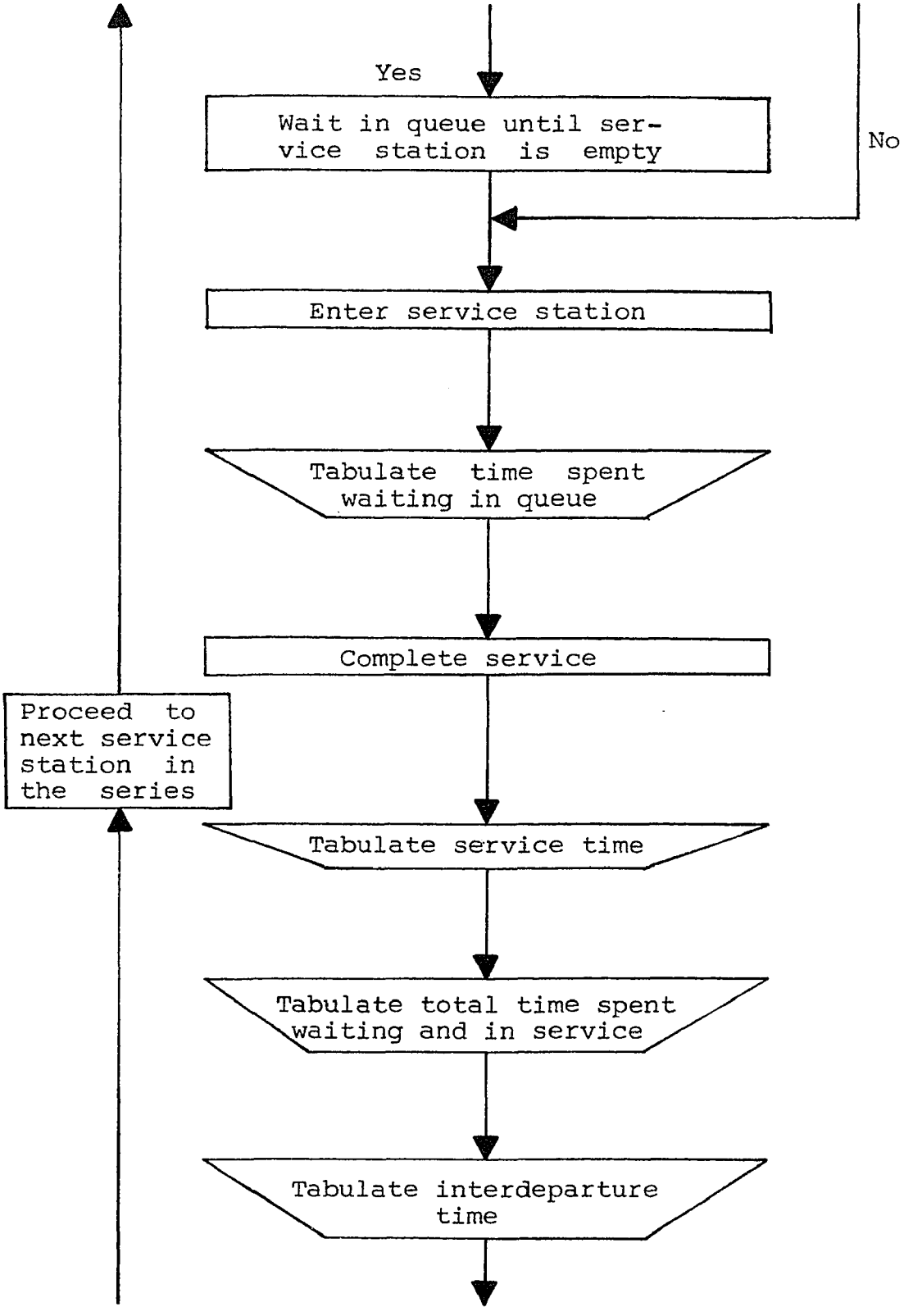


Fig. 2--Continued

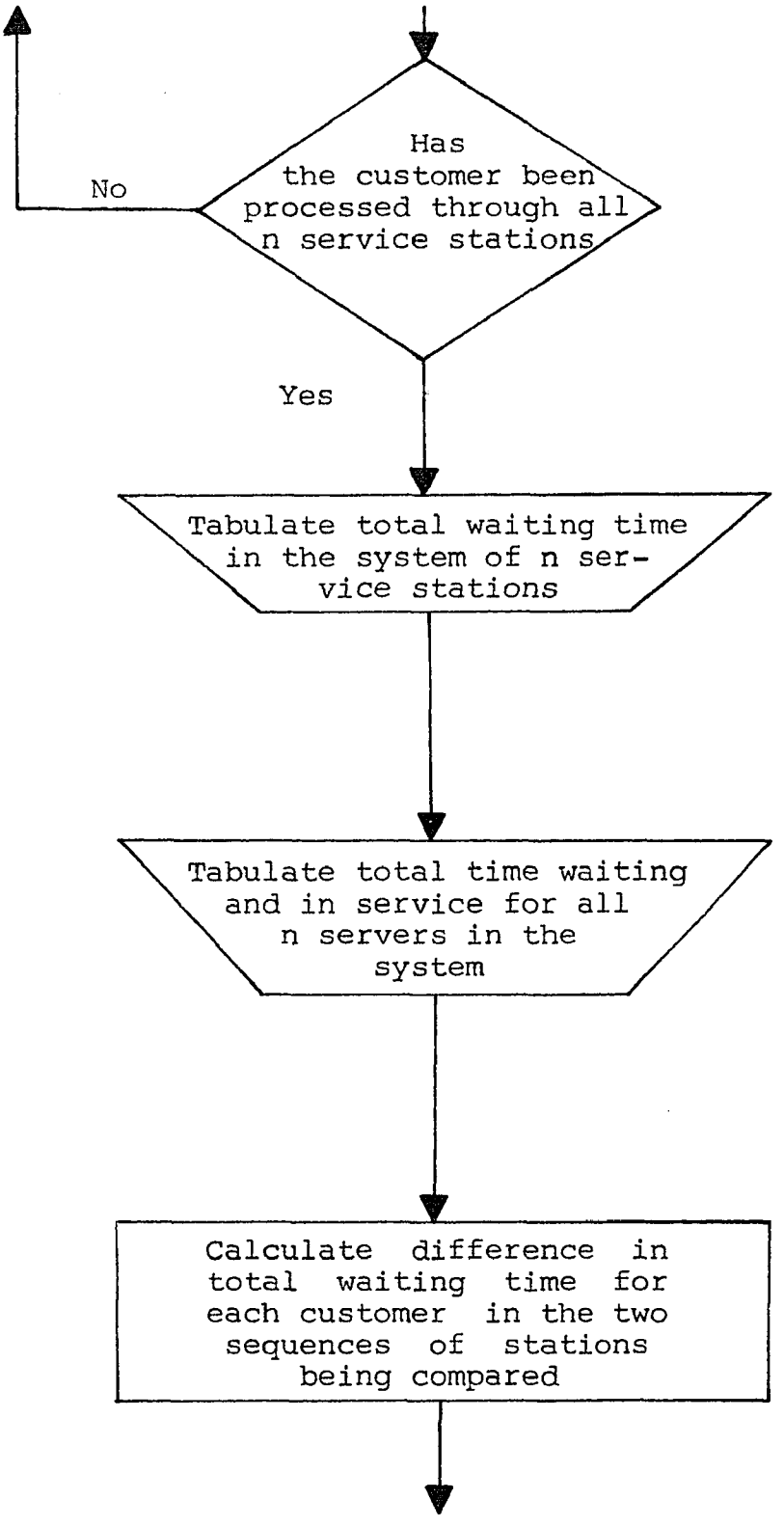


Fig. 2--Continued

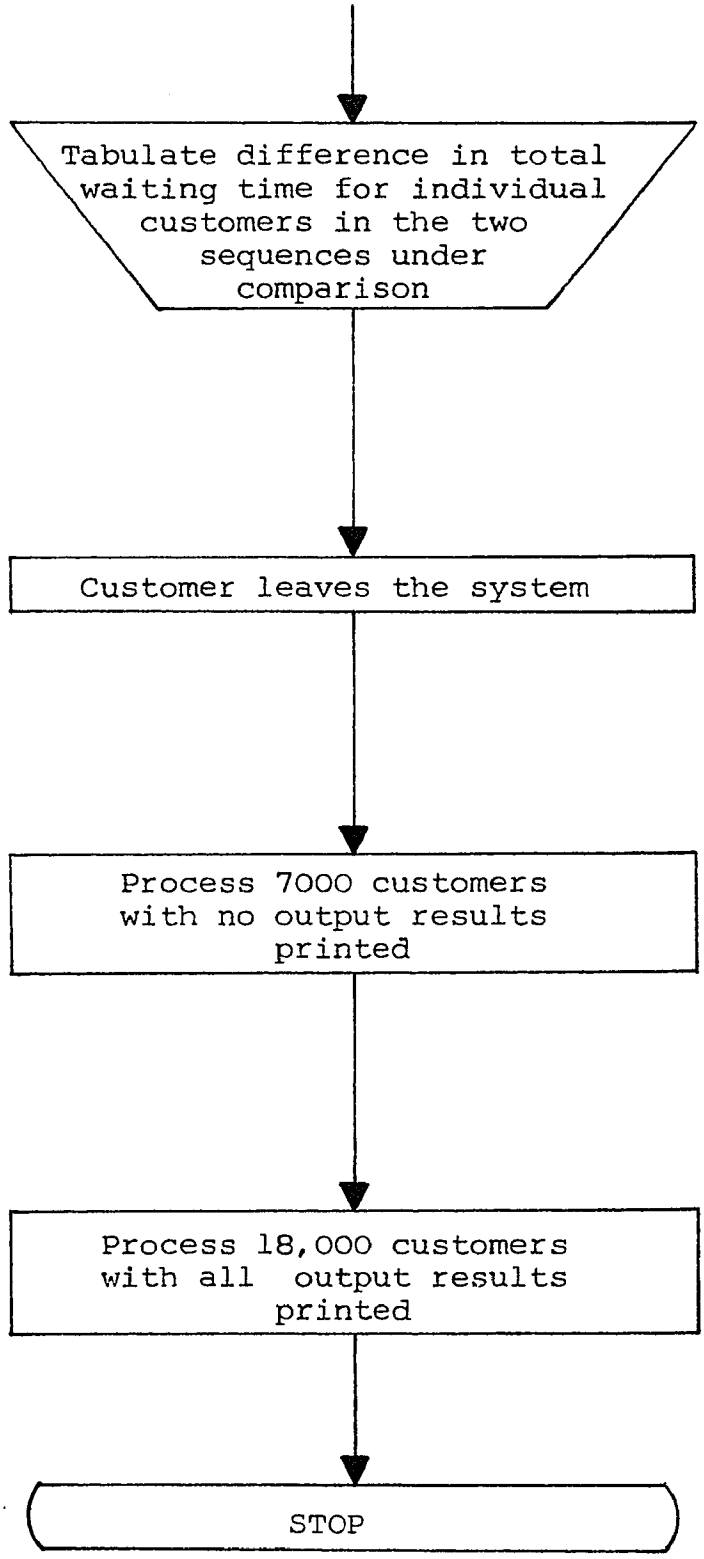


Fig. 2--Continued

Experimental Design

The inability to use mathematical analysis alone as a solution technique in tandem queueing systems, calls for experimental results from system simulations. Simulation, as a supplement to mathematical analysis, can be useful in obtaining a solid understanding of queueing systems to serve as the foundation for general theory of system behavior. Results in the form of empirically tested hypotheses and tables of data serve as comprehensive and experimental theory.

The specific experimental design was decided upon with the idea of choosing parameters representing systems that are realistic while at the same time selecting diverse sets of parameters. In addition, the computing time serves as a constraint on the number of experiments that are feasible. Because of the large number of parameters, the number of levels for any given factor was limited. They were not the only combination of possibilities that existed, but the results of these experiments served the purpose of this paper.

The simulation experiments were conducted with the two-service station model discussed in Chapter II. Two sequences were considered. In sequence A, all arriving units went to service station 1 first and then to service station 2. Sequence B had service station 2 preceding service station 1. In all experiments the distribution of the system's interarrival process was exponential with

mean interarrival time, $1/\lambda$, of 100 time units of the simulator. The exponential interarrival distribution, widely used in queueing research and quite frequent in real life problems, allowed experimental results to be compared with known theoretical results. The choice of the mean of 100 time units made determination and control of utilization rates an easy task.

Another factor in the design was the choice of the form of the service processes at the two stations. Steady state properties with regard to order of stations have been analytically obtained for the systems in which either one or both stations have a constant service time or both have exponential service distributions. It was decided to use a number of service processes that vary in terms of randomness and the Erlangian family of probability distributions with degree k was chosen. The Erlang density function

$$\left[(k\mu)^k/(k-1)!\right]\exp(-k\mu t)t^{k-1} \qquad (t\geq 0,\ k \text{ integer} \geq 1)$$

has a mean of $1/\mu$ and variance of $1/k$ its mean squared. The usefulness of this distribution rests in the fact that for a constant mean service time the family of Erlangian distributions interpolates infinitely many distributions between the completely random negative exponential ($k=1$) and completely regular constant service time ($k=\infty$). Ghosal[23] observed that for k greater than 4, the Erlang

---

[23] A. Ghosal at the International Conference on Stochastic Processes, "Isomorphic Queueing Systems," University of Maryland, 1975, p. 6. (Mimeographed.)

distribution gives results similar to a normal distribution. Morse[24] claimed that the Erlang distributions will not fit all possible service time distributions, but they will fit many, if not most, of the ones encountered in practice.

The following Erlangian distributions were used in this study. These were chosen to be representative of varying degrees of randomness.

Parameter k	(Coefficient of Variation)^2
1	1
2	1/2
3	1/3
4	1/4
9	1/9

Each experimental run is identified by a number and a letter. The number refers to a particular experiment with different stations in a series while the letter distinguishes the two sequences of the same stations (either A or B). In all sets of experiments, both possible sequences of the same servers were compared in order to determine the optimal ordering. The first set of simulation runs were designed to test the effect of the variance of the service distributions on the order of the queueing stations. The utilization rates at each of the two service stations were equal and arbitrarily chosen to be 0.75. This value

---

[24]Philip M. Morse, Queues, Inventories and Maintenance (New York: John Wiley & Sons, 1958), p. 41.

was neither high nor low but one which was realistic and proved to make an interesting queueing problem. As a result, the mean service times at the two stations were $1/\mu_1 = 1/\mu_2 = 75$ time units. A comparison of the steady state output of the following experimental runs were made:

Experiment	Sequence A	Sequence B	$\sigma_{s_1}^2$	$\sigma_{s_2}^2$
1	$M/M/1 \rightarrow ./E_2/1$	$M/E_2/1 \rightarrow ./M/1$	5625.0	2812.5
2	$M/E_2/1 \rightarrow ./E_3/1$	$M/E_3/1 \rightarrow ./E_2/1$	2812.5	1875.0
3	$M/E_3/1 \rightarrow ./E_4/1$	$M/E_4/1 \rightarrow ./E_3/1$	1875.0	1406.25
4	$M/E_4/1 \rightarrow ./E_9/1$	$M/E_9/1 \rightarrow ./E_4/1$	1406.25	625.0

The second set of experiments was designed to test the effect of utilization rates on the order of service stations. The variance of the service distributions were held constant and utilization ratios varied from 0.30 to 0.90. All possible combinations of different service station distributions were considered. The case of exponential service at both stations, for which analytic results are already known, played a part in validating the simulation model. Again the steady state output of the two sequences for each experimental system was compared. The following experiments were conducted:

Experiment	Sequence A	Sequence B	ρ_1	ρ_2
5	$M/M/1 \rightarrow ./E_2/1$	$M/E_2/1 \rightarrow ./M/1$	0.30	$0.30\sqrt{2}$
6	$M/M/1 \rightarrow ./E_3/1$	$M/E_3/1 \rightarrow ./M/1$	0.30	$0.30\sqrt{3}$
7	$M/M/1 \rightarrow ./E_9/1$	$M/E_9/1 \rightarrow ./M/1$	0.30	0.90
8	$M/E_2/1 \rightarrow ./E_3/1$	$M/E_3/1 \rightarrow ./E_2/1$	$0.30\sqrt{2}$	$0.30\sqrt{3}$

Experiment	Sequence A	Sequence B	ρ_1	ρ_2
9	$M/E_2/1 \rightarrow ./E_9/1$	$M/E_9/1 \rightarrow ./E_2/1$	$0.30\sqrt{2}$	0.90
10	$M/E_3/1 \rightarrow ./E_9/1$	$M/E_9/1 \rightarrow ./E_3/1$	$0.30\sqrt{3}$	0.90

Based upon the initial simulation results and the analytical derivations presented in Chapter IV, more experiments were conducted. These were intended to test the relationships found for optimal server sequencing through an in-depth study of two separate systems of service stations in series.

One set of experiments dealt with the two station series consisting of Erlang service distributions with parameters 2 and 3. In addition to the results of experiment 2 on this system, the following simulation runs were made:

Experiment	Sequence A	Sequence B	ρ_1	ρ_2
11	$M/E_2/1 \rightarrow ./E_3/1$	$M/E_3/1 \rightarrow ./E_2/1$	0.6841	0.75
12	$M/E_2/1 \rightarrow ./E_3/1$	$M/E_3/1 \rightarrow ./E_2/1$	0.6370	0.75
13	$M/E_2/1 \rightarrow ./E_3/1$	$M/E_3/1 \rightarrow ./E_2/1$	0.5900	0.75
14	$M/E_2/1 \rightarrow ./E_3/1$	$M/E_3/1 \rightarrow ./E_2/1$	0.2000	0.75
15	$M/E_2/1 \rightarrow ./E_3/1$	$M/E_3/1 \rightarrow ./E_2/1$	0.1000	0.75

The final simulations were conducted on a system in which one service station had an exponential service distribution and the second station an Erlang service distribution with parameter 2. These experimental runs were:

Experiment	Sequence A	Sequence B	ρ_1	ρ_2
16	$M/M/1 \rightarrow ./E_2/1$	$M/E_2/1 \rightarrow ./M/1$	0.05	0.10
17	$M/M/1 \rightarrow ./E_2/1$	$M/E_2/1 \rightarrow ./M/1$	0.05	0.20

Experiment	Sequence A	Sequence B	ρ_1	ρ_2
18	M/M/1 → ./E_2/1	M/E_2/1 → ./M/1	0.05	0.30
19	M/M/1 → ./E_2/1	M/E_2/1 → ./M/1	0.10	0.15
20	M/M/1 → ./E_2/1	M/E_2/1 → ./M/1	0.15	0.10
21	M/M/1 → ./E_2/1	M/E_2/1 → ./M/1	0.15	0.15
22	M/M/1 → ./E_2/1	M/E_2/1 → ./M/1	0.20	0.10
23	M/M/1 → ./E_2/1	M/E_2/1 → ./M/1	0.20	0.15
24	M/M/1 → ./E_2/1	M/E_2/1 → ./M/1	0.20	0.25

In Table 1, a summary of the service distributions used in each experimental run is presented. Although the choice of parameters limited the results found, exhaustive experimentation would have been an impossible task. The main consideration in planning the design of these experiments was to find a range of possible queueing systems that represent realistic situations. The experimental results were used to base hypotheses describing the behavior of the model in terms of the specific set of parameters. This led to continued experimentation geared to refining the hypotheses, adding new numerical results, and leading to general predictive theory.

TABLE 1

A SUMMARY OF THE SERVICE DISTRIBUTIONS USED IN THE SIMULATION EXPERIMENTS

Experiment	Station 1		Station 2	
	k	$1/\mu$	k	$1/\mu$
1	1	75.00	2	75.00
2	2	75.00	3	75.00
3	3	75.00	4	75.00
4	4	75.00	9	75.00
5	1	30.00	2	42.43
6	1	30.00	3	51.96
7	1	30.00	9	90.00
8	2	42.43	3	51.96
9	2	42.43	9	90.00
10	3	51.96	9	90.00
11	2	68.41	3	75.00
12	2	63.70	3	75.00
13	2	59.00	3	75.00
14	2	20.00	3	75.00
15	2	10.00	3	75.00
16	1	5.00	2	10.00
17	1	5.00	2	20.00
18	1	5.00	2	30.00
19	1	10.00	2	15.00
20	1	15.00	2	10.00
21	1	15.00	2	15.00
22	1	20.00	2	10.00
23	1	20.00	2	15.00
24	1	20.00	2	25.00

NOTE: For all experiments, the interarrival distribution is exponential with mean $1/\lambda=100$. For service station i, the utilization rate $\rho_i=(1/\mu_i)/100$ and the variance $\sigma^2_{s_i}=1/(k_i\mu_i^2)$.

CHAPTER IV

RESULTS IN THE SEQUENCING OF STATIONS IN SERIES

The Effects of Server Utilization Rates and Variance of Service Distribution on the Sequence of Stations

The observed values of the system waiting time statistics for each of the simulated systems A1 through A10 and B1 through B10 are summarized in Tables 2 and 3. The mean and standard deviation of the total waiting times are shown for each system. In addition, for sequences A and B of the same service stations, the mean difference in total waiting time for individual customers in the two sequences, $\bar{w}_{q(A-B)}$, and the standard deviation of the differences about the mean, $\sigma_{w_{q(A-B)}}$, are reported. These were found by using the mean and variance from each experimental run and averaging over all the replications. For each replication, a total of 18,000 customers were processed. Working with this data, the mean difference was tested to see if it was significantly different than zero. The values of "t" which are significant at the .01 level are indicated by *.

The two sequences of servers for each experiment were compared on the basis of their waiting time distribution functions for a test of stochastic dominance in

TABLE 2

A TEST OF SERVICE DISTRIBUTION VARIANCES
ON STATION SEQUENCES

Experiment	Sequence	System	$\bar{w}_q$	σ_{w_q}	Differences
1	A	$M/M/1 \rightarrow ./E_2/1$	397.836	368.401	$\bar{w}_{q(A-B)} = 31.523$
	B	$M/E_2/1 \rightarrow ./M/1$	364.438	357.143	$\sigma_{w_{q(A-B)}} = 371.123$
					t=37.796*
2	A	$M/E_2/1 \rightarrow ./E_3/1$	285.274	274.873	$\bar{w}_{q(A-B)} = 7.345$
	B	$M/E_3/1 \rightarrow ./E_2/1$	276.877	269.962	$\sigma_{w_{q(A-B)}} = 242.309$
					t=13.488*
3	A	$M/E_3/1 \rightarrow ./E_4/1$	238.657	239.536	$\bar{w}_{q(A-B)} = 0.579$
	B	$M/E_4/1 \rightarrow ./E_3/1$	237.413	235.901	$\sigma_{w_{q(A-B)}} = 189.763$
					t=1.358
4	A	$M/E_4/1 \rightarrow ./E_9/1$	206.094	211.199	$\bar{w}_{q(A-B} = 8.118$
	B	$M/E_9/1 \rightarrow ./E_4/1$	197.742	206.440	$\sigma_{w_{q(A-B)}} = 148.809$
					t=24.339*

*Indicates significance at the 1% level.

TABLE 3

A TEST OF SERVER UTILIZATION RATES
ON STATION SEQUENCES

Experiment	Sequence	System	$\bar{w}_q$	σ_{w_q}	Differences
5	A	$M/M/1 \rightarrow ./E_2/1$	36.023	57.460	$\bar{w}_{q(A-B)} = 2.335$
	B	$M/E_2/1 \rightarrow ./M/1$	33.333	53.607	$\sigma_{w_{q(A-B)}} = 64.901$
					t=12.771*
6	A	$M/M/1 \rightarrow ./E_3/1$	48.840	70.656	$\bar{w}_{q(A-B)} = 4.386$
	B	$M/E_3/1 \rightarrow ./M/1$	44.044	64.894	$\sigma_{w_{q(A-B)}} = 72.800$
					t=22.862*
7	A	$M/M/1 \rightarrow ./E_9/1$	450.308	450.206	$\bar{w}_{q(A-B)} = 5.519$
	B	$M/E_9/1 \rightarrow ./M/1$	444.240	453.910	$\sigma_{w_{q(A-B)}} = 275.089$
					t=9.705*
8	A	$M/E_2/1 \rightarrow ./E_3/1$	53.427	73.764	$\bar{w}_{q(A-B)} = 2.012$
	B	$M/E_3/1 \rightarrow ./E_2/1$	51.115	71.804	$\sigma_{w_{q(A-B)}} = 74.772$
					t=10.211*
9	A	$M/E_2/1 \rightarrow ./E_9/1$	450.222	472.726	$\bar{w}_{q(A-B)} = 0.190$
	B	$M/E_9/1 \rightarrow ./E_2/1$	449.971	474.471	$\sigma_{w_{q(A-B)}} = 276.929$
					t=0.332
10	A	$M/E_3/1 \rightarrow ./E_9/1$	458.870	466.553	$\bar{w}_{q(A-B)} = 0.975$
	B	$M/E_9/1 \rightarrow ./E_3/1$	457.541	476.900	$\sigma_{w_{q(A-B)}} = 297.746$
					t=1.584

*Indicates significance at the 1% level.

Figures 30 through 39 of Appendix B. In Figures 40 through 49 of Appendix C, the distribution of the difference in customer waiting times under the two sequences are shown for each experimental run. This represents customer waiting time in sequence A minus customer waiting time in sequence B, referred to as $W_{q(A-B)}$.

In experiments 1 through 4 the utilization rates of each service station were set equal and the variance of the service distributions allowed to vary. In each case, the larger mean and variance of the waiting time occurs under sequence A of the same servers. Except for experiment 3, the difference in mean waiting times is significant at the .01 level. On the basis of their waiting time distributions, $B \overset{(1)}{\leq} A$ in all cases. These results conclude that, with utilization rates held constant, the optimal sequence of service stations is the one with the smallest variance of the service distribution placed first in the order.

In studying the effect of utilization rates on sequencing, experiments 5 through 10 present results from which conclusions cannot be so obviously drawn. Here the variance of the service distribution for each station was fixed and the utilization ratios varied from 0.30 to 0.90. Systems 5 through 8 display the optimality of sequence B, where the station with the larger utilization rate is first in the order. Under this sequence the mean waiting time is significantly smaller than under sequence A at the

.01 level and the variance of the waiting times is smaller for all but system 7. On comparing the distribution functions of both sequences, $B \overset{(1)}{\leq} A$ for each case. In system 7 the two distribution functions merge together in the area of the larger waiting times.

Systems 9 and 10 exhibit situations in which the mean waiting time is smaller for sequence B but not significantly. In addition, the variance of the waiting times is larger for sequence B and the waiting time distributions for the two sequences overlap so that no sequence is optimal with respect to first degree stochastic dominance. In the area of smaller waiting times, sequence B dominates, while sequence A dominates in the area of larger waiting times.

The initial experimentation implied the following conditions for an optimal sequence: (1) the station with the smaller service distribution variance be first in order, and, though not as conclusive (2) the station with the larger utilization rate be first in order. These findings seem to agree with the known results in sequencing. It has been proven that if two stations are being ordered and one has a constant service time (variance of zero), it should be placed first in order to create a sequence with the smallest waiting time. Finding (1) appears to be an extension of this idea. In addition, two stations of finite service times should be ordered by having the longer service time station performed first. Again, finding (2) serves to extend the knowledge in sequencing of servers.

Evaluation of each of these factors separately on several series queueing systems, including the case of two exponential servers, showed that variance, σ_s^2, and utilization rate, ρ, did not operate independently but that some relationship existed between the two measures that determined the optimality of sequences of servers. This led to the development of mathematical approximations for optimality of sequences with respect to total waiting time moments and rules of stochastic dominance on the basis of the relationship between σ_s^2 and ρ for each service station in the queueing series.

Mathematical Approximations for Optimal Sequencing

The initial experimentation conducted on sequences of service stations in series queueing systems implied two rules for ordering. (1) When $\rho_1=\rho_2$, sequence A dominates sequence B if $\sigma_{s_1}^2<\sigma_{s_2}^2$. (2) When $\sigma_{s_1}^2=\sigma_{s_2}^2$, sequence A dominates sequence B if $1/\rho_1<1/\rho_2$. As always, in sequence A, station 1 is followed by station 2 and in sequence B station 2 is followed by station 1.

These observations suggested the following hypothesis for consideration:

> If $\alpha_1<\alpha_2$ (where α_1 and α_2 are the ratios σ_s^2/ρ for stations 1 and 2 respectively), then sequence A dominates sequence B with respect to waiting time. If $\alpha_1=\alpha_2$, waiting times are

indifferent to ordering. If $\alpha_1 > \alpha_2$, then sequence B dominates sequence A.

This hypothesis was evaluated in terms of the known results for the series queueing systems under study and found to be true for some systems but not for all. It became apparent that such a simple relationship does not exist for all systems of service stations.

A new hypothesis was formulated that stated:

If $\alpha_1 < f(\alpha_2)$, then sequence A dominates sequence B. If $\alpha_1 = f(\alpha_2)$, then waiting times are indifferent to ordering. If $\alpha_1 > f(\alpha_2)$, then sequence B dominates sequence A. In each case, $f(\alpha_2)$ is some function of α_2.

The investigation on the basis of this hypothesis was carried out by using the results of Fraker,[1] who developed an approximation for the mean waiting time in tandem queueing systems. For an $M/E_{k_1}/1 \rightarrow ./E_{k_2}/1$ system, the total waiting time is given by formulas (2.2) and 2.3) in Chapter II. The technique used to study sequencing involved mathematically finding the relationship between α_1 and α_2 that results in a mean total waiting time which is the same for sequences A and B of the two service stations. Careful investigation of this relationship led to ordering rules on the basis of mean waiting time as well as stochastic dominance.

---

[1] John R. Fraker, "Approximate Techniques for the Analysis of Tandem Queueing Systems," (Ph.D. dissertation, Clemson University, 1971).

Three categories of situations were considered separately in the investigation:

(1) Both stations in the system have service distributions which are different from each other and not exponential.

(2) Both service stations have the same service distribution.

(3) One of the service stations has an exponential service distribution.

(1) Different service distributions.

It was not possible to study in depth many series queueing systems because of the cost of computer experimentation. The system considered was one in which station 1 had an Erlang service distribution with parameter 2 and station 2 had an Erlang service distribution with parameter 3. Some simulation results had already been obtained for this system. For a number of other systems, mathematical derivations were performed that seemed to suggest results similar to those found for the system under study.

For sequence A, the system is $M/E_2/1 \rightarrow ./E_3/1$ and has mean total waiting time

$$\begin{aligned}\bar{w}_{q_A} = (3/4)\left[\rho_1/(1-\rho_1)\mu_1\right] + \left[\lambda/2(1-\rho_2)\right]\Big\{&2/3\mu_2^2 - \rho_1^2/2\mu_2^2 \\ &-(1-\rho_2)\rho_1^2/6\mu_2^2 + 2/3\mu_2^2 - (1-\rho_2)\rho_1^2/18\mu_2^2 \\ &-\left[2(1-\rho_2)\rho_1^2/3\mu_2^2\right]\left[(2-\rho_1^2)/2\right]\Big\}.\end{aligned} \tag{4.1}$$

Sequence B is the system $M/E_3/1 \rightarrow ./E_2/1$ with mean total waiting time

$$\begin{aligned}\bar{w}_{q_B} = (2/3)\left[\rho_2/(1-\rho_2)\mu_2\right] + \left[\lambda/2(1-\rho_1)\right]\Big\{&1/\mu_1^2 - 2\rho_2^2/3\mu_1^2 \\ &-(1-\rho_1)\rho_2^2/3\mu_1^2 + 1/2\mu_1^2 - (1-\rho_1)\rho_2^2/12\mu_1^2 \\ &-\left[2(1-\rho_1)\rho_2^2/3\mu_1^2\right]\left[(3-2\rho_2^2)/3\right]\Big\}.\end{aligned} \tag{4.2}$$

If the expressions are simplified by letting $1/\lambda = 1$, then $1/\mu_1 = \rho_1$ and $1/\mu_2 = \rho_2$. Using these equalities, formulas (4.1) and (4.2) are equated and the result is

$$24(1-\rho_2) - 18(1-\rho_1) + (7+12\rho_1^2-16\rho_2^2)(1-\rho_1)(1-\rho_2) = 0. \quad (4.3)$$

This is a third order equation in ρ_1 and ρ_2 which expresses equal waiting time for both sequences of service stations. In terms of α, this is a representation of the equation $\alpha_1=f(\alpha_2)$ which is stated in the hypothesis.

Table 4 shows some values of ρ_1 and ρ_2 satisfying equation (4.3). By definition $\alpha_i=\sigma_{s_i}^2/\rho_i$ which is equivalent to $(\rho_i/k_i) \times 10^4$. Values of α_1 and α_2 are also reported in this table. In Figure 3 the relationship is graphed in terms of α_1 and α_2. Mathematically analyzing equation (4.3), it was found that the slope of the curve is always positive.[2] Therefore the curve is constantly increasing and divides the area of possible values into two parts as shown in Figure 3.

The relationship between α_1 and α_2 yields indifference between sequences on the basis of mean waiting time. For a given α_2, an α_1 can be found so that the two sequences have the same mean waiting time. Some preliminary tests of this relationship showed agreement with the stated hypothesis. That is, for a given value of α_2, if the actual value of α_1 was less than that predicted for indifference, placing station 1 first in the sequence would yield the

---

[2] In Appendix D this analysis is shown for the general indifference equation for Erlang service distributions.

TABLE 4

SOME VALUES FROM THE WAITING TIME INDIFFERENCE EQUATION FOR A SYSTEM WITH E_2 AND E_3 SERVICE DISTRIBUTIONS

ρ_1	ρ_2	α_1	α_2
0	0.374	0	1247
0.160	0.450	800	1500
0.255	0.500	1275	1667
0.341	0.550	1705	1833
0.423	0.600	2115	2000
0.500	0.649	2500	2163
0.578	0.700	2890	2333
0.652	0.750	3260	2500
0.724	0.800	3620	2667
0.795	0.850	3975	2833
0.864	0.900	4320	3000
0.933	0.950	4665	3167
1.000	1.000	5000	3333

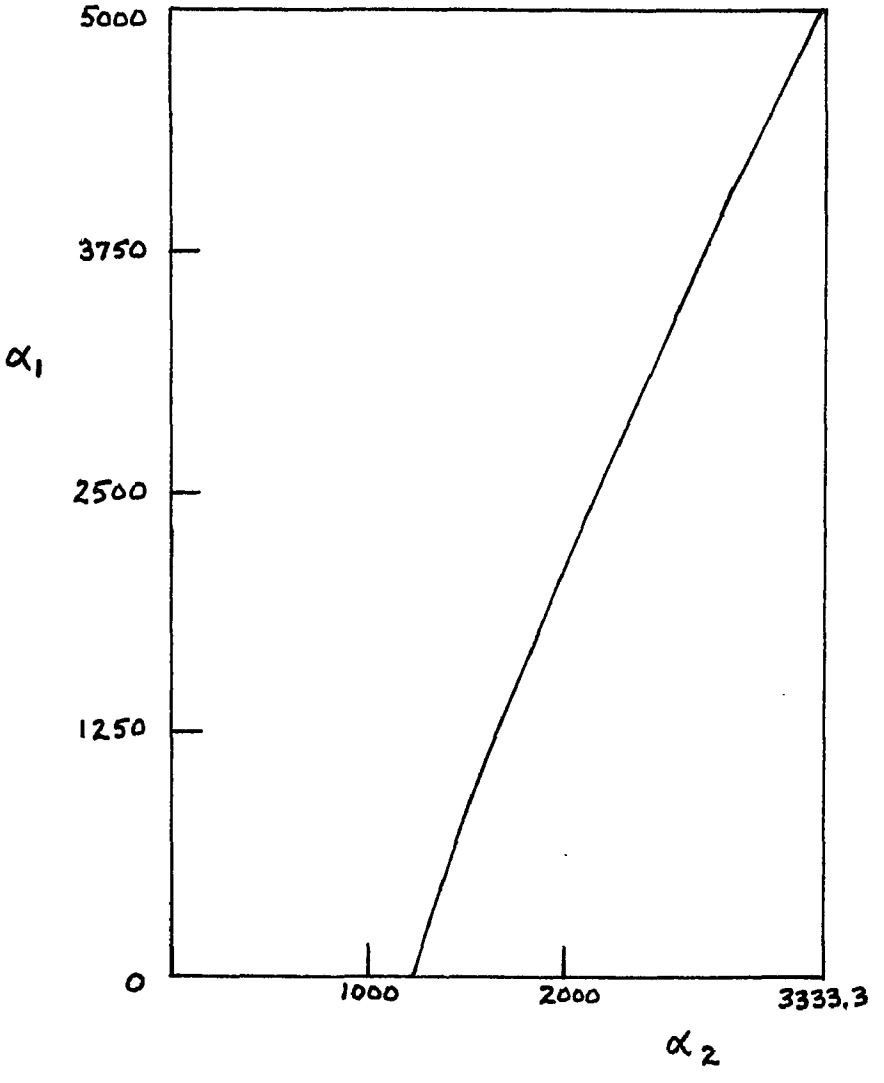


Fig. 3. The Third Order Indifference Equation in α_1 and α_2 for a System with E_2 and E_3 Service Distributions

smaller waiting time. If α_1 was greater than that predicted, station 1 should be second in the sequence to result in the smaller waiting time.

The findings were then extended to the consideration of stochastic dominance. The available evidence from tests showed that for a given α_2, a value of α_1 near the indifferent value resulted in some overlapping of the distribution functions of the two sequences, while a value of α_1 much larger or smaller than the indifferent value resulted in first degree stochastic dominance. Based upon this analysis, it would be logical to set up intervals around the indifference curve. It was theorized that an actual value of α_1, for a given value of α_2, within the interval around the indifference curve, would lead to a situation of second degree stochastic dominance between sequences. A value of α_1 outside this interval would result in first degree stochastic dominance.

At this point the difficulty of working with a third order equation became apparent. It was decided that a straight line approximation of equation (4.3) would prove to be beneficial in conducting further investigation. Firstly, constructing intervals about the linear indifference curve on the basis of statistical analysis is an easy task as compared to working with the third order equation. Also the linear approximation is useful in comparing the indifference relationships for different systems of service stations. Finally, for a particular

system, working with a linear relationship in order to conduct tests on hypotheses is most practical.

Equation (4.3) was approximated by a straight line using the least squares method and extended to α_1 and α_2 for the data in Table 4. The notation and formulas used are from Hays and Winkler.[3] Using the linear regression model, $Y = a + bX$, and letting α_2 be the X_i's and α_1 be the Y_i's, the following coefficients were found:

$$a = -2666.096 \qquad \text{and} \qquad b = 2.341.$$

The slope of this line, b, was significantly different than zero at the .01 level. Therefore, the equation of the line was

$$\alpha_1 = -2666.096 + 2.341\alpha_2.$$

This straight line approximation shows a relationship between α_1 and α_2 that yields an indifference between sequences on the basis of the mean waiting time.

A 95% confidence interval for the actual value of α_1 was constructed based upon the data in Table 4. Thus confidence bands were drawn about the indifference line. The hypothesized areas of first and second degree stochastic dominance between sequences are shown in Figure 4.

The theory was tested where the utilization ratio of the second service station, with Erlang service distribution of parameter 3, was chosen to be 0.75. Therefore,

---

[3]William L. Hays and Robert L. Winkler, Statistics, Vol. II (New York: Holt, Rinehart & Winston, 1970), pp. 1-48.

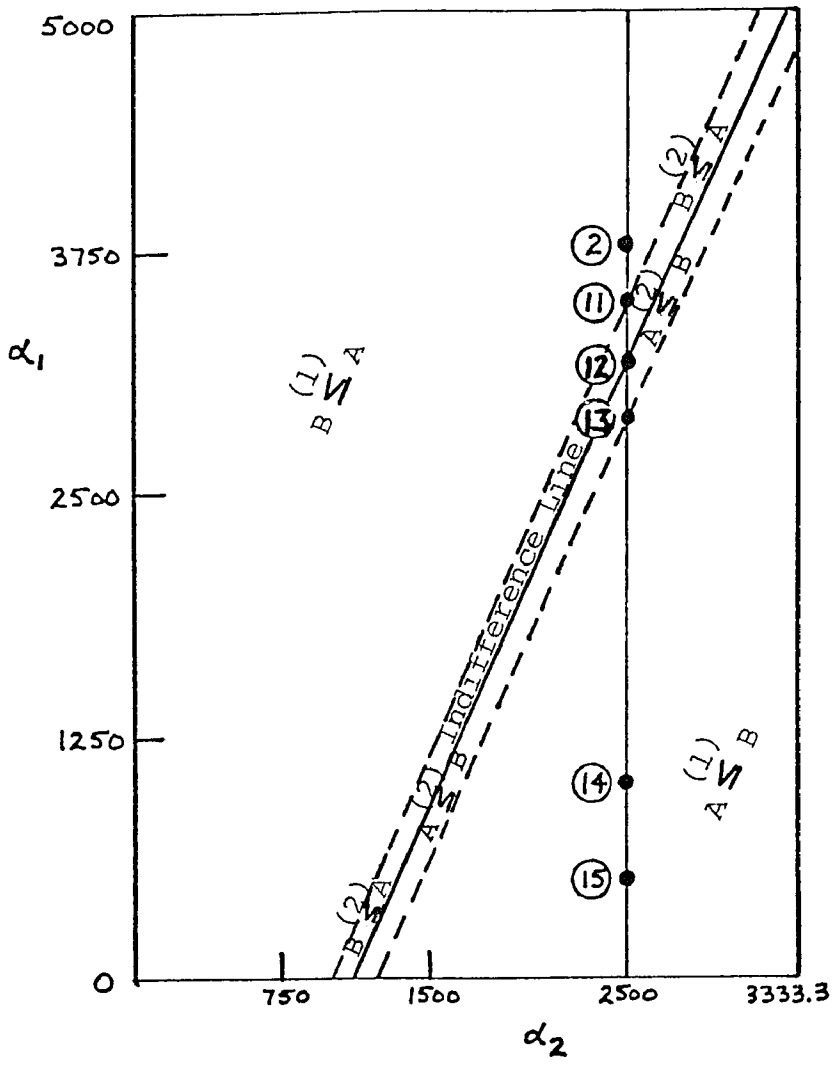


Fig. 4. The Linear Relationship between α_1 and α_2 for a System with E_2 and E_3 Service Distributions

α_2 = 2500. The following experiments were conducted for these values of α_1:

Experiment	$\alpha_1 = \sigma^2_{s_1}/\rho_1$
2	3750.00
11	3420.54
12	3185.20
13	2949.86
14	1000.00
15	500.00

The values of α_1 for each experiment are shown in Figure 4 in relation to the indifference line and confidence bands drawn about that line. These suggest the expected results of the simulation runs for these systems. The actual results of the simulation runs are presented in Table 5 where the t values significantly different than zero at the .01 level are indicated by *. The comparison of the distribution functions for both sequences A and B are displayed in Figure 5.

The results indicate that the method developed for determining optimal sequencing of stations in a series queueing system is valid and useful, even to the extent of predicting the relationship between sequences on the basis of stochastic dominance of waiting times.

The use of an indifference line is actually an approximation of the third order relationship between α_1 and α_2 for the series system. In addition, the use of

TABLE 5

A TEST OF THE RELATIONSHIP BETWEEN α_1 AND α_2 FOR A SYSTEM WITH E_2 AND E_3 SERVICE DISTRIBUTIONS

Experiment	Sequence	System	$\bar{w}_q$	σ_{w_q}	Differences
2	A	$M/E_2/1 \rightarrow ./E_3/1$	285.274	274.873	$\bar{w}_{q(A-B)} = 7.345$
	B	$M/E_3/1 \rightarrow ./E_2/1$	276.877	269.962	$\sigma_{w_{q(A-B)}} = 242.309$
					t=13.488*
11	A	$M/E_2/1 \rightarrow ./E_3/1$	238.751	236.860	$\bar{w}_{q(A-B)} = 5.283$
	B	$M/E_3/1 \rightarrow ./E_2/1$	232.087	233.301	$\sigma_{w_{q(A-B)}} = 208.769$
					t=5.544*
12	A	$M/E_2/1 \rightarrow ./E_3/1$	206.679	217.319	$\bar{w}_{q(A-B)} = 0.783$
	B	$M/E_3/1 \rightarrow ./E_2/1$	205.833	210.655	$\sigma_{w_{q(A-B)}} = 190.956$
					t=1.556
13	A	$M/E_2/1 \rightarrow ./E_3/1$	179.812	210.915	$\bar{w}_{q(A-B)} = -1.051$
	B	$M/E_3/1 \rightarrow ./E_2/1$	180.670	194.471	$\sigma_{w_{q(A-B)}} = 188.990$
					t=-1.218
14	A	$M/E_2/1 \rightarrow ./E_3/1$	146.458	175.293	$\bar{w}_{q(A-B)} = -4.014$
	B	$M/E_3/1 \rightarrow ./E_2/1$	150.141	187.449	$\sigma_{w_{q(A-B)}} = 178.189$
					t=-4.935*
15	A	$M/E_2/1 \rightarrow ./E_3/1$	142.278	180.404	$\bar{w}_{q(A-B)} = -6.235$
	B	$M/E_3/1 \rightarrow ./E_2/1$	148.144	190.688	$\sigma_{w_{q(A-B)}} = 173.925$
					t=-7.854*

*Indicates significance at the 1% level.

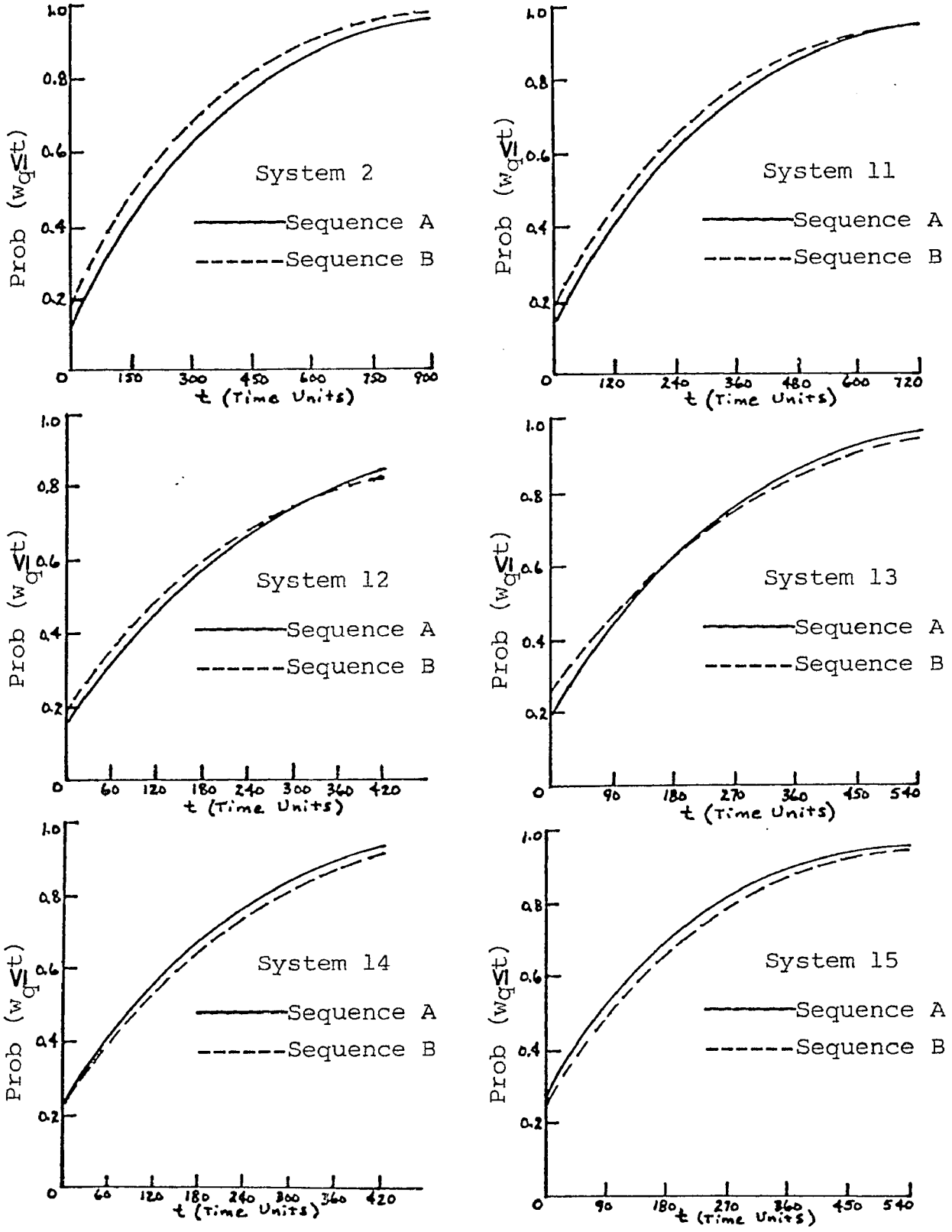


Fig. 5. A Comparison of Cumulative Distributions Under Sequences A and B for a System with E_2 and E_3 Service Distributions with Different α's

confidence bands, and the form of those bands about the indifference line, to represent the separation between regions of first and second order stochastic dominance are issues that require further investigation. They are certainly valuable in developing the relationship between sequences of servers and for predicting an optimal ordering on the basis of total waiting time.

Similar analysis was conducted on a system with an Erlang service distribution with parameter 2 in station 1 and Erlang distribution with parameter 9 in station 2. The equation of indifference for the mean total waiting time was

$$144(1-\rho_2) - 81(1-\rho_1) + (77+72\rho_1^2-128\rho_2^2)(1-\rho_1)(1-\rho_2) = 0. \quad (4.4)$$

Using a straight line approximation in terms of α_1 and α_2 resulted in the indifference relationship

$$\alpha_1 = -5765.053 + 9.849\alpha_2.$$

In Figure 6 this relationship is shown with confidence bands constructed to predict the optimal sequence on the basis of stochastic dominance.

For the system where station 1 had an Erlang service distribution with parameter 3 and station 2 an Erlang distribution with parameter 9, the indifference equation was found to be

$$108(1-\rho_2) - 81(1-\rho_1) + (35+96\rho_1^2-128\rho_2^2)(1-\rho_1)(1-\rho_2) = 0 \quad (4.5)$$

and the linear approximation gave

$$\alpha_1 = -1599.253 + 4.514\alpha_2.$$

Figure 7 shows this system and the range of optimality of each of the sequences in terms of stochastic dominance.

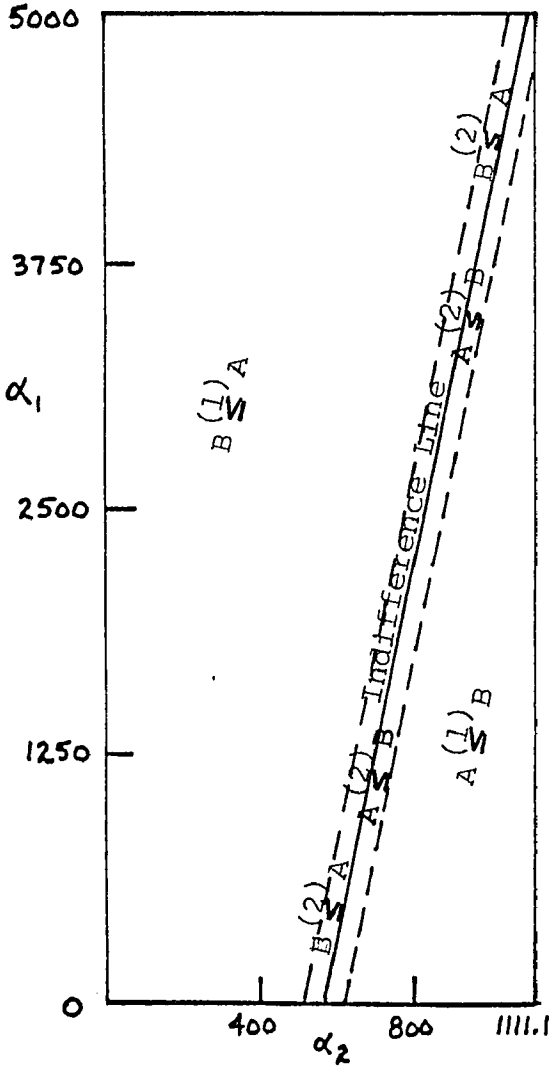


Fig. 6. The Relationship Between α_1 and α_2 for a System with E_2 and E_9 Service Distributions

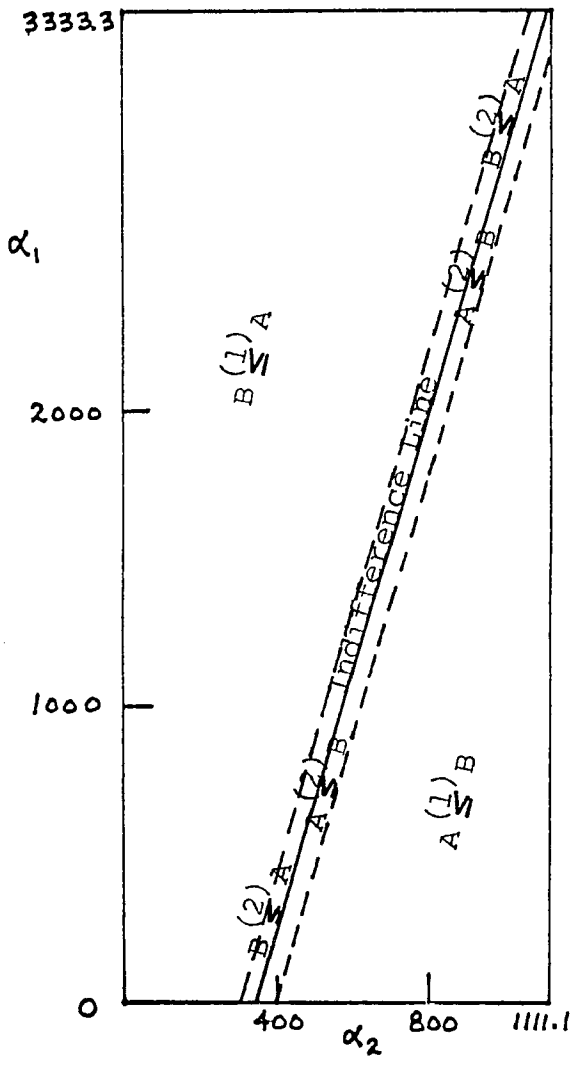


Fig. 7. The Relationship Between α_1 and α_2 for a System with E_3 and E_9 Service Distributions

In comparing the three systems studied, it was found that the slope of the third order indifference curve for each system was always positive and the graph of these equations took on a similar form. In trying to formulate some generalizations for all the systems, the linear approximation for the indifference equation was used. After constructing a 95% confidence interval for the true slope of each indifference equation, no overlapping among the intervals were found to occur. As a result, it seemed that no general expression could be developed to represent all systems of service stations. Although the method of study is the same, each system of service stations in series must be analyzed individually to determine the optimality of sequences.

(2) The same service distributions at both stations.

The situation in which both stations in the system have exponential service distributions is known to be indifferent to ordering with respect to mean waiting time, no matter what the utilization rates are. However this is not true for any other service distributions. Where stations 1 and 2 both had Erlang distributions with parameter 2, the equation for indifference was

$$2(1-\rho_2) - 2(1-\rho_1) + (\rho_1^2-\rho_2^2)(1-\rho_1)(1-\rho_2) = 0.$$

Analysis of this equation yielded the relationship

$$\alpha_1 = \alpha_2 ,$$

a straight line with a slope of +1.

No more investigation of this system was conducted since the results were obvious. Through simulation experimentation, the areas separating first and second degree dominance between the sequences could be determined. The relationship between α_1 and α_2 looks similar to that derived in the previous section where the system was composed of different service distributions. (See Fig. 8.) This applies to all systems where both service stations have the same distribution form.

(3) An exponential service distribution at one station.

Analysis using the mean total waiting time approximation formulas (2.2) and (2.3) proved that if one station in the sequence had an exponential service distribution, it should always be placed second in the system. Considering a system with an exponential distribution for service in station 1 and an Erlang with parameter 2 in station 2, the indifference equation was

$$\rho_1^2 \rho_2^2 (2-\rho_1) = 0.$$

Since the utilization rates were restricted to less than one, no situations existed where the two sequences of servers had equal mean waiting times. Instead, sequence B, with the exponential server last, was always optimal.

This system was analyzed further to search for a possible separation between first and second degree stochastic dominance. It was assumed that second degree

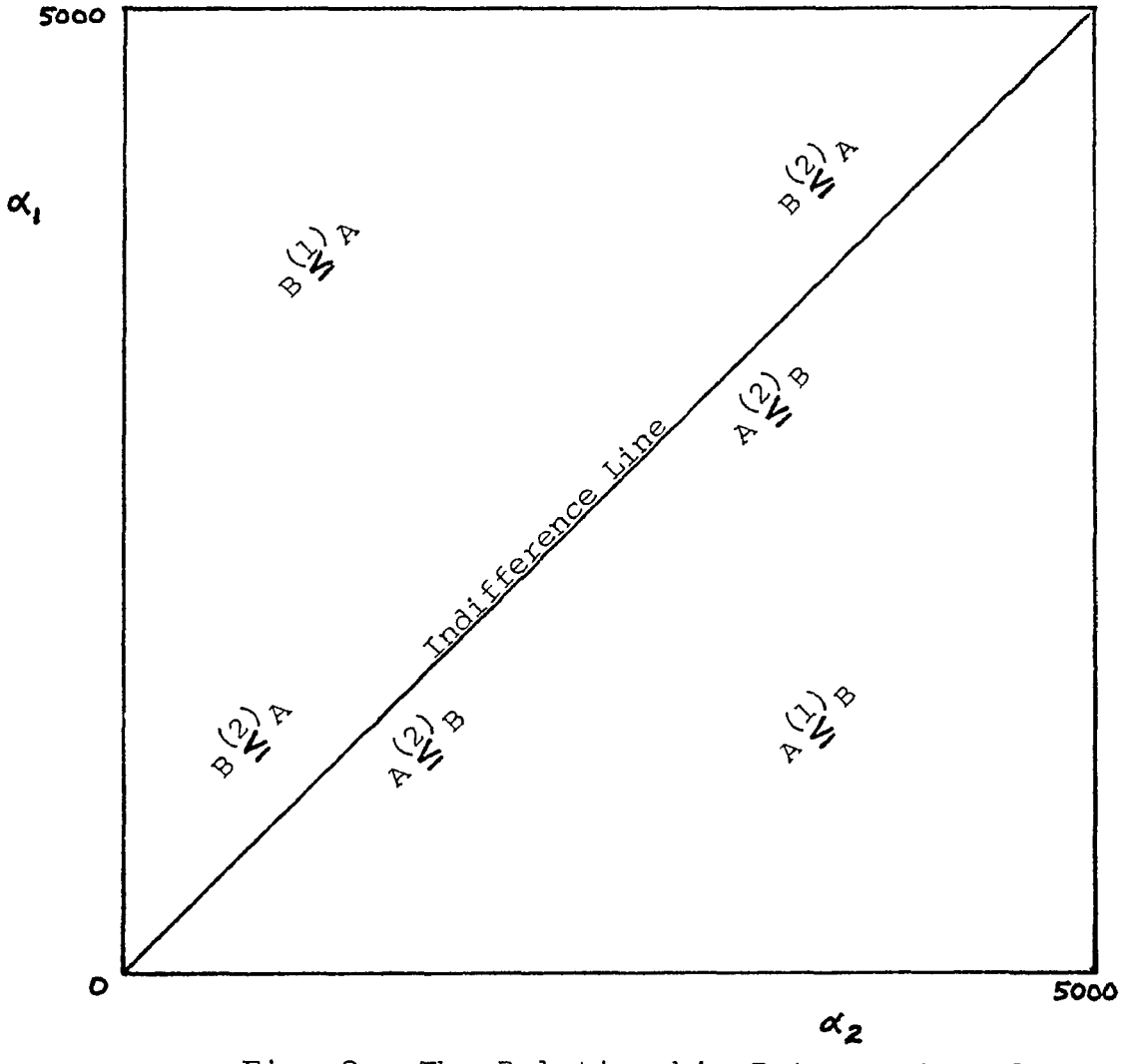


Fig. 8. The Relationship Between α_1 and α_2 for a System with E_2 and E_2 Service Distributions

dominance in the waiting time distribution would result when the difference in mean waiting times between the two sequences was small. Simulation experiments were conducted to determine a range of α_1 and α_2 values that produced small differences in mean waiting times. The relationship used was

$$\bar{w}_{q(A-B)} = \rho_1^2\rho_2^2(2-\rho_1) \leq K\sigma_{w_{q(A-B)}}$$

where K was arbitrarily chosen as 3. Thus differences in mean waiting times that are smaller than three standard deviations were expected to result in second degree dominance between the waiting time distribution functions of the two sequences.

The method of investigation was by computer simulation. Several values of ρ_1 and ρ_2 were chosen and the system simulated. Clearly, low values of ρ_1 and ρ_2 produce the smallest differences in waiting times between the two sequences. Results of the simulation experiments are shown in Table 6. Sequence A represents the ordering M/M/1 -► ./E_2/1 and sequence B is M/E_2/1 -► ./M/1. Included in this table are results of two previous experiments performed on this system. Figure 9 shows the areas of stochastic dominance for this system. Although based on a limited number of experimental results, the idea of estimating second degree stochastic dominance on the basis of small differences in mean waiting times for the two sequences is generally supported by the results found for the distribution functions. There is still a considerable

TABLE 6

DETERMINATION OF THE RELATIONSHIP BETWEEN α_1 AND α_2 FOR A SYSTEM WITH M AND E_2 SERVICE DISTRIBUTIONS

Experiment	ρ_1	ρ_2	α_1	α_2	$\bar{w}_{q(A-B)}$	$\sigma_{w_{q(A-B)}}$	t
16	0.05	0.10	500	500	0.036	5.632	1.71
17	0.05	0.20	500	1000	0.206	14.039	3.94
18	0.05	0.30	500	1500	0.292	26.625	3.29
19	0.10	0.15	1000	750	0.128	11.839	2.90
20	0.15	0.10	1500	500	0.147	14.582	2.70
21	0.15	0.15	1500	750	0.190	16.500	3.09
22	0.20	0.10	2000	500	0.380	22.125	4.61
23	0.20	0.15	2000	750	0.262	23.500	3.34
24	0.20	0.25	2000	1250	0.475	29.687	4.80
5	0.30	$0.30\sqrt{2}$	3000	2121	2.335	64.901	12.77
1	0.75	0.75	7500	3750	31.523	371.123	37.80

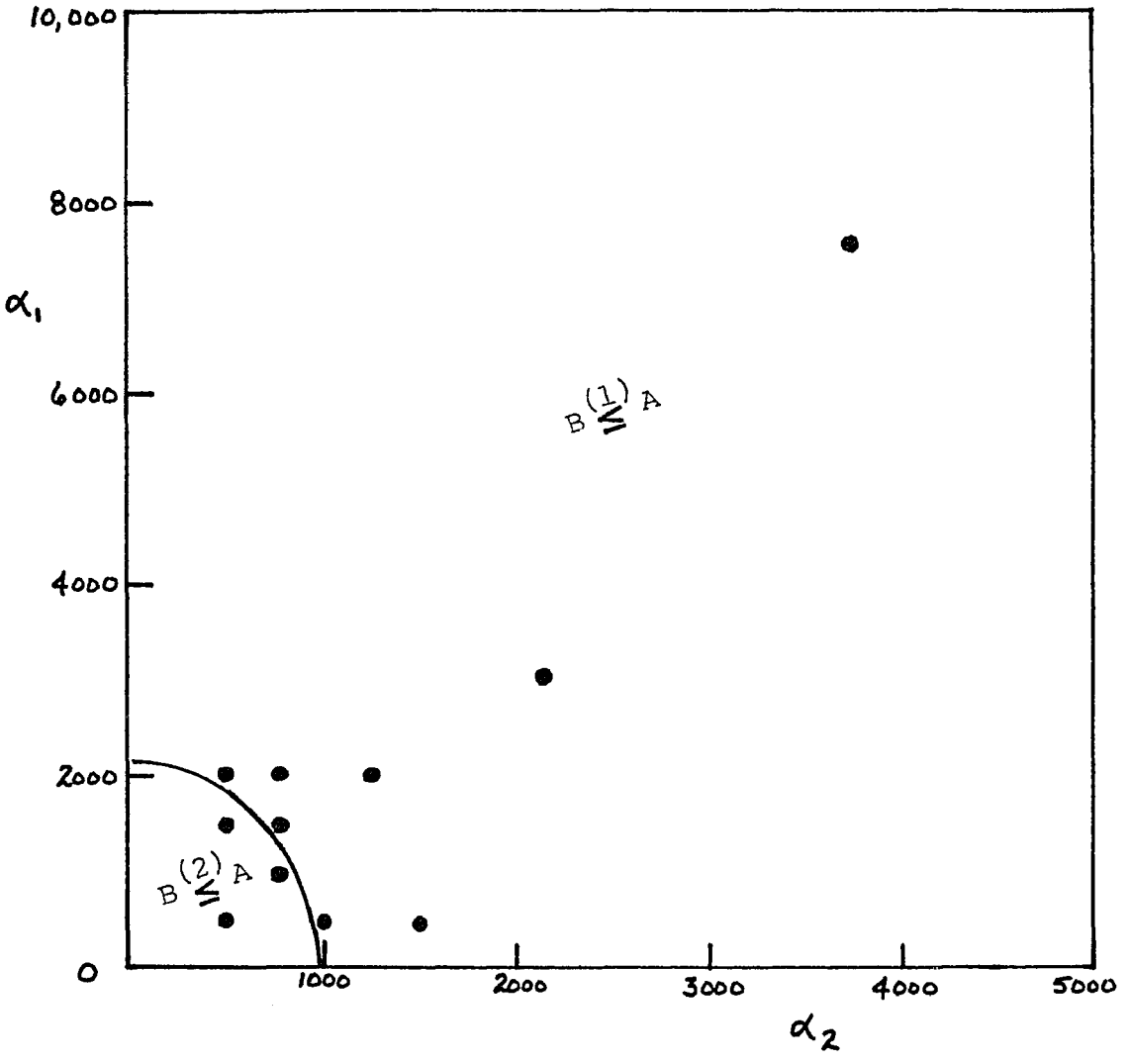


Fig. 9. The Relationship Between α_1 and α_2 for a System with M and E_2 Service Distributions

amount of research that is needed. However, the ability to determine these relationships and predict optimality between sequences on the basis of stochastic dominance is a valuable tool in the study and design of series queueing systems.

CHAPTER V

ISOMORPHIC SYSTEMS

Discussion of Concepts

Isomorphism refers to similarity in form between two substances with respect to certain sets of properties. Isomorphism in queueing processes was introduced by Ghosal[1,2] while studying cybernetic systems. The basic intention of isomorphic research is to develop simpler problems for complicated real-life problems which are isomorphic with respect to certain properties of interest. In this way study of the real-life problem, often impossible because of its complexity, can be conducted on the simpler, isomorphic problem.

Queueing systems have been defined as input-output mechanisms. The input elements include the interarrival process, service process, queue discipline, number of servers, size of waiting room, and any other regulating conditions. The output consists of the departure process, idle time distribution, waiting time and number in the

---

[1] A. Ghosal, Some Aspects of Queueing and Storage Systems, Lecture Notes in Operations Research and Mathematical Systems, vol. 23 (Heidelberg and New York: Springer-Verlag, 1970), pp. 62-66.

[2] A. Ghosal, "Some Problems in Applied Cybernetics," SCIMA 2 (1973):36-40.

queue. The following definitions apply to two queueing systems, Q_1 and Q_2, with X_1 and X_2 the respective input elements to each system, and Y_1 and Y_2 the output elements:

(1) If two systems, with equivalent input characteristics, have equivalent output characteristics, then they are strictly isomorphic. This is expressed as

$$X_1 \frown X_2 \text{ and } Y_1 \frown Y_2$$

where "$\frown$" refers to equivalence with respect to the probability distribution functions of particular characteristics.

(2) If two systems do not have the same input characteristics but are approximately equivalent with respect to at least one output element, then they are isomorphic in the restricted sense. This is expressed as

$$Y_1 \overset{p}{\frown} Y_2$$

where "$\overset{p}{\frown}$" refers to equivalence with repect to the probability distribution functions of one or more, but not all, or the output elements of Y_1 and Y_2.

Statistical tests are necessary to test equivalence of output characteristics. If $F_1(t)$ and $F_2(t)$ are the probability distribution functions from systems Q_1 and Q_2 of some output element, then the two systems are restricted isomorphs if $\left|F_1(t) - F_2(t)\right|$ is not significantly different from zero for all t. This is a failure to reject the hypothesis $F_1(t) = F_2(t)$. Statistical tests can also be performed on their frequency distributions to show similar conclusions.

Analytical work has been limited in this recently developed area of isomorphism. Ghosal[3] reported finding M/M/1 restricted isomorphs of GI/M/1 systems with respect to waiting time and queue size and M/M/1 restricted isomorphs of M/G/1 systems with respect to idle time. Working with a two service station series system, Ghosal[4] was able to develop a technique for finding an M/M/1 restricted isomorph for the waiting time in the second queue and with more difficulty, introduced the possibility of finding $M/E_K/1$ restricted isomorphs of the same output characteristic.

The purpose of this research is to develop methods of determining simpler systems which are isomorphic with the actual complex system. In this study single server queues are sought as restricted isomorphs of various two-server series queueing systems. The particular output characteristic of interest was the total waiting time in both queues. Most series queueing systems do not have known analytical solutions for their waiting time distribution. But if single server queues can be found to approximately represent the waiting time, then decision making involved with the series queueing system can be

---

[3]A. Ghosal at the International Conference on Stochastic Processes, "Isomorphic Queueing Systems," University of Maryland, 1975. (Mimeographed.)

[4]A. Ghosal, ed., "Isomorphic Queueing Systems and Related Problems," Working Paper G 1/76, Graduate Center of Management, Baruch College, New York, 1976.

based on study of the simpler, single server isomorph.

This study developed techniques for finding restricted isomorphs of the complex systems. Statistical tests were used to determine if no significant differences existed in the output of the actual system and its isomorph. If the test was satisfied, general rules were to be found to estimate the parameters of the isomorph in terms of parameters and observations of the complex system. The systems used for study were those same queueing systems simulated in Chapter IV with regard to optimal sequencing of servers. The simulation experiments run for the series queueing systems are referred to using the same identification scheme--A1, A2,..., B1, B2,.... However, since each system was treated individually and order of the servers was disregarded, the notation used to describe a queueing system is:

λ = the mean arrival rate

μ_1 = the mean service rate for the first service station in the series

μ_2 = the mean service rate for the second service station in the series

$s(t)$ = the probability density function of service times

$w_q(t)$ = the probability density function for the total time waiting in the system

$W_q(t)$ = $\text{Prob}(w_q \leq t)$ = the cumulative distribution function for the total time waiting in the system

$H_q(t)$ = $\text{Prob}(w_q > t) = 1 - W_q(t)$

$\bar{w}_q$ = the mean total waiting time in the system

$\sigma^2_{w_q}$ = the variance of the total waiting time in the system

σ_{w_q} = the standard deviation of the total waiting time in the system

$C^2_{w_q}$ = the square of the coefficient of variation (ratio of the variance to the square of the mean total waiting time) for the system

ρ = the utilization factor for the system (equal to the probability of a customer not waiting in any of the queues)

k_1 = the number of phases in the Erlang service distribution for the first service station

k_2 = the number of phases in the Erlang service distribution for the second service station

γ = the weight assigned to a sum of exponentials in a hyperexponential distribution.

In most cases the phrase "isomorph" has been used for "restricted isomorph."

Methods of Determining Single Server Isomorphs

Two methods were employed in finding single station isomorphs of the complex series queueing systems. These two were: (1) a method of fitting distribution functions and (2) a method of fitting moments.

(1) The method of fitting distribution functions

had been suggested by Ghosal.[5] The attempt is to fit the total waiting time distribution of the series queueing system with an isomorphic M/M/1 system. It is known that the waiting time distribution function in an M/M/1 system is

$$W_q(t) = 1 - \rho \exp\left[-(\mu-\lambda)t\right] \qquad \text{where } \rho = \lambda/\mu < 1.$$

If one defines

$$H_q(t) = 1 - W_q(t) = \rho \exp\left[-(\mu-\lambda)t\right]$$

then $$\ln H_q(t) = \ln \rho - (\mu-\lambda)t$$

whose graph is a straight line. Plotting the results of the simulation experiments of series systems may not give $\ln H_q(t)$ that adheres to a straight line pattern, but can be estimated by a straight line using the least squares method of curve fitting.

The straight line approximation represents an M/M/1 isomorph whose parameters can be found from the intercept and slope of the line. The intercept of the line, a, gives ρ from

$$a = \ln \rho$$

and the slope, b, gives μ and λ from

$$b = -(\mu-\lambda).$$

In these studies, the intercept of the line was fixed and set equal to the ρ observed in the series queueing system. The result was that both the M/M/1 isomorph and the original queueing network had exactly the same utilization rate. The slope of the line was found by least squares and used to determine μ and λ of the isomorph.

---

[5] A. Ghosal, "Isomorphic Queueing Systems."

(2) The method of fitting moments attempts to find an M/G/1 isomorph of the series queueing system. In this sense it is more flexible than the method of fitting distribution functions since it is not restricted to finding an M/M/1 isomorphic system. The basic idea is to find an M/G/1 isomorph whose first two moments of the waiting time and utilization rate are equivalent to the moments of the total waiting time and utilization rate of the complex system.

Two general service distributions were considered--the Erlang and hyperexponential. The Erlang distribution with density function

$$s(t) = \left[(k\mu)^k/(k-1)!\right]\exp(-k\mu t)\,t^{k-1} \quad (t\geq 0,\ k \text{ integer} \geq 1)$$

has a mean $1/\mu$ and variance $1/k$ its mean squared. For different values of the parameter k, it represents an infinite number of distributions between the completely regular constant service time ($k=\infty$), and completely random negative exponential ($k=1$). The hyperexponential distribution, as defined by Morse,[6] has a density function

$$s(t) = 2\gamma^2\mu\exp(-2\gamma\mu t) + 2(1-\gamma)^2\mu\exp(-2\left[1-\gamma\right]\mu t)$$

$$(t\geq 0,\ 0<\gamma\leq \tfrac{1}{2})$$

with mean $1/\mu$ and variance $\left\{1+\left[(1-2\gamma)^2/2\gamma(1-\gamma)\right]\right\}/\mu^2$. These distributions are more random than the exponential distribution. This becomes an exponential distribution when $\gamma=\frac{1}{2}$ with variance of $1/\mu^2$. When $\gamma\to 0$ the variance

---

[6]Philip M. Morse, Queues, Inventories and Maintenance (New York: John Wiley & Sons, 1958), pp. 51-55.

approaches infinity.

The moments of the waiting time distribution of these M/G/1 systems can be found from the relationship derived by Riordan.[7] He showed that

$$(1-\rho)nw_q^{(n-1)} = \lambda\sum_{i=2}^{i=n}\binom{n}{i}w_q^{(n-i)}s^{(i)} \qquad \text{where } n>1$$

where $w_q^{(n)}$ is the n^{th} moment of the waiting time distribution, and $s^{(n)}$ is the n^{th} moment of the service distribution.

For the $M/E_k/1$ system where the moment generating function of the service distribution is

$$M_t(z) = E(e^{zt}) = \left[k\mu/(k\mu-z)\right]^k$$

the following relationships can be found:

$$\bar{w}_q = \rho(k+1)/2k\mu(1-\rho)$$

$$\sigma^2_{w_q} = \rho(k+1)(8+4k-5\rho-\rho k)/12k^2\mu^2(1-\rho)^2$$

and

$$C^2_{w_q} = (8+4k-5\rho-\rho k)/3\rho(k+1). \qquad (5.1)$$

For the M/H/1 system, the moment generating function of the service distribution is

$$M_t(z) = (2\gamma^2\mu)/(2\gamma\mu-t) + \left[2(1-\gamma)^2\mu\right]/\left[2(1-\gamma)\mu-t\right]$$

and these relationships are derived:

$$\bar{w}_q = \rho/\left[4\mu\gamma(1-\gamma)(1-\rho)\right]$$

$$\sigma^2_{w_q} = \left[\rho^2+4\rho(1-\rho)(1-2\gamma+2\gamma^2)\right]/4\left[2\gamma(1-\gamma)\right]^2\mu^2(1-\rho)^2$$

and

$$C^2_{w_q} = \left[\rho+4(1-\rho)(1-2\gamma+2\gamma^2)\right]/\rho. \qquad (5.2)$$

From these relationships a set of formulas had been developed by simultaneously solving for the unknown parameters of the isomorph. One must obtain the following

---

[7] John Riordan, Stochastic Service Systems (New York: John Wiley & Sons, 1962), p. 48.

information from the particular series queueing system: the mean total waiting time, $\bar{w}_q$, the square of the coefficient of variation of the waiting time, $C^2_{w_q}$, and the utilization rate, ρ. If an $M/E_k/1$ isomorph is required, λ, μ, and k can be found from these three equations:

$$\lambda = \rho\mu \tag{5.3}$$

$$\mu^3 - (0.375\lambda C^2_{w_q}+1.625\lambda)\mu^2 + (0.375\lambda^2 C^2_{w_q}+0.625\lambda^2-0.25\lambda/\bar{w}_q)\mu + (0.25\lambda^2/\bar{w}_q) = 0 \tag{5.4}$$

and

$$k = \lambda/(2\mu^2\bar{w}_q-2\mu\lambda\bar{w}_q-\lambda). \tag{5.5}$$

If an M/H/1 isomorph is needed, the following equations are used to find λ, μ, and γ:

$$\lambda = \rho\mu \tag{5.6}$$

$$\mu^2 - \left[(3\lambda\bar{w}^2_q+\lambda C^2_{w_q})/4\bar{w}^2_q\right]\mu - (\lambda/2\bar{w}_q) = 0 \tag{5.7}$$

and

$$\gamma = \tfrac{1}{2} - \sqrt{\tfrac{1}{4}-\left[\lambda/4\mu(\mu-\lambda)\bar{w}_q\right]}. \tag{5.8}$$

The work involved in applying this method had been considerably reduced because of the results developed by Sphicas and Shimshak.[8] They studied the coefficient of variation of the waiting time distribution for various M/G/1 systems. For the Erlang and hyperexponential service distributions, they found non-overlapping bounds on this measure. For the $M/E_k/1$ system the limits on $C^2_{w_q}$ are

$$1 + \tfrac{4}{3}\left[(1-\rho)/\rho\right] \leq C^2_{w_q} \leq 1 + 2\left[(1-\rho)/\rho\right]. \tag{5.9}$$

---

[8] Georghios P. Sphicas and Daniel G. Shimshak, "Waiting Time Variability in Some Simple Queueing Systems," Working Paper, Graduate Center of Management, Baruch College, New York, 1976.

For the M/H/1 system these limits are

$$1 + 2\left[(1-\rho)/\rho\right] \leq C_{w_q}^2 \leq 1 + 4\left[(1-\rho)/\rho\right].^9 \qquad (5.10)$$

From the results of each simulated queueing system, it was only necessary to find $C_{w_q}^2$, and knowing the utilization rate of the system, it was possible to determine, by applying these bounds, whether an $M/E_k/1$ or an M/H/1 isomorph existed with equivalent moments and utilization rate. If one did, formulas (5.3) through (5.5) or (5.6) through (5.8) were used to find the parameters of the isomorph.

Results

A total of 20 of the experimental runs on series queueing systems were used in this analysis. These systems were simulated and their pertinent characteristics gathered and analyzed for possible fit by single server isomorphs. The discussion of these systems and the steps involved in their simulation can be found in Chapter III. The results of fitting the series queueing system distribution function to find an M/M/1 isomorph are presented in Table 7. Each system and its isomorph is described by its parameters, mean waiting time, and standard deviation of the waiting time. In Figures 10 through 29, the straight line approximation of $\ln H_q(t)$ for each system is shown. Table 8 summarizes the results of using the method of fitting

---

[9] A derivation of these limits can be found in Appendix E.

TABLE 7

RESULTS FOUND BY METHOD OF FITTING DISTRIBUTION FUNCTIONS

Experiment	Series Queueing System	Isomorphic System
A1	$M/M/1 \rightarrow ./E_2/1$ $1/\lambda$=100; $1/\mu_1$=75.0; $1/\mu_2$=75.0 $\bar{w}_q$ = 397.836 σ_{w_q} = 368.401	M/M/1 $1/\lambda$ 49.29; $1/\mu$=44.12; ρ=0.895 $\bar{w}_q$ = 376.070 σ_{w_q} = 417.828
A2	$M/E_2/1 \rightarrow ./E_3/1$ $1/\lambda$=100; $1/\mu_1$=75.0; $1/\mu_2$=75.0 $\bar{w}_q$ = 285.274 σ_{w_q} = 274.873	M/M/1 $1/\lambda$=45.05; $1/\mu$=39.19; ρ=0.870 $\bar{w}_q$ = 262.272 σ_{w_q} = 298.928
A3	$M/E_3/1 \rightarrow ./E_4/1$ $1/\lambda$=100; $1/\mu_1$=75.0; $1/\mu_2$=75.0 $\bar{w}_q$ = 238.657 σ_{w_q} = 239.536	M/M/1 $1/\lambda$=44.94; $1/\mu$=38.38; ρ=0.854 $\bar{w}_q$ = 224.497 σ_{w_q} = 260.064
A4	$M/E_4/1 \rightarrow ./E_9/1$ $1/\lambda$=100; $1/\mu_1$=75.0; $1/\mu_2$=75.0 $\bar{w}_q$ = 206.094 σ_{w_q} = 211.199	M/M/1 $1/\lambda$=44.37; $1/\mu$=37.18; ρ=0.838 $\bar{w}_q$ = 192.326 σ_{w_q} = 226.494

TABLE 7--Continued

Experiment	Series Queueing System	Isomorphic System
A5	$M/M/1 \rightarrow ./E_2/1$ $1/\lambda=100;\ 1/\mu_1=30.0;\ 1/\mu_2=42.43$ $\bar{w}_q = 36.023$ $\sigma_{w_q} = 57.460$	$M/M/1$ $1/\lambda=73.95;\ 1/\mu=36.53;\ \rho=0.494$ $\bar{w}_q = 35.664$ $\sigma_{w_q} = 62.269$
A6	$M/M/1 \rightarrow ./E_3/1$ $1/\lambda=100;\ 1/\mu_1=30.0;\ 1/\mu_2=51.96$ $\bar{w}_q = 48.840$ $\sigma_{w_q} = 70.656$	$M/M/1$ $1/\lambda=68.82;\ 1/\mu=38.27;\ \rho=0.556$ $\bar{w}_q = 47.924$ $\sigma_{w_q} = 77.222$
A7	$M/M/1 \rightarrow ./E_9/1$ $1/\lambda=100;\ 1/\mu_1=30.0;\ 1/\mu_2=90.0$ $\bar{w}_q = 450.308$ $\sigma_{w_q} = 450.206$	$M/M/1$ $1/\lambda=52.48;\ 1/\mu=47.39;\ \rho=0.903$ $\bar{w}_q = 441.167$ $\sigma_{w_q} = 486.287$
A8	$M/E_2/1 \rightarrow ./E_3/1$ $1/\lambda=100;\ 1/\mu_1=42.43;\ 1/\mu_2=51.96$ $\bar{w}_q = 53.427$ $\sigma_{w_q} = 73.764$	$M/M/1$ $1/\lambda=60.48;\ 1/\mu=35.69;\ \rho=0.590$ $\bar{w}_q = 51.359$ $\sigma_{w_q} = 79.387$

TABLE 7--Continued

Experiment	Series Queueing System	Isomorphic System
A9	$M/E_2/1 \rightarrow ./E_9/1$ $1/\lambda=100$; $1/\mu_1=42.43$; $1/\mu_2=90.0$ $\bar{w}_q = 450.222$ $\sigma_{w_q} = 472.726$	M/M/1 $1/\lambda=52.11$; $1/\mu=47.16$; $\rho=0.905$ $\bar{w}_q = 449.261$ $\sigma_{w_q} = 494.156$
A10	$M/E_3/1 \rightarrow ./E_9/1$ $1/\lambda=100$; $1/\mu_1=51.96$; $1/\mu_2=90.0$ $\bar{w}_q = 458.870$ $\sigma_{w_q} = 466.553$	M/M/1 $1/\lambda=52.22$; $1/\mu=47.31$; $\rho=0.906$ $\bar{w}_q = 455.988$ $\sigma_{w_q} = 501.043$
B1	$M/E_2/1 \rightarrow ./M/1$ $1/\lambda=100$; $1/\mu_1=75.0$; $1/\mu_2=75.0$ $\bar{w}_q = 364.438$ $\sigma_{w_q} = 357.143$	M/M/1 $1/\lambda=50.66$; $1/\mu=44.89$; $\rho=0.886$ $\bar{w}_q = 348.882$ $\sigma_{w_q} = 391.180$
B2	$M/E_3/1 \rightarrow ./E_2/1$ $1/\lambda=100$; $1/\mu_1=75.0$; $1/\mu_2=75.0$ $\bar{w}_q = 276.877$ $\sigma_{w_q} = 269.962$	M/M/1 $1/\lambda=44.84$; $1/\mu=38.92$; $\rho=0.868$ $\bar{w}_q = 255.928$ $\sigma_{w_q} = 292.249$

TABLE 7--Continued

Experiment	Series Queueing System	Isomorphic System
B3	$M/E_4/1 \rightarrow ./E_3/1$ $1/\lambda = 100$; $1/\mu_1 = 75.0$; $1/\mu_2 = 75.0$ $\bar{w}_q = 237.413$ $\sigma_{w_q} = 235.901$	M/M/1 $1/\lambda = 45.08$; $1/\mu = 38.41$; $\rho = 0.852$ $\bar{w}_q = 221.117$ $\sigma_{w_q} = 256.673$
B4	$M/E_9/1 \rightarrow ./E_4/1$ $1/\lambda = 100$; $1/\mu_1 = 75.0$; $1/\mu_2 = 75.0$ $\bar{w}_q = 197.742$ $\sigma_{w_q} = 206.440$	M/M/1 $1/\lambda = 44.40$; $1/\mu = 37.03$; $\rho = 0.834$ $\bar{w}_q = 186.042$ $\sigma_{w_q} = 219.960$
B5	$M/E_2/1 \rightarrow ./M/1$ $1/\lambda = 100$; $1/\mu_1 = 42.43$; $1/\mu_2 = 30.0$ $\bar{w}_q = 33.333$ $\sigma_{w_q} = 53.607$	M/M/1 $1/\lambda = 68.73$; $1/\mu = 33.75$; $\rho = 0.491$ $\bar{w}_q = 32.556$ $\sigma_{w_q} = 57.072$
B6	$M/E_3/1 \rightarrow ./M/1$ $1/\lambda = 100$; $1/\mu_1 = 51.96$; $1/\mu_2 = 30.0$ $\bar{w}_q = 44.044$ $\sigma_{w_q} = 64.894$	M/M/1 $1/\lambda = 61.77$; $1/\mu = 34.28$; $\rho = 0.555$ $\bar{w}_q = 42.754$ $\sigma_{w_q} = 68.989$

TABLE 7--Continued

Experiment	Series Queueing System	Isomorphic System
B7	$M/E_9/1 \rightarrow ./M/1$ $1/\lambda=100$; $1/\mu_1=90.0$; $1/\mu_2=30.0$ $\bar{w}_q = 444.240$ $\sigma_{w_q} = 453.910$	M/M/1 $1/\lambda=51.94$; $1/\mu=46.91$; $\rho=0.903$ $\bar{w}_q = 436.698$ $\sigma_{w_q} = 481.279$
B8	$M/E_3/1 \rightarrow ./E_2/1$ $1/\lambda=100$; $1/\mu_1=51.96$; $1/\mu_2=42.43$ $\bar{w}_q = 51.115$ $\sigma_{w_q} = 71.804$	M/M/1 $1/\lambda=58.63$; $1/\mu=34.36$; $\rho=0.586$ $\bar{w}_q = 48.635$ $\sigma_{w_q} = 75.542$
B9	$M/E_9/1 \rightarrow ./E_2/1$ $1/\lambda=100$; $1/\mu_1=90.0$; $1/\mu_2=42.43$ $\bar{w}_q = 449.971$ $\sigma_{w_q} = 474.471$	M/M/1 $1/\lambda=53.36$; $1/\mu=48.18$; $\rho=0.903$ $\bar{w}_q = 448.521$ $\sigma_{w_q} = 494.379$
B10	$M/E_9/1 \rightarrow ./E_3/1$ $1/\lambda=100$; $1/\mu_1=90.0$; $1/\mu_2=51.96$ $\bar{w}_q = 457.541$ $\sigma_{w_q} = 476.900$	M/M/1 $1/\lambda=53.00$; $1/\mu=47.97$; $\rho=0.905$ $\bar{w}_q = 456.977$ $\sigma_{w_q} = 502.639$

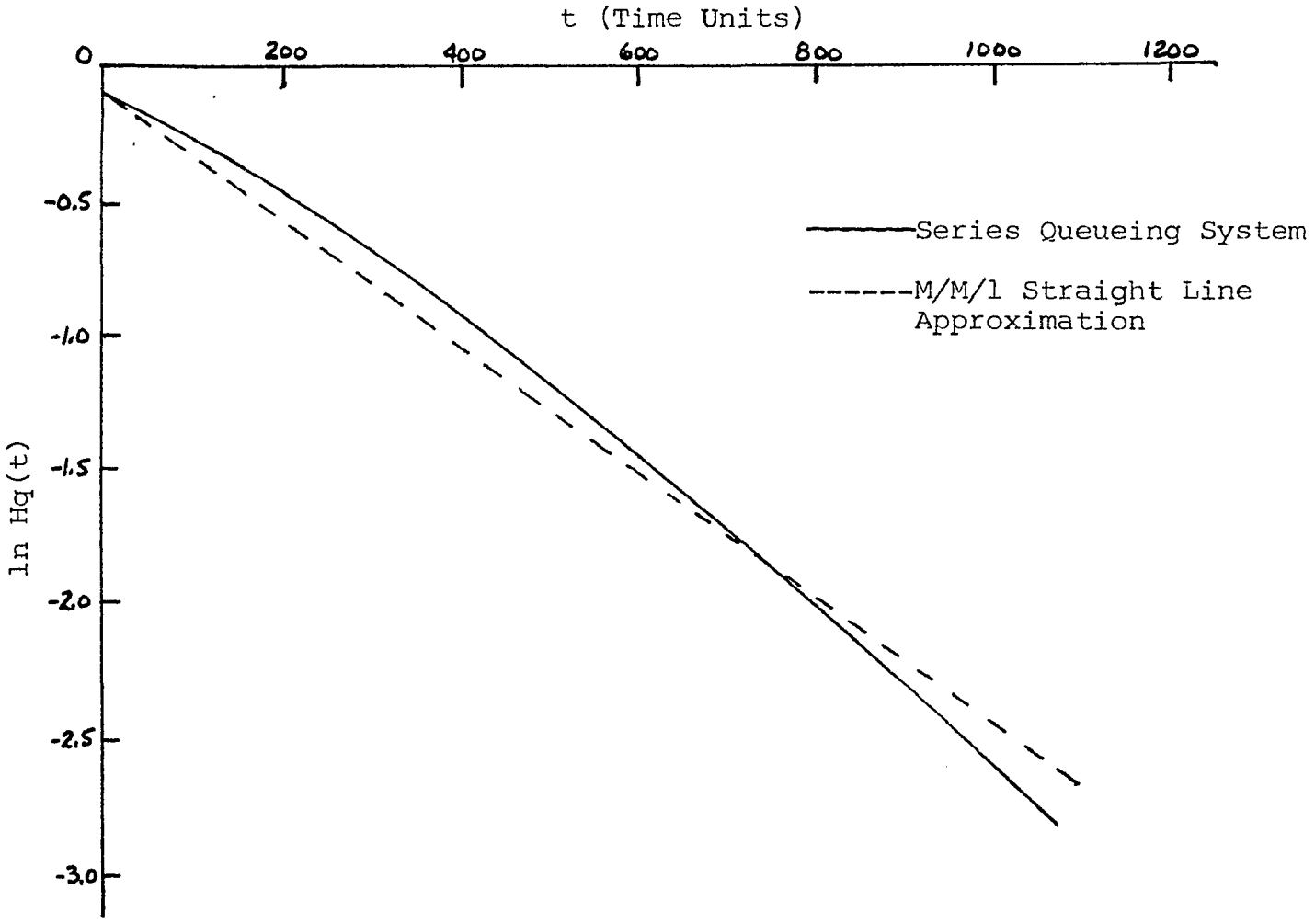


Fig. 10. Determination of M/M/1 Isomorph for Queueing System A1

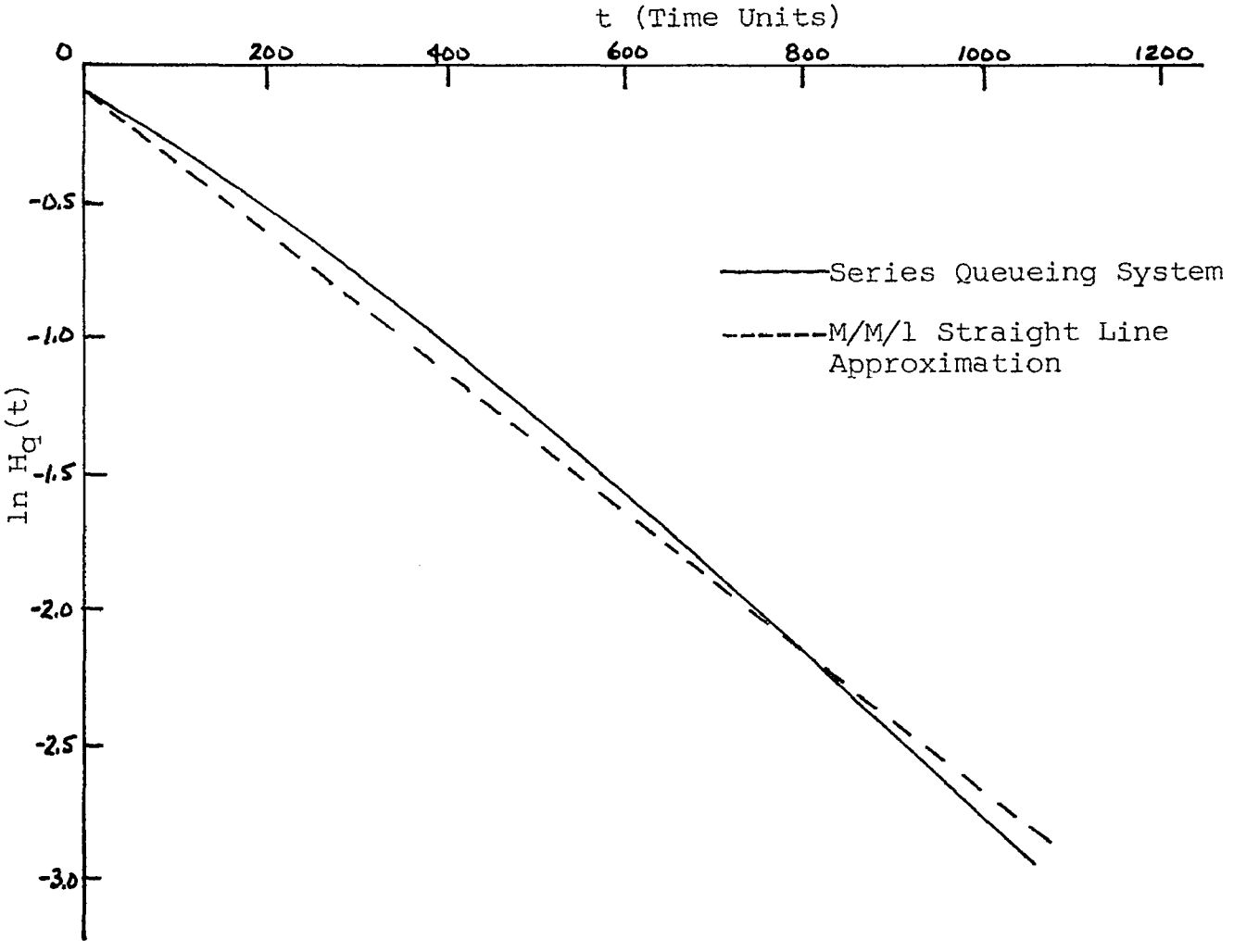


Fig. 11. Determination of M/M/1 Isomorph for Queueing System B1

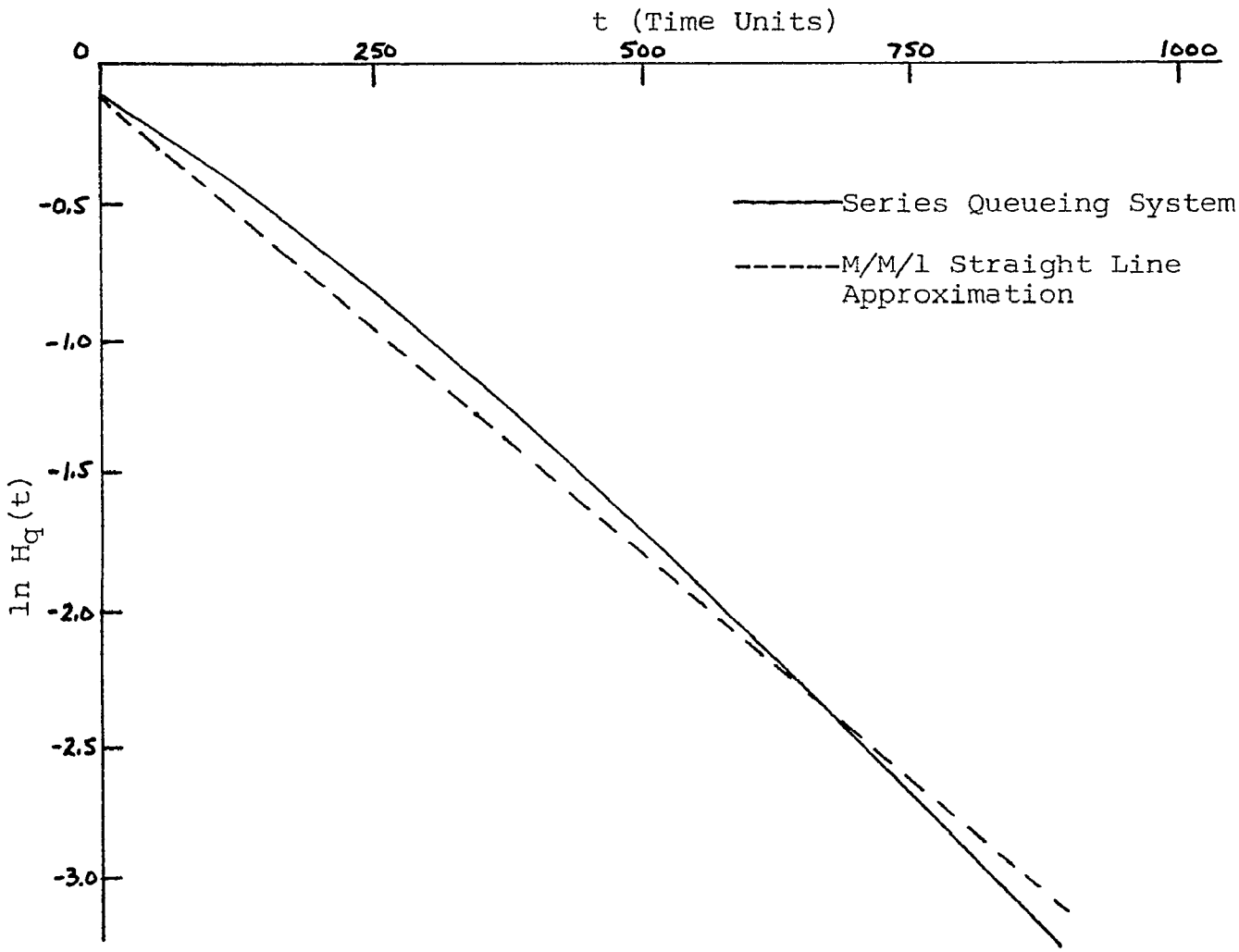


Fig. 12. Determination of M/M/1 Isomorph for Queueing System A2

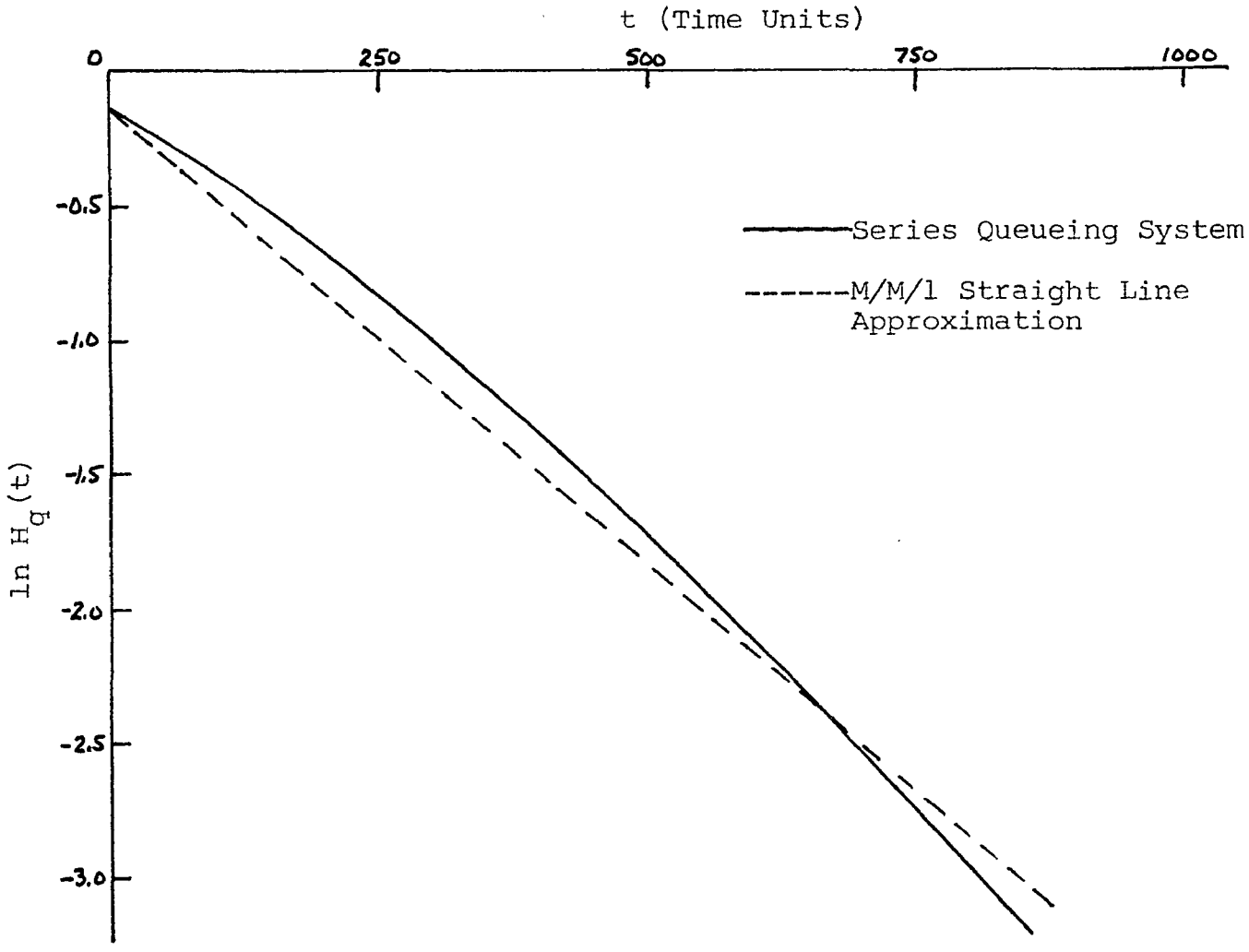


Fig. 13. Determination of M/M/1 Isomorph for Queueing System B2

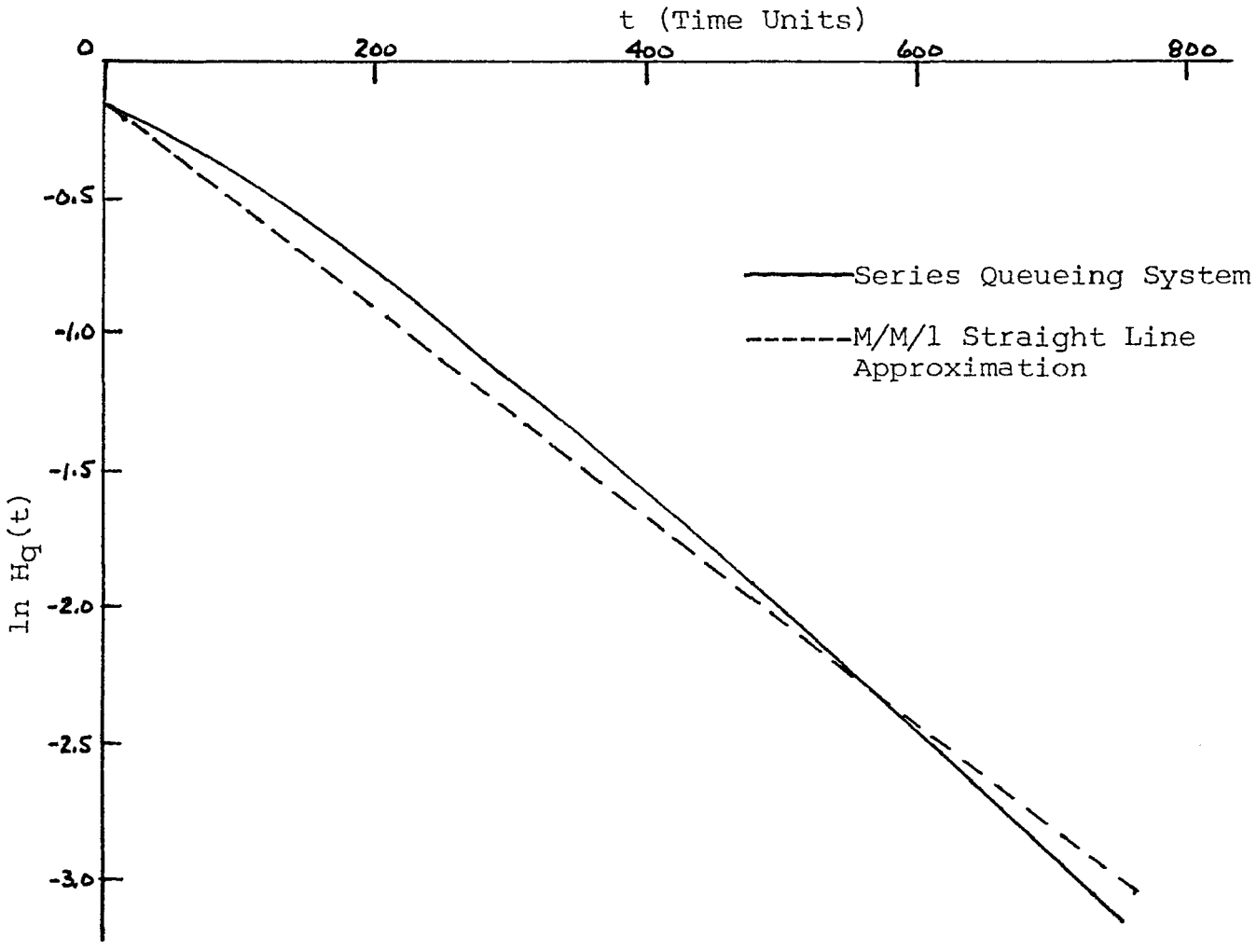


Fig. 14. Determination of M/M/1 Isomorph for Queueing System A3

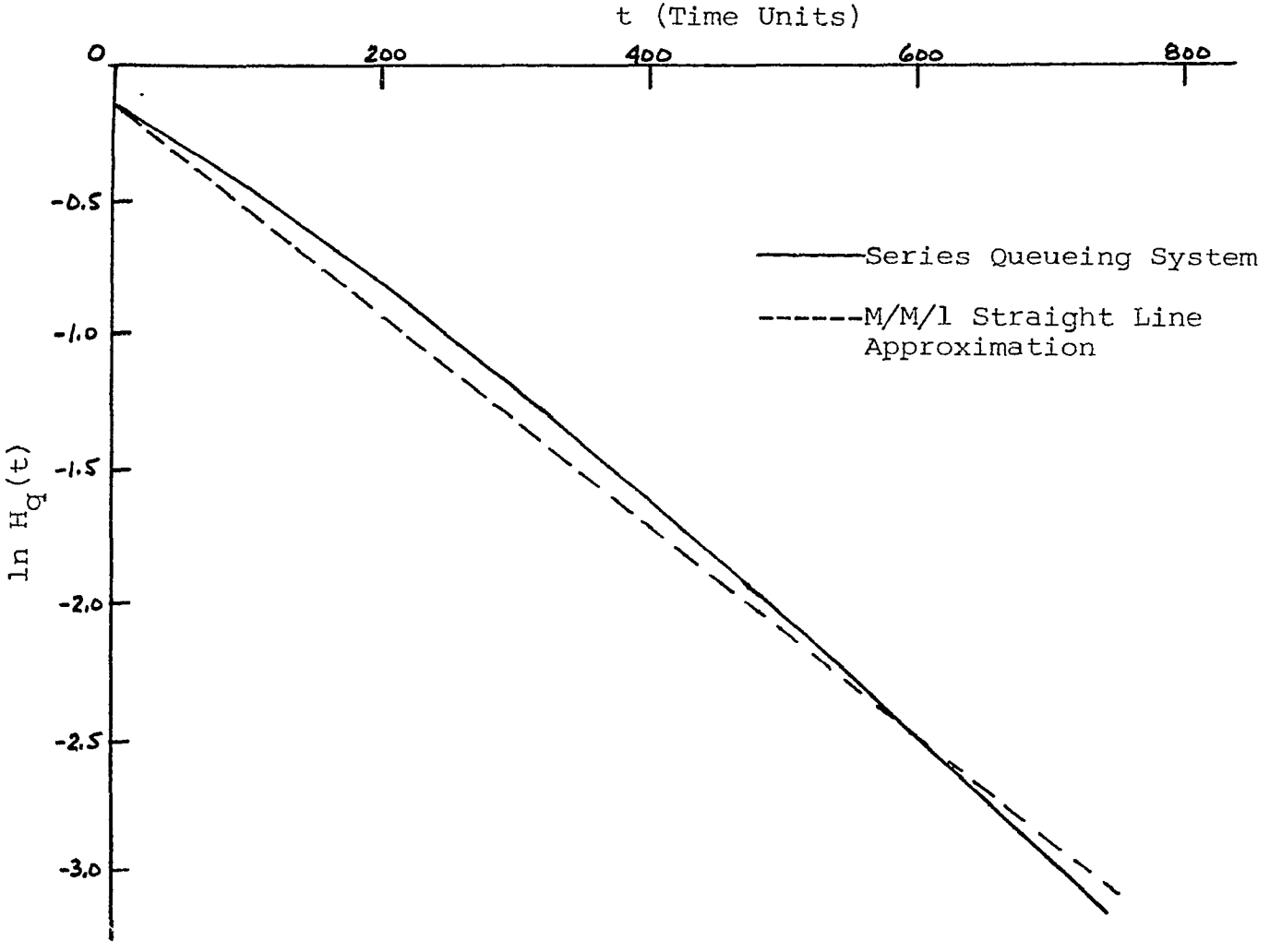


Fig. 15. Determination of M/M/1 Isomorph for Queueing System B3

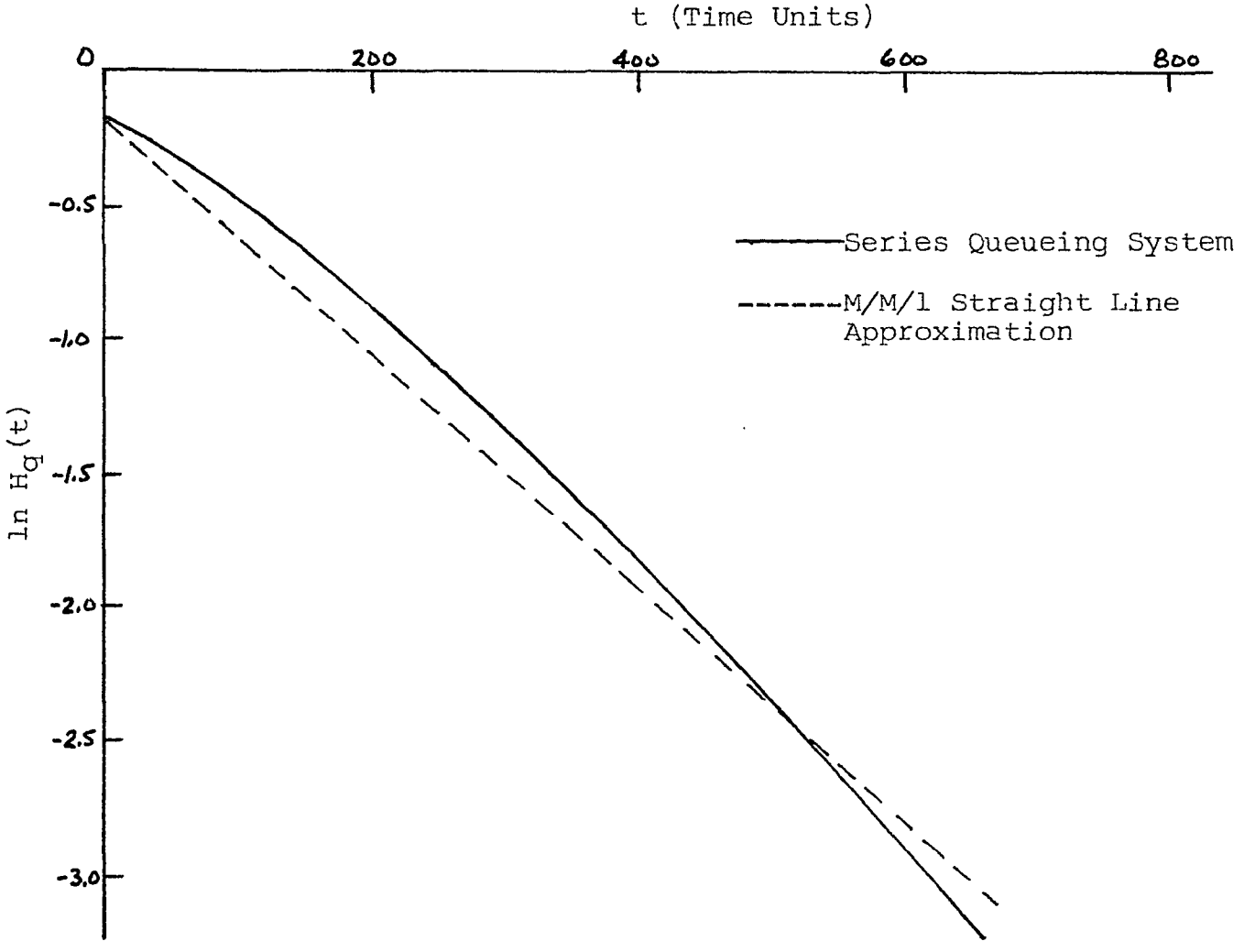


Fig. 16. Determination of M/M/1 Isomorph for Queueing System A4

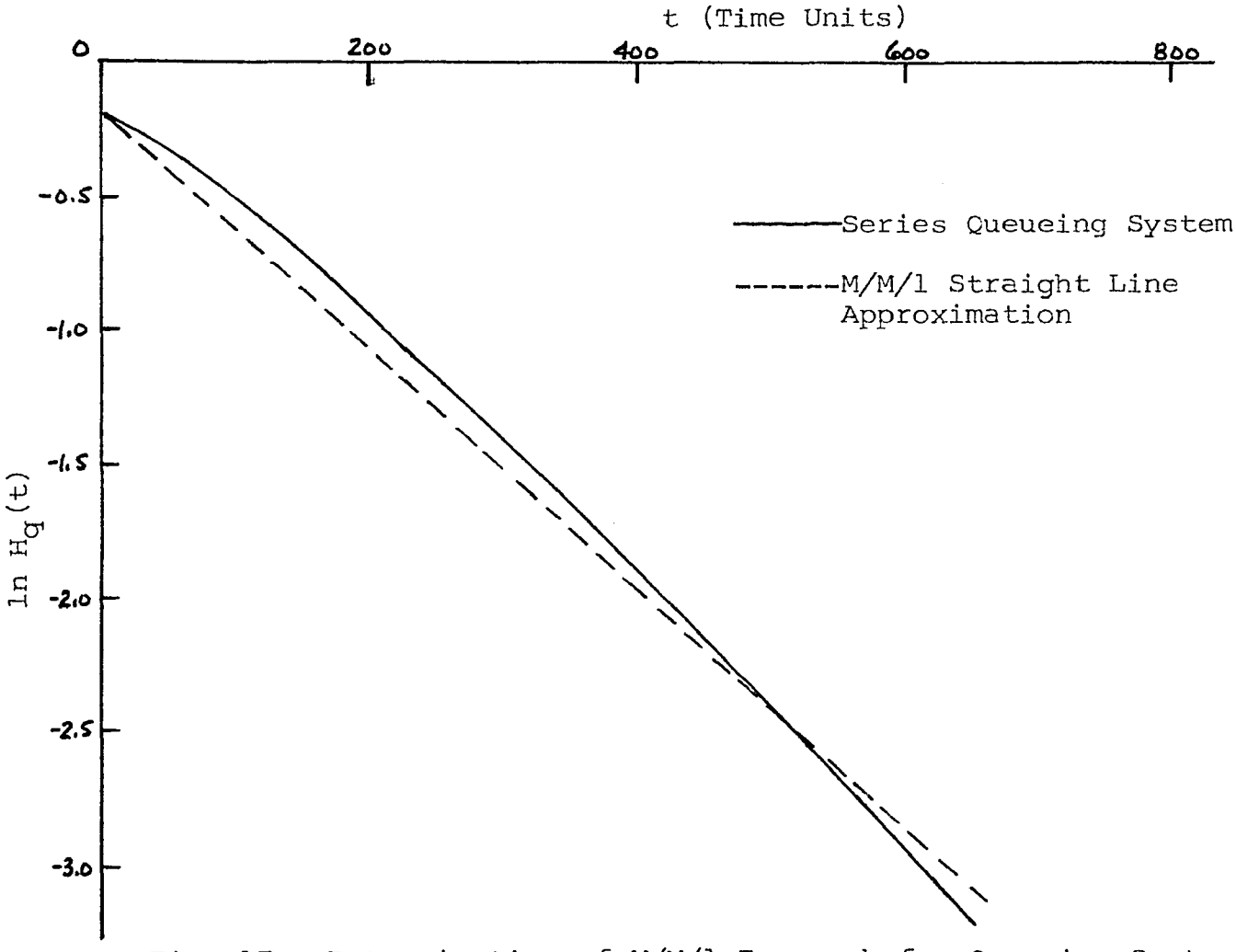


Fig. 17. Determination of M/M/1 Isomorph for Queueing System B4

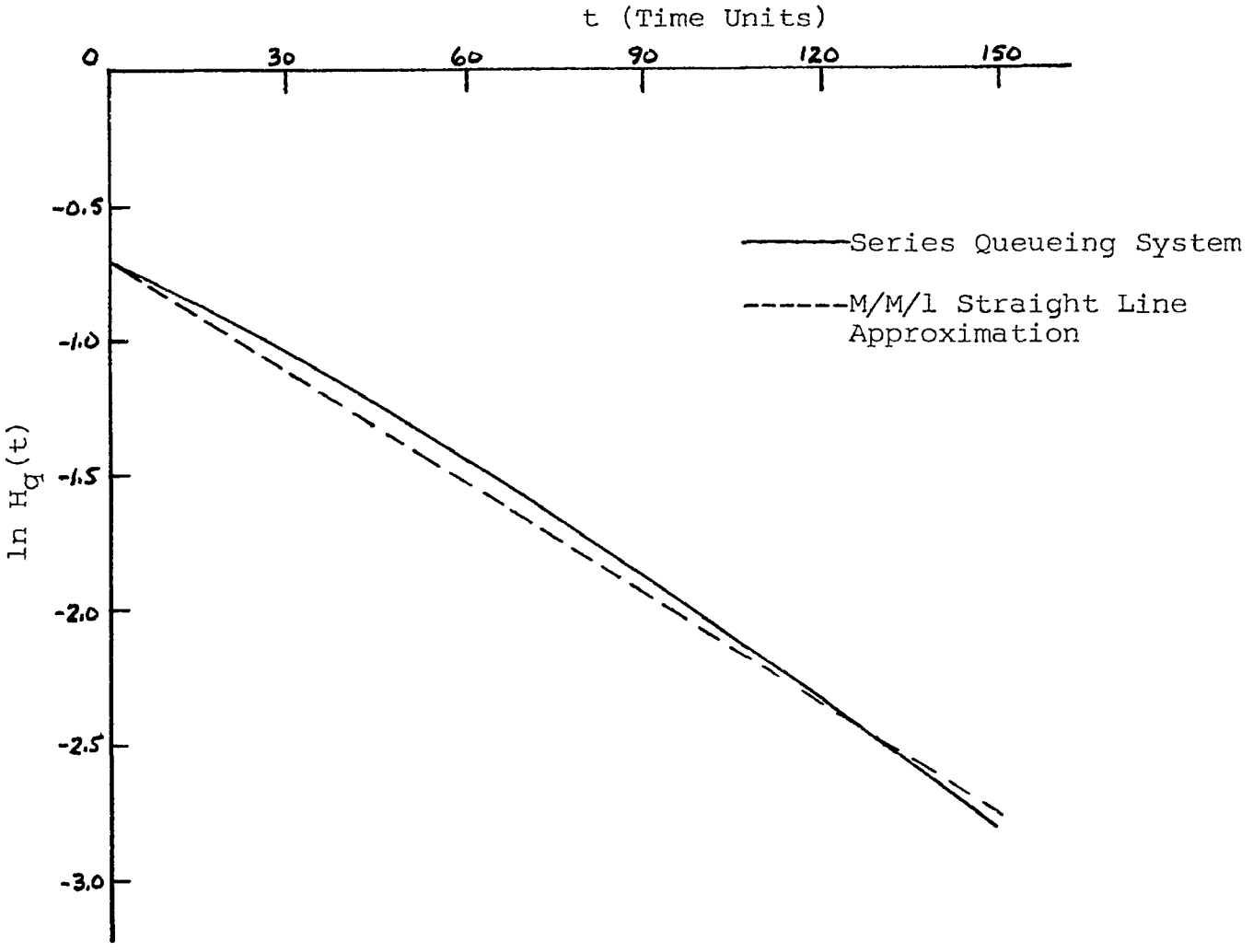


Fig. 18. Determination of M/M/1 Isomorph for Queueing System A5

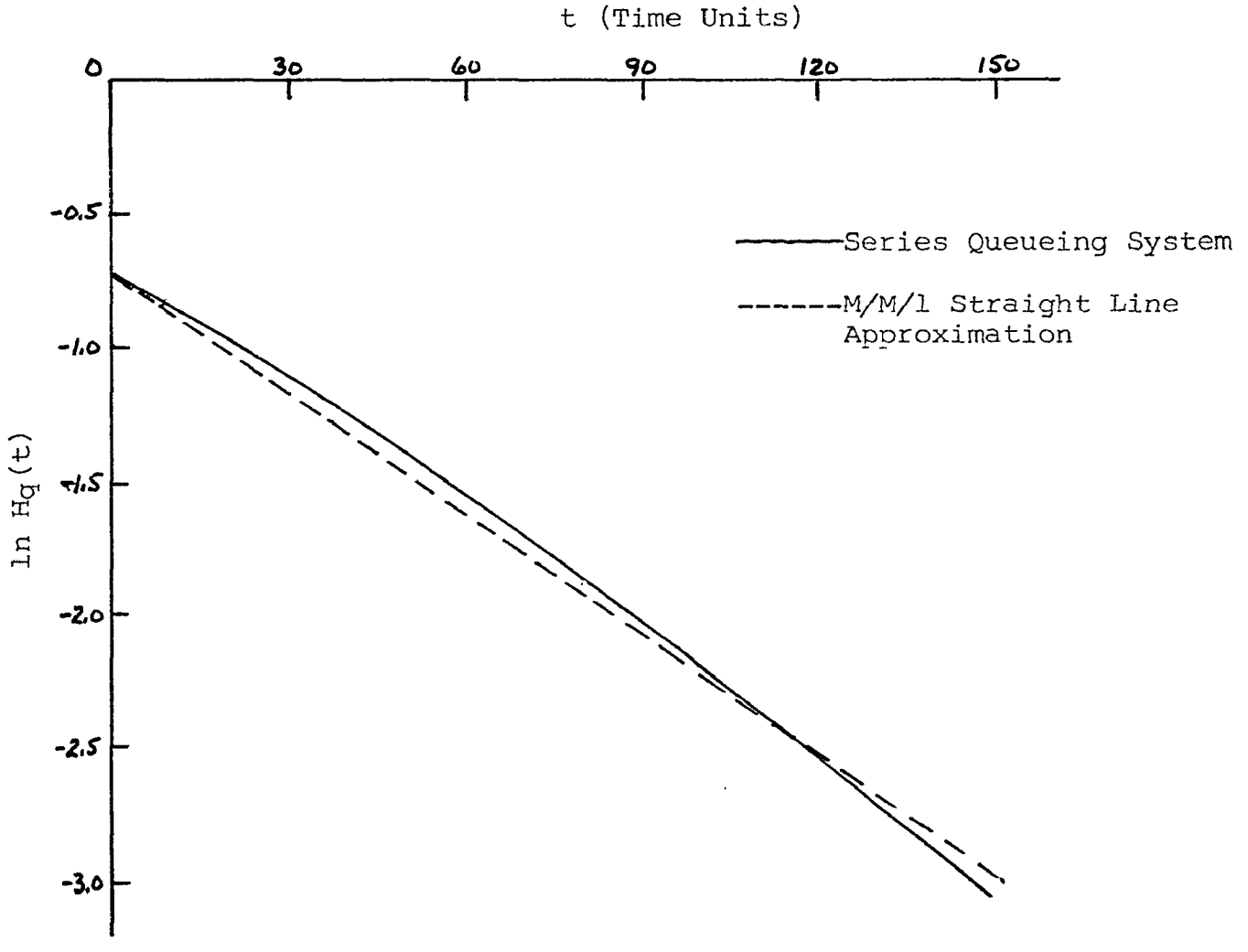


Fig. 19. Determination of M/M/1 Isomorph for Queueing System B5

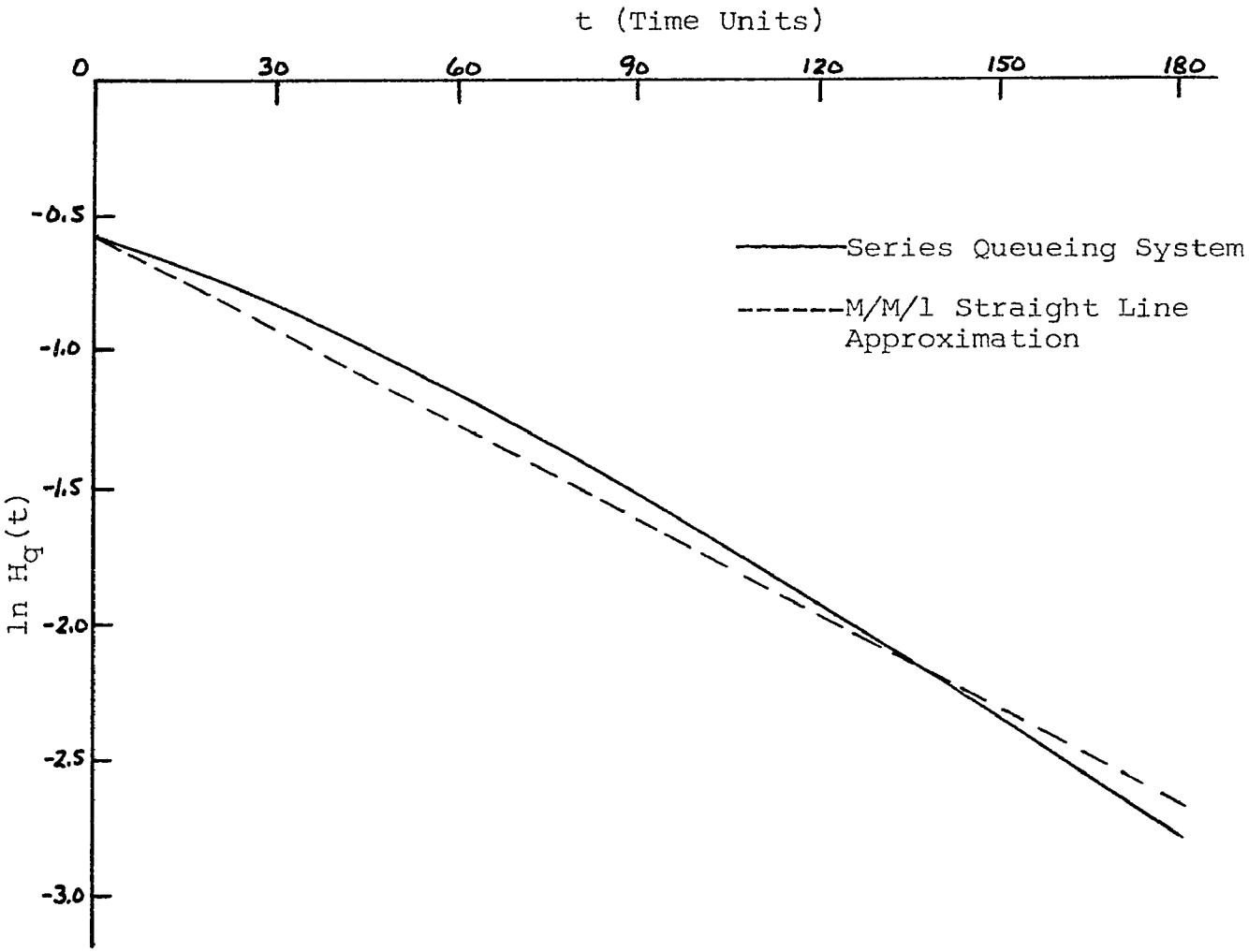


Fig. 20. Determination of M/M/1 Isomorph for Queueing System A6

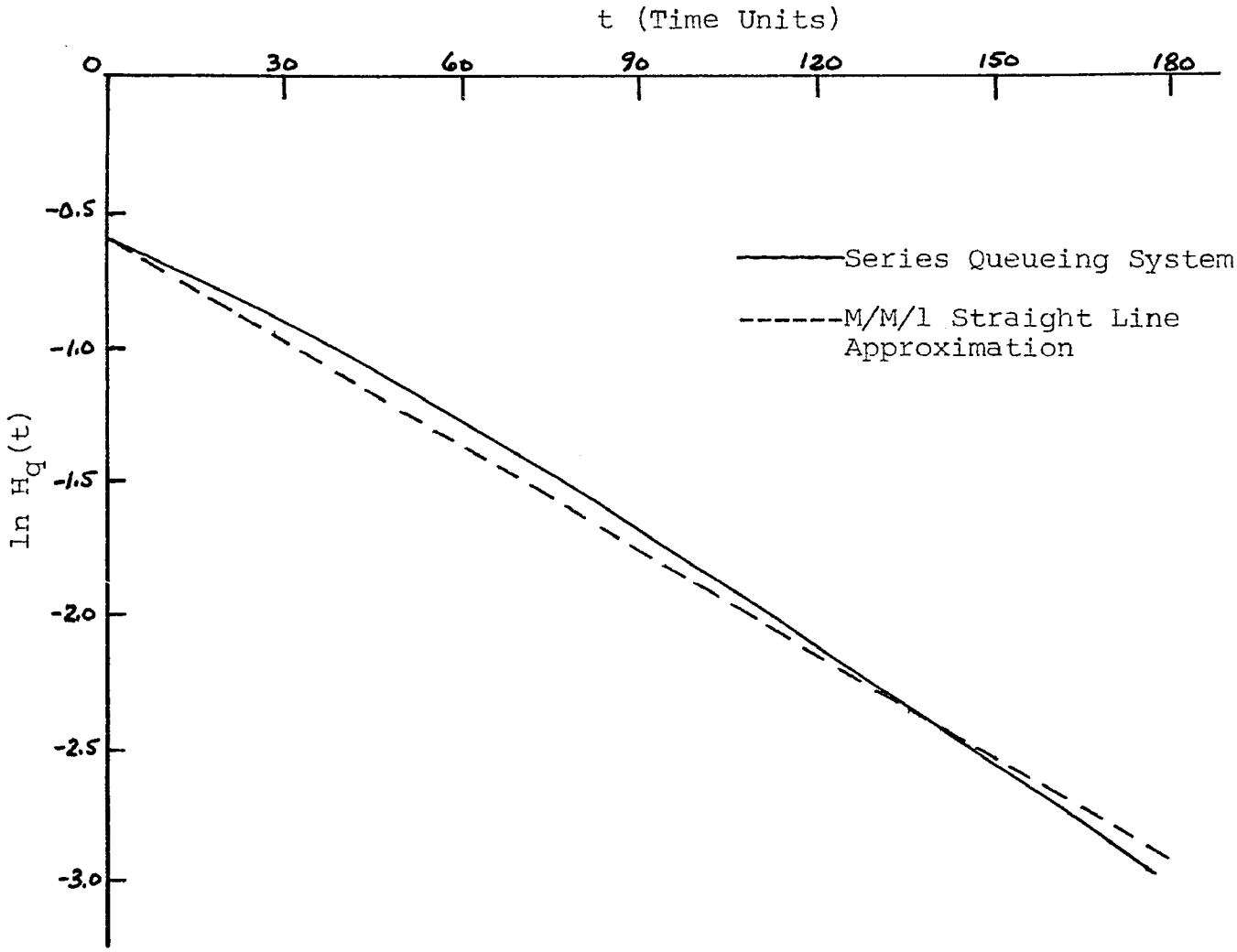


Fig. 21. Determination of M/M/1 Isomorph for Queueing System B6

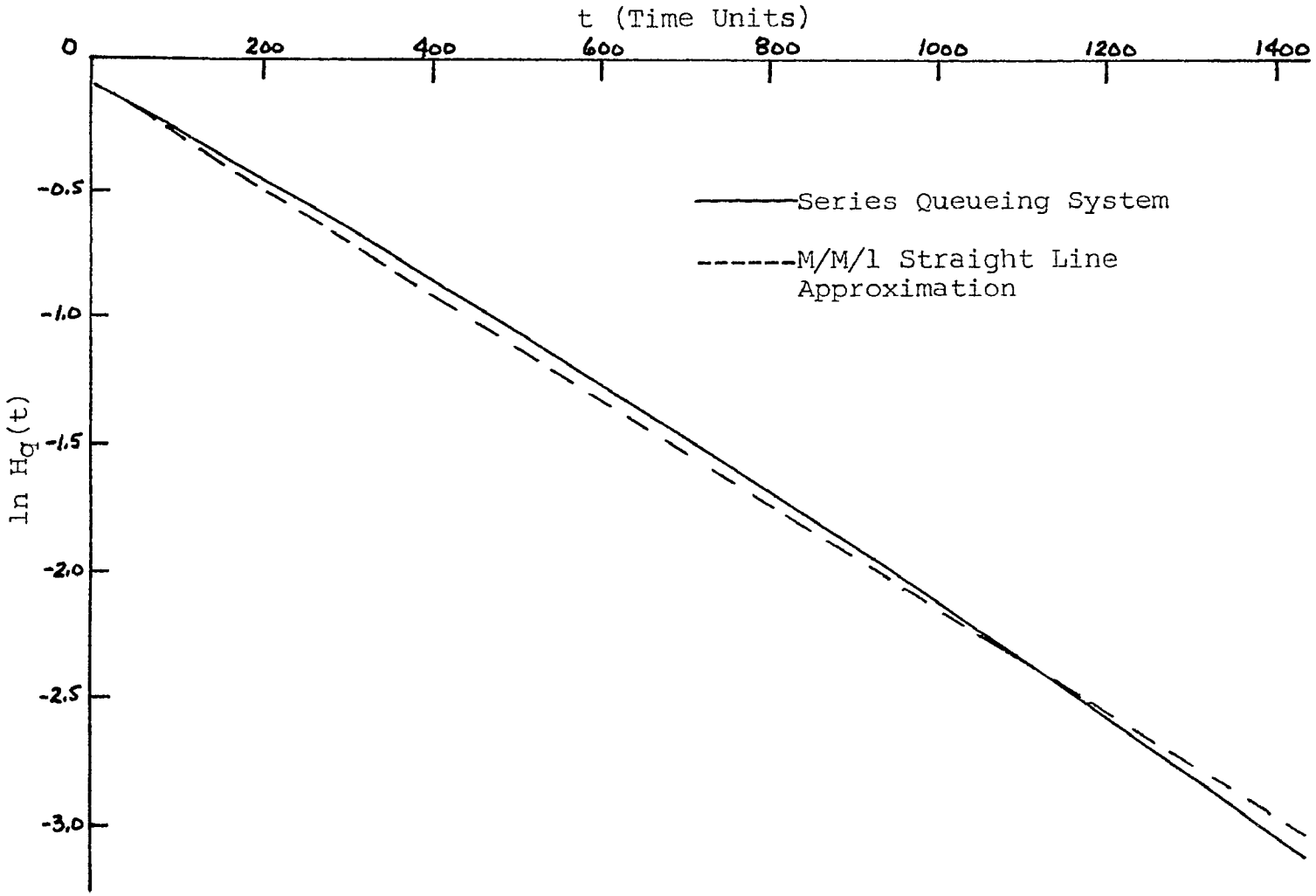


Fig. 22. Determination of M/M/1 Isomorph for Queueing System A7

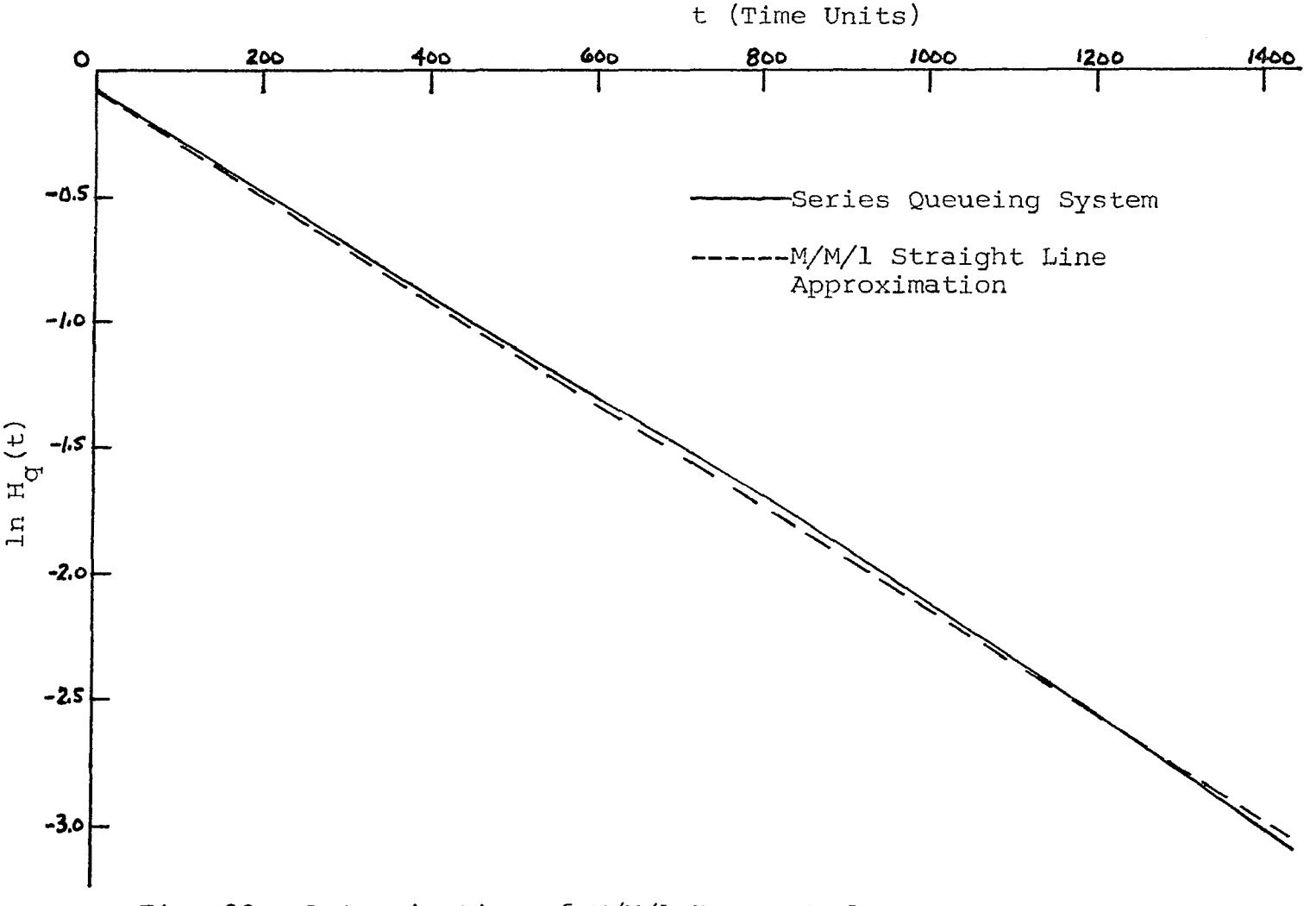


Fig. 23. Determination of M/M/1 Isomorph for Queueing System B7

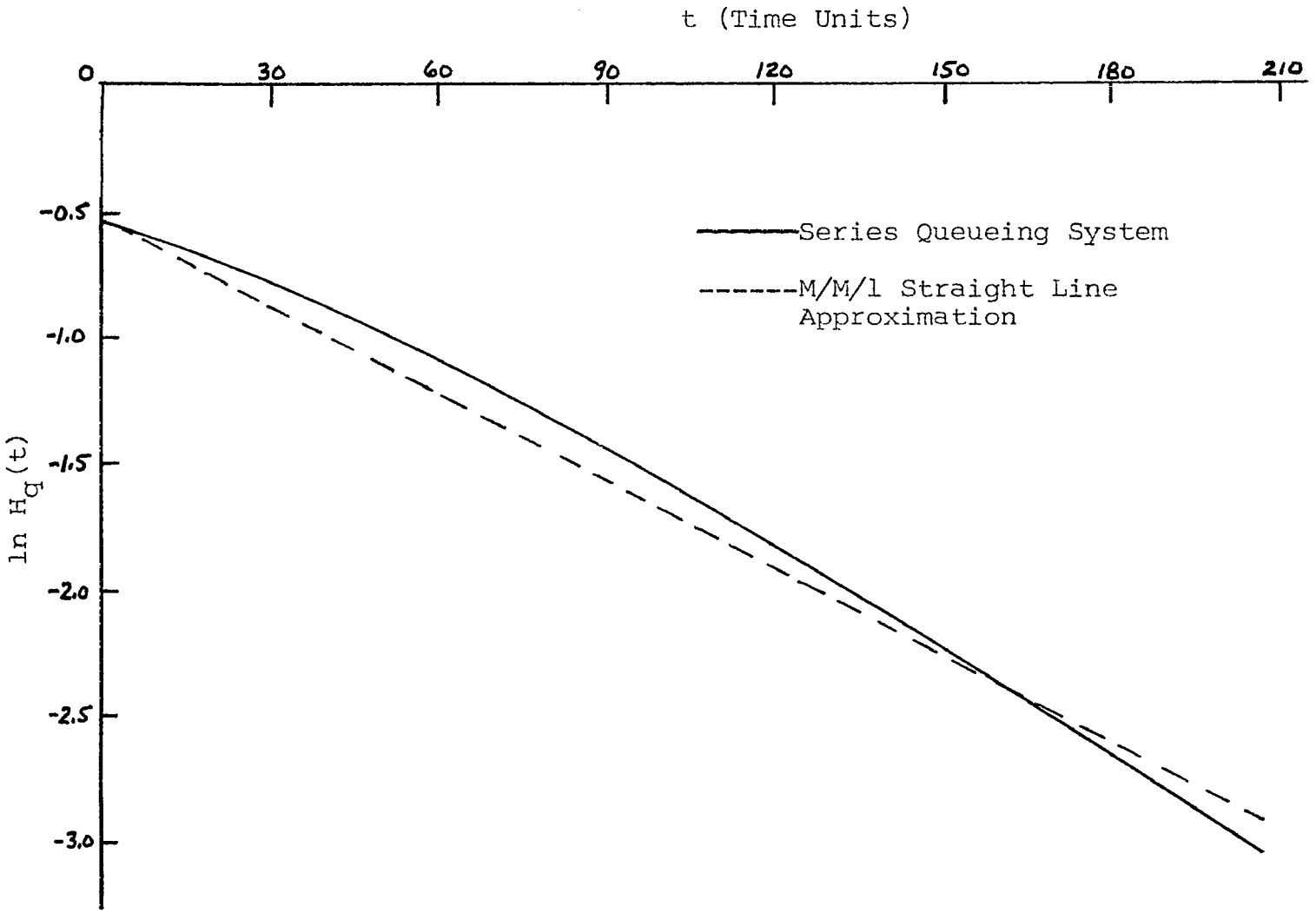


Fig. 24. Determination of M/M/1 Isomorph for Queueing System A8

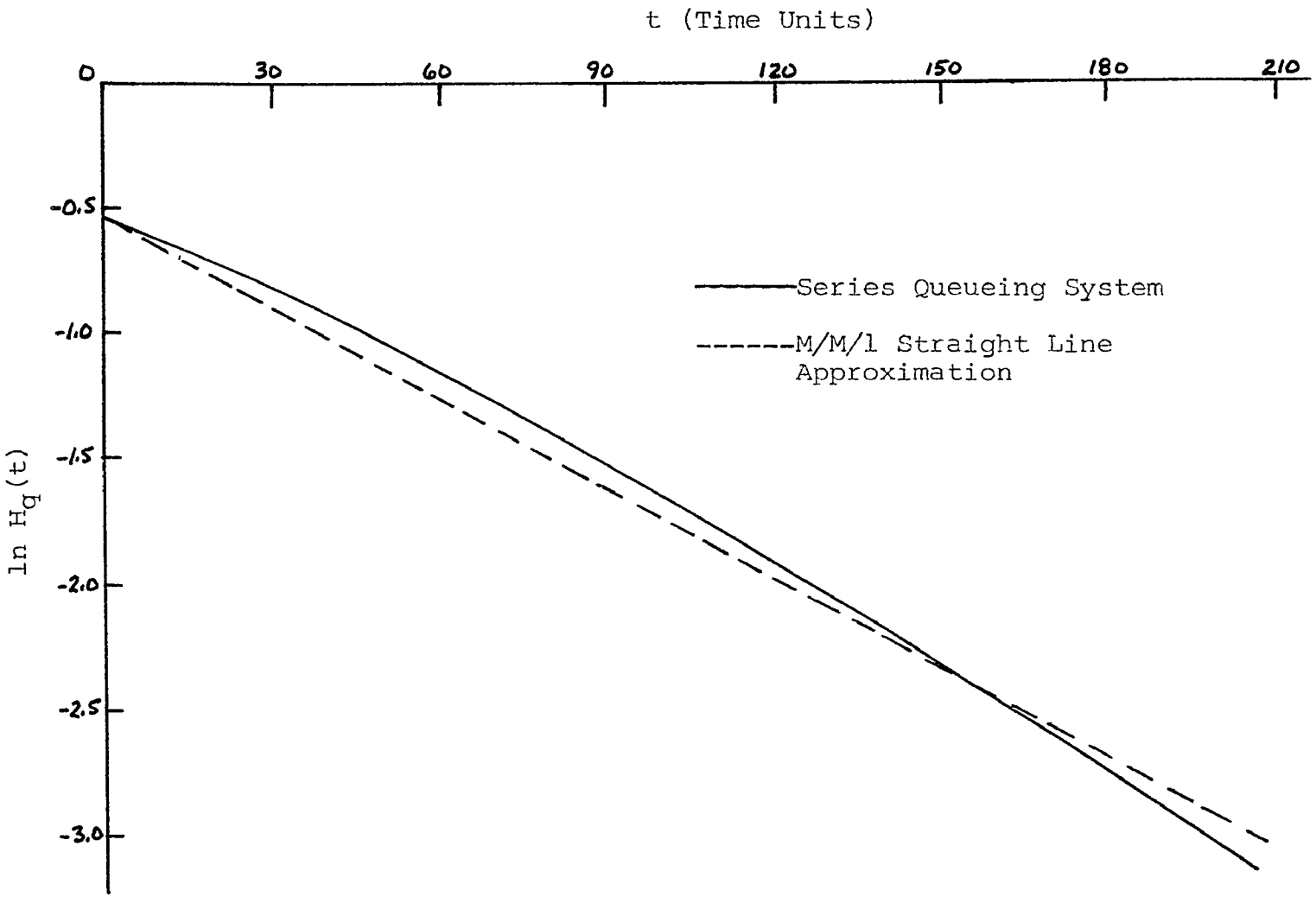


Fig. 25. Determination of M/M/1 Isomorph for Queueing System B8

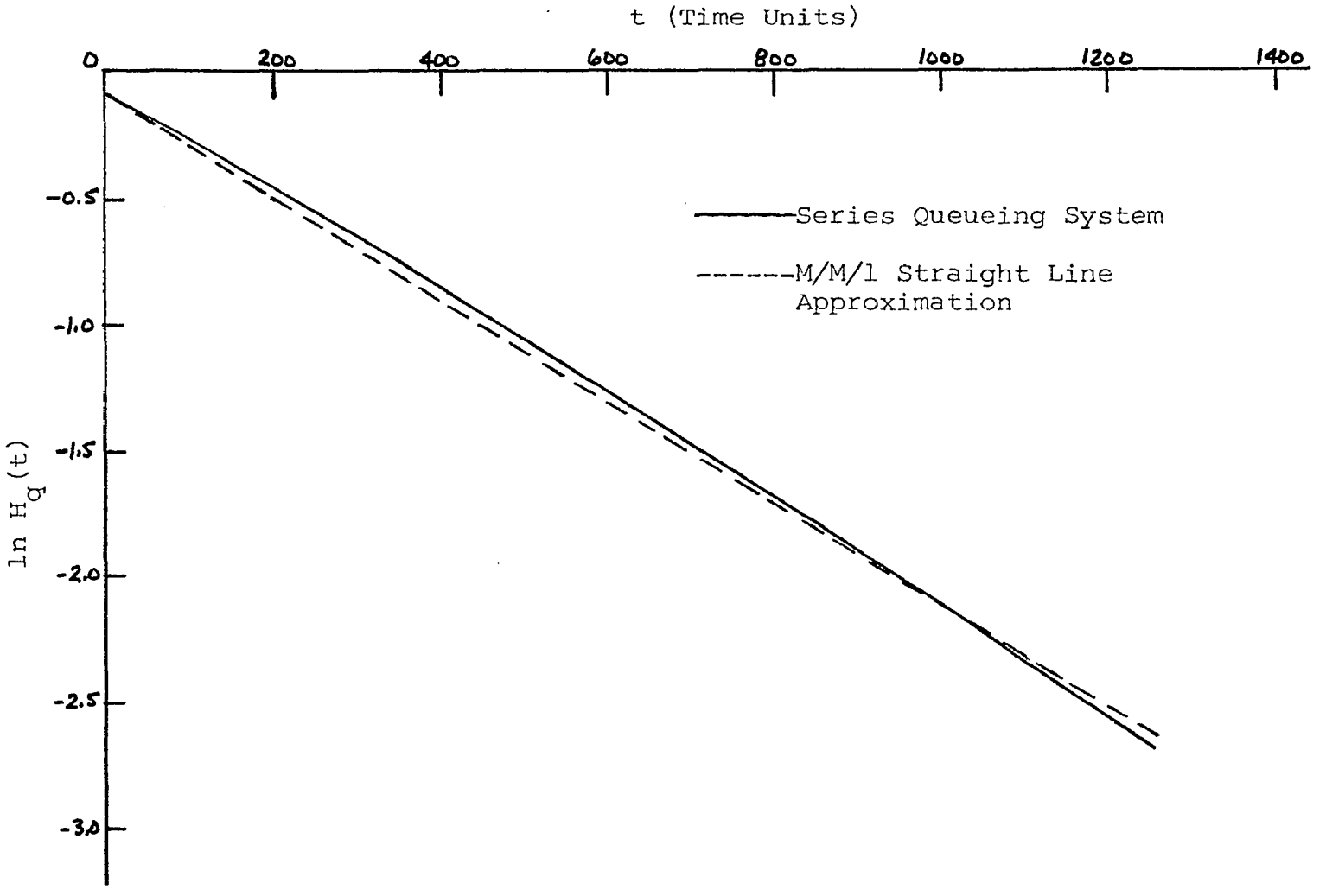


Fig. 26. Determination of M/M/1 Isomorph for Queueing System A9

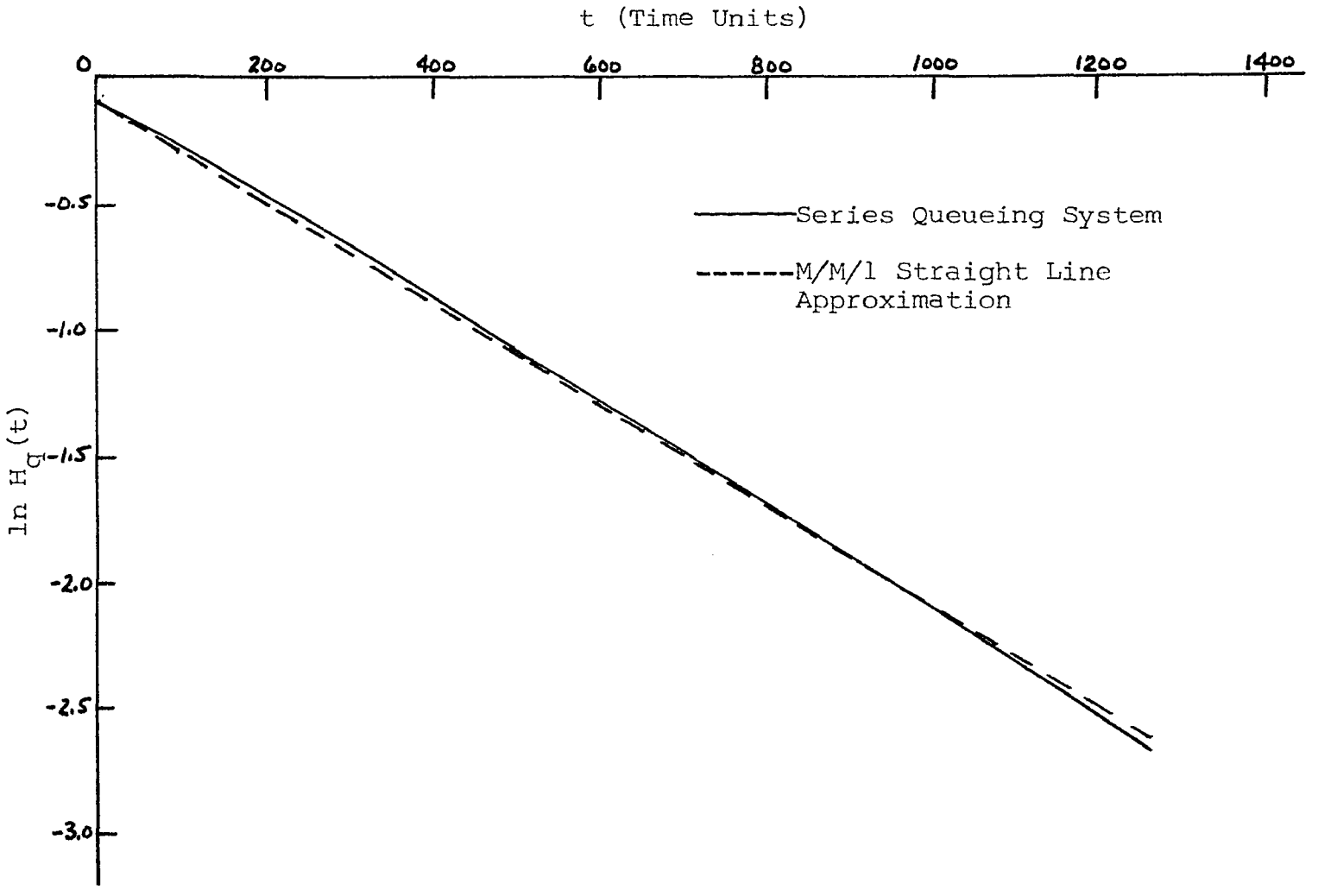


Fig. 27. Determination of M/M/1 Isomorph for Queueing System B9

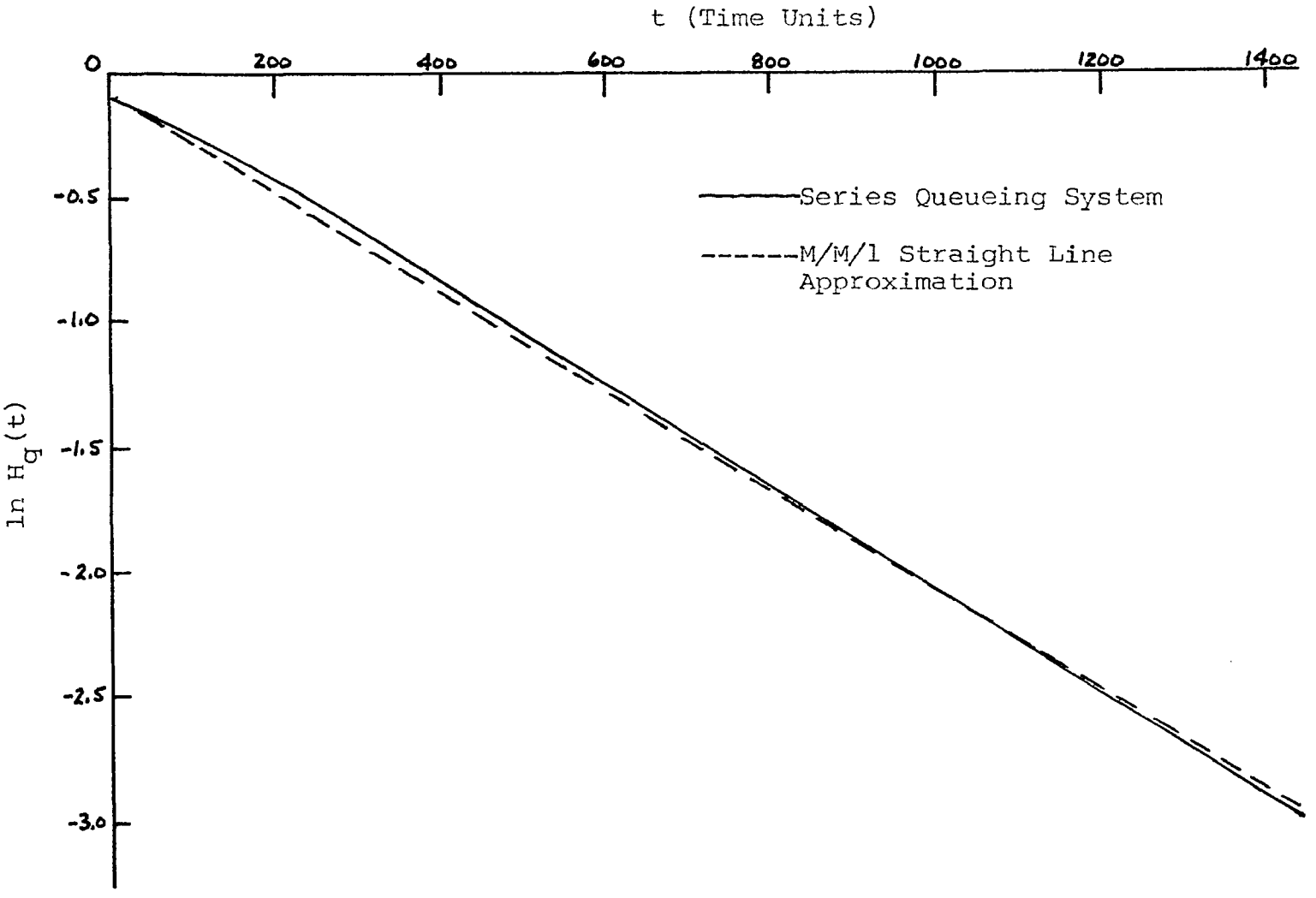


Fig. 28. Determination of M/M/1 Isomorph for Queueing System A10

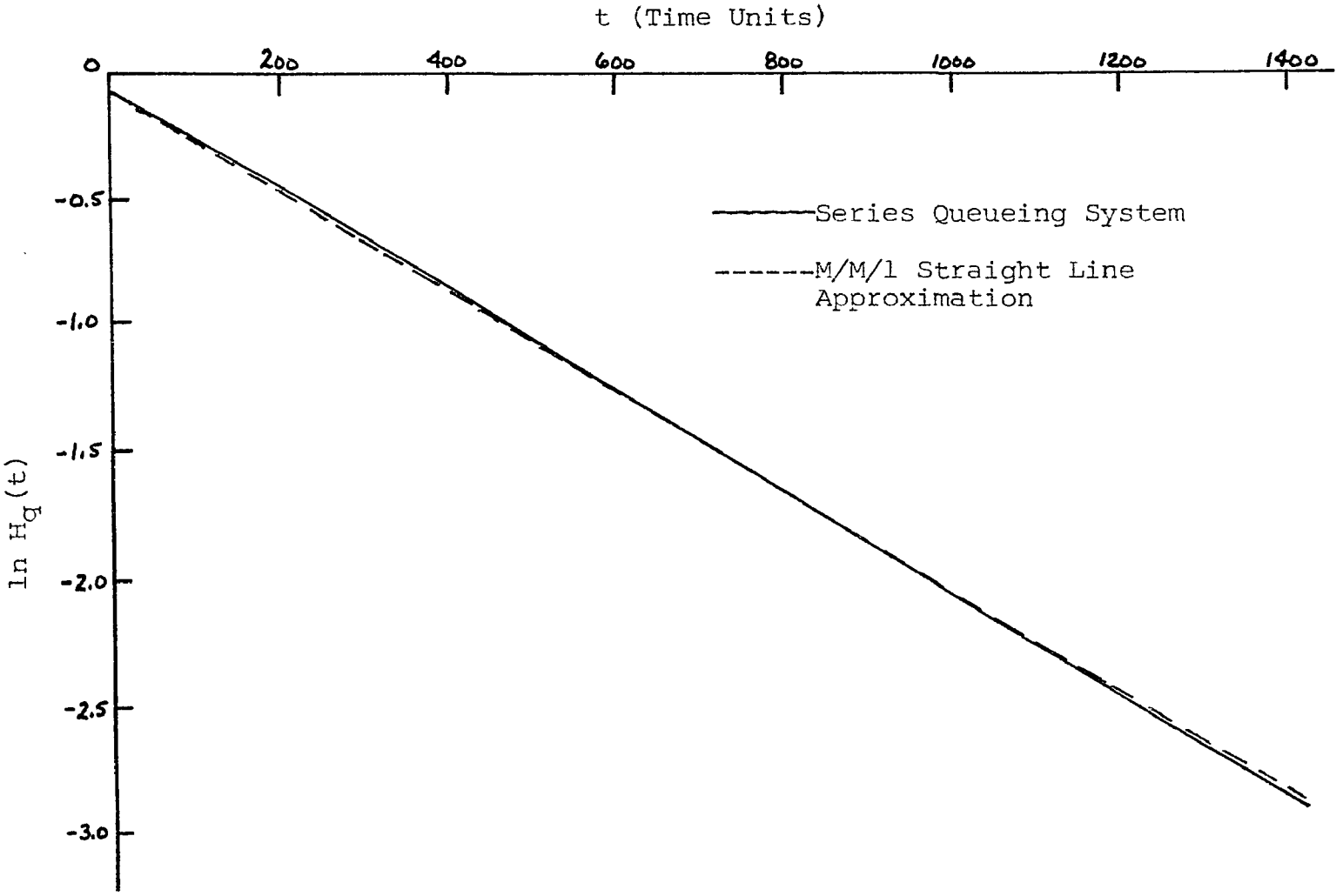


Fig. 29. Determination of M/M/1 Isomorph for Queueing System B10

TABLE 8

RESULTS FOUND BY METHOD OF FITTING MOMENTS

Experiment	Series Queueing System	Isomorphic System
A5	$M/M/1 \rightarrow ./E_2/1$ $1/\lambda = 100$; $1/\mu_1 = 30.0$; $1/\mu_2 = 42.43$ $\bar{w}_q = 36.023$ and $\sigma_{w_q} = 57.460$	$M/E_{6.6}/1$ $1/\lambda = 129.89$; $1/\mu = 64.16$; $\rho = 0.494$
A6	$M/M/1 \rightarrow ./E_3/1$ $1/\lambda = 100$; $1/\mu_1 = 30.0$; $1/\mu_2 = 51.96$ $\bar{w}_q = 48.840$ and $\sigma_{w_q} = 70.656$	$M/E_{36.7}/1$ $1/\lambda = 136.58$; $1/\mu = 75.94$; $\rho = 0.556$
A8	$M/E_2/1 \rightarrow ./E_3/1$ $1/\lambda = 100$; $1/\mu_1 = 42.43$; $1/\mu_2 = 51.96$ $\bar{w}_q = 53.427$ and $\sigma_{w_q} = 73.764$	$M/E_\infty/1$ $1/\lambda = 121.98$; $1/\mu = 72.62$; $\rho = 0.590$
B5	$M/E_2/1 \rightarrow ./M/1$ $1/\lambda = 100$; $1/\mu_1 = 42.43$; $1/\mu_2 = 30.0$ $\bar{w}_q = 33.333$ and $\sigma_{w_q} = 53.607$	$M/E_{5.8}/1$ $1/\lambda = 120.00$; $1/\mu = 58.92$; $\rho = 0.491$
B6	$M/E_3/1 \rightarrow ./M/1$ $1/\lambda = 100$; $1/\mu_1 = 51.96$; $1/\mu_2 = 30.0$ $\bar{w}_q = 44.044$ and $\sigma_{w_q} = 64.894$	$M/E_{9.5}/1$ $1/\lambda = 115.12$; $1/\mu = 63.89$; $\rho = 0.555$

TABLE 8

RESULTS FOUND BY METHOD OF FITTING MOMENTS

Experiment	Series Queueing System	Isomorphic System
A5	$M/M/1 \rightarrow ./E_2/1$ $1/\lambda = 100$; $1/\mu_1 = 30.0$; $1/\mu_2 = 42.43$ $\bar{w}_q = 36.023$ and $\sigma_{w_q} = 57.460$	$M/E_{6.6}/1$ $1/\lambda = 129.89$; $1/\mu = 64.16$; $\rho = 0.494$
A6	$M/M/1 \rightarrow ./E_3/1$ $1/\lambda = 100$; $1/\mu_1 = 30.0$; $1/\mu_2 = 51.96$ $\bar{w}_q = 48.840$ and $\sigma_{w_q} = 70.656$	$M/E_{36.7}/1$ $1/\lambda = 136.58$; $1/\mu = 75.94$; $\rho = 0.556$
A8	$M/E_2/1 \rightarrow ./E_3/1$ $1/\lambda = 100$; $1/\mu_1 = 42.43$; $1/\mu_2 = 51.96$ $\bar{w}_q = 53.427$ and $\sigma_{w_q} = 73.764$	$M/E_\infty/1$ $1/\lambda = 121.98$; $1/\mu = 72.62$; $\rho = 0.590$
B5	$M/E_2/1 \rightarrow ./M/1$ $1/\lambda = 100$; $1/\mu_1 = 42.43$; $1/\mu_2 = 30.0$ $\bar{w}_q = 33.333$ and $\sigma_{w_q} = 53.607$	$M/E_{5.8}/1$ $1/\lambda = 120.00$; $1/\mu = 58.92$; $\rho = 0.491$
B6	$M/E_3/1 \rightarrow ./M/1$ $1/\lambda = 100$; $1/\mu_1 = 51.96$; $1/\mu_2 = 30.0$ $\bar{w}_q = 44.044$ and $\sigma_{w_q} = 64.894$	$M/E_{9.5}/1$ $1/\lambda = 115.12$; $1/\mu = 63.89$; $\rho = 0.555$

TABLE 8--Continued

Experiment	Series Queueing System	Isomorphic System
B8	$M/E_3/1 \rightarrow ./E_2/1$ $1/\lambda=100$; $1/\mu_1=51.96$; $1/\mu_2=42.43$ $\bar{w}_q = 51.115$ and $\sigma_{w_q} = 71.804$	$M/E_{29.3}/1$ $1/\lambda=119.19$; $1/\mu=69.84$; $\rho=0.586$

moments. With this method, isomorphs were found for only some of the series queueing systems after applying the findings of Sphicas and Shimshak.[10]

Both methods of fitting queueing systems provide some isomorphs that closely represent the actual system. Table 9 shows a comparison of the total waiting time distribution functions of system B10 and its M/M/1 isomorph found by fitting distribution functions. In Table 10 the same system and its isomorph are compared on a basis of their frequency distributions. In simulating system B10, thirteen replications were required, each generating 18,000 customer arrivals.

A series queueing system and its M/G/1 isomorph found by fitting moments are also compared. For system A8 and its isomorph, $M/E_{\infty}/1$, the total waiting time distribution functions are shown in Table 11 and their frequency distributions in Table 12.

A comparison of the waiting time distribution functions for each series queueing system and its M/M/1 isomorph are summarized in Table 13. Here the largest absolute vertical deviation between the two distribution functions is presented where $D=\max\left|W_{q_S}(t)-W_{q_I}(t)\right|$ and $W_{q_S}(t)$ is the distribution function of the series queueing system and $W_{q_I}(t)$ is the distribution function of the isomorph. In Table 14, the measure D is given for each system and its

---

[10]Sphicas and Shimshak, "Waiting Time Variability."

TABLE 9

COMPARISON OF DISTRIBUTION FUNCTIONS
OF SERIES SYSTEM B10 AND
ITS M/M/1 ISOMORPH

t (Time units)	$W_q(t)$ Series System	$W_q(t)$ Isomorph
0	.095	.095
120	.266	.286
240	.427	.437
360	.552	.556
480	.650	.650
600	.726	.724
720	.783	.783
840	.827	.829
960	.863	.865
1080	.892	.892
1200	.916	.916
1320	.934	.934
1440	.949	.948

TABLE 10

COMPARISON OF FREQUENCY DISTRIBUTIONS
OF SERIES SYSTEM B10 AND
ITS M/M/1 ISOMORPH

Waiting Time Class Limits	Series System Proba-bilities	Isomorph Proba-bilities	Series System Frequencies	Isomorph Frequencies
0	.095	.095	22,230	22,230
$0<t\leq 120$	.171	.191	40,014	44,694
$120<t\leq 240$	.161	.151	37,674	35,334
$240<t\leq 360$	.125	.119	29,250	27,846
$360<t\leq 480$	.098	.094	22,932	21,996
$480<t\leq 600$	.076	.074	17,784	17,316
$600<t\leq 720$	.057	.059	13,338	13,806
$720<t\leq 840$	.044	.046	10,296	10,764
$840<t\leq 960$	.036	.036	8,424	8,424
$960<t\leq 1080$	.029	.027	6,786	6,318
$1080<t\leq 1200$	.024	.024	5,616	5,616
$1200<t\leq 1320$	.018	.018	4,212	4,212
$1320<t\leq 1440$	.015	.014	3,410	3,276
$1440<t$	.051	.052	11,934	12,168

TABLE 11

COMPARISON OF DISTRIBUTION FUNCTIONS
OF SERIES SYSTEM A8 AND
ITS M/G/1 ISOMORPH

t (Time units)	$W_q(t)$ Series System	$W_q(t)$ Isomorph
0	.410	.405
15	.475	.458
30	.542	.517
45	.602	.585
60	.666	.662
75	.720	.740
90	.767	.780
105	.807	.817
120	.841	.850
135	.870	.879
150	.893	.900
165	.913	.918
180	.930	.933
195	.942	.945
210	.954	.955

TABLE 12

COMPARISON OF FREQUENCY DISTRIBUTIONS
OF SERIES SYSTEM A8 AND
ITS M/G/1 ISOMORPH

Waiting Time Class Limits	Series System Proba-bilities	Isomorph Proba-bilities	Series System Frequencies	Isomorph Frequencies
0	.410	.405	59,040	58,320
$0 < t \leq 15$	.065	.053	9,360	7,632
$15 < t \leq 30$	.067	.059	9,648	8,496
$30 < t \leq 45$	.060	.068	8,640	9,792
$45 < t \leq 60$	.064	.077	9,216	11,088
$60 < t \leq 75$	.064	.078	7,776	11,232
$75 < t \leq 90$	.047	.040	6,768	5,760
$90 < t \leq 105$	.040	.037	5,760	5,328
$105 < t \leq 120$	.034	.033	4,896	4,752
$120 < t \leq 135$	.029	.029	4,176	4,176
$135 < t \leq 150$	.023	.021	3,312	3,024
$150 < t \leq 165$	.020	.018	2,880	2,592
$165 < t \leq 180$	.017	.015	2,448	2,160
$180 < t \leq 195$	.012	.012	1,728	1,728
$195 < t \leq 210$	.012	.010	1,728	1,440
$210 < t$	.046	.045	6,624	6,480

TABLE 13

LARGEST ABSOLUTE VERTICAL DEVIATION BETWEEN DISTRIBUTION FUNCTIONS OF THE SERIES SYSTEMS AND THEIR M/M/1 ISOMORPHS

System	$D = \left\lvert W_{q_S}(t) - W_{q_I}(t) \right\rvert$
A1	.080
A2	.085
A3	.070
A4	.071
A5	.024
A6	.036
A7	.040
A8	.042
A9	.035
A10	.032
B1	.063
B2	.075
B3	.069
B4	.063
B5	.026
B6	.032
B7	.024
B8	.041
B9	.022
B10	.020

TABLE 14

LARGEST ABSOLUTE VERTICAL DEVIATION BETWEEN DISTRIBUTION FUNCTIONS OF THE SERIES SYSTEMS AND THEIR M/G/1 ISOMORPHS

System	$D = \left\| W_{q_S}(t) - W_{q_I}(t) \right\|$
A5	.031
A6	.021
A8	.025
B5	.034
B6	.024
B8	.021

M/G/1 isomorph found by fitting moments.

A problem arose when statistical tests were used to determine significant differences in the waiting time frequency distributions of the series system and its isomorph. Since the number of customers in each simulation experiment was over 100,000, the differences in the frequencies would have to be exceptionally small to pass any statistical test. As a result, significant differences appeared in each system when comparing waiting time frequency distributions.

The inability to find isomorphic systems that passed the statistical tests for goodness of fit does not demean the techniques developed for determining isomorphs. Rather it is a result of the large sample sizes necessary to accurately determine output in series queueing systems. As mathematical analysis continues into the study of queues in series, approximations for the waiting time distribution function may be developed that will yield more conclusive proof in favor of the existence of isomorphs for these systems.

Observation of the graphs of the $\ln H_q(t)$ for the series queueing systems in Figures 10 through 29 indicate that the M/M/1 isomorphs found have better fits in the areas of larger waiting times. This suggests that perhaps the isomorphs are useful in predicting only portions of the waiting time distribution. Also $\ln H_q(t)$ may be fit by a system of broken lines so that one M/M/1 isomorph is

useful for studying smaller waiting times and another M/M/1 isomorph for the larger waiting times. In addition, many of the graphs of ln $H_q(t)$ that significantly depart from linearity can be fit by curves of higher order than one. Ghosal[11] found that the waiting time distribution function of an $M/E_k/1$ system follows a hyperexponential distribution and has the form expressed by many of these curves. However, the distribution function of this single server queue cannot be expressed in a closed mathematical form but is a function to the power k. Even for small values of k, no algorithms are known for determining the parameters of the $M/E_k/1$ isomorph from the series queueing system.

In applying the method of moments, results have been limited because of the few service distributions considered in the M/G/1 system. Some work has been done by the researcher on approximation formulas for the moments of the waiting time in the $E_j/E_k/1$ system. Again the difficulty lies with determining the waiting time distribution function. The investigation of new single server systems and future analytical developments in their analysis can only lead to further developments in isomorphic studies.

---

[11] A. Ghosal, "Isomorphic Queueing Systems and Related Problems."

useful for studying smaller waiting times and another M/M/1 isomorph for the larger waiting times. In addition, many of the graphs of ln $H_q(t)$ that significantly depart from linearity can be fit by curves of higher order than one. Ghosal[11] found that the waiting time distribution function of an $M/E_k/1$ system follows a hyperexponential distribution and has the form expressed by many of these curves. However, the distribution function of this single server queue cannot be expressed in a closed mathematical form but is a function to the power k. Even for small values of k, no algorithms are known for determining the parameters of the $M/E_k/1$ isomorph from the series queueing system.

In applying the method of moments, results have been limited because of the few service distributions considered in the M/G/1 system. Some work has been done by the researcher on approximation formulas for the moments of the waiting time in the $E_j/E_k/1$ system. Again the difficulty lies with determining the waiting time distribution function. The investigation of new single server systems and future analytical developments in their analysis can only lead to further developments in isomorphic studies.

---

[11] A. Ghosal, "Isomorphic Queueing Systems and Related Problems."

CHAPTER VI

CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER STUDIES

The purpose of this investigation has been the development of techniques which are useful for the analysis of systems of service stations in series. The approach has been directed into two areas. One deals with the determination of an optimal sequence of the servers in the system. The second is concerned with finding a single server queue which is isomorphic with the series system in terms of total waiting time.

Sequencing is a topic of extreme usefulness and importance. There are numerous real-life problems that can be represented by a system of servers in series. For many of them, the ability exists to arrange the servers in a particular order. Because of the high costs involved with waiting in the system, the ability to determine a sequence that can minimize the waiting time is a great asset to the applied researcher and systems designer. In addition, comparison of sequences on the basis of rules of stochastic dominance is a major contribution since many cost functions involved in real problems are not linear. With costs related directly to the waiting time distribution

functions, the introduction of stochastic dominance adds new depth to the evaluation of sequences of servers in series.

The evaluation of a system of service stations in series requires knowledge of the variance of the service distribution and utilization rate for each station. Determination of an optimal sequence in terms of the mean total waiting time is based upon some analytical derivations. Through additional statistical analysis and simulation work, stochastic dominance is included in the analysis of the sequences.

For a given system, knowledge of the service distribution and utilization rates for each service station would allow the study of sequences through computer simulation. However, the procedure developed in this study offers certain advantages over simulation alone. Whereas simulation can only study a particular system, the analytical technique can analyze sequences for any utilization rates in a two-station series system. Not only is this method more comprehensive, but it is more efficient with regard to computing time.

This investigation serves to extend the current knowledge in the sequencing of servers. Previous work had analyzed only stations with constant and exponential service distributions. Now considering the family of Erlang service distributions, a system of two stations in series can be evaluated to give a range of parameters

where each of the two possible sequences of these servers is optimal on the basis of mean waiting times. Further, these regions are broken down in terms of first and second degree stochastic dominance. When one of the service stations is exponential, the sequence with the exponential server placed second in the order is always optimal, again by first and second degree dominance.

There exists the need for further study in several areas. The consideration of additional networks of queues with arrival patterns other than Poisson and service times other than Erlang can be researched in a manner similar to this study. Allowing queues in parallel as well as in series would prove to be a challenging problem. Another obvious extension is the consideration of finite queues as suggested in Chapter II. Each of these would serve to broaden the investigation into new and practical areas. Ultimately, the determination of heuristic rules for ordering with application to any system of service stations in series will become a reality.

The benefits derived from the study of isomorphism have been mentioned throughout the paper. The ability to find single server queues that display characteristics which are equivalent to more complex systems of servers in series, makes it possible to study the series system by investigating its isomorph. Research that might otherwise be impossible because of the complexity of the series system can now be carried out. The effects of changes in

the order of stations, arrival or service rates, or other elements in the system can be analyzed by working with the single server isomorph.

Two methods of determining isomorphs have been developed. In each, total waiting time is the basis of determining the isomorph. The first method estimates the parameters of the isomorph through fitting the waiting time distribution function of the series system. The second attempts to fit moments of the waiting time density function of the series system.

The results obtained from the application of these techniques on several series queueing systems have shown limited success. Apparantly not all systems have isomorphs, and those that do suggest their usefulness may be in predicting only portions of the waiting time distribution functions. It seems that the isomorphs have good fits in the area of larger waiting times. This would still prove to be useful to the system analyst who is concerned with non-linear cost functions that are related to waiting times.

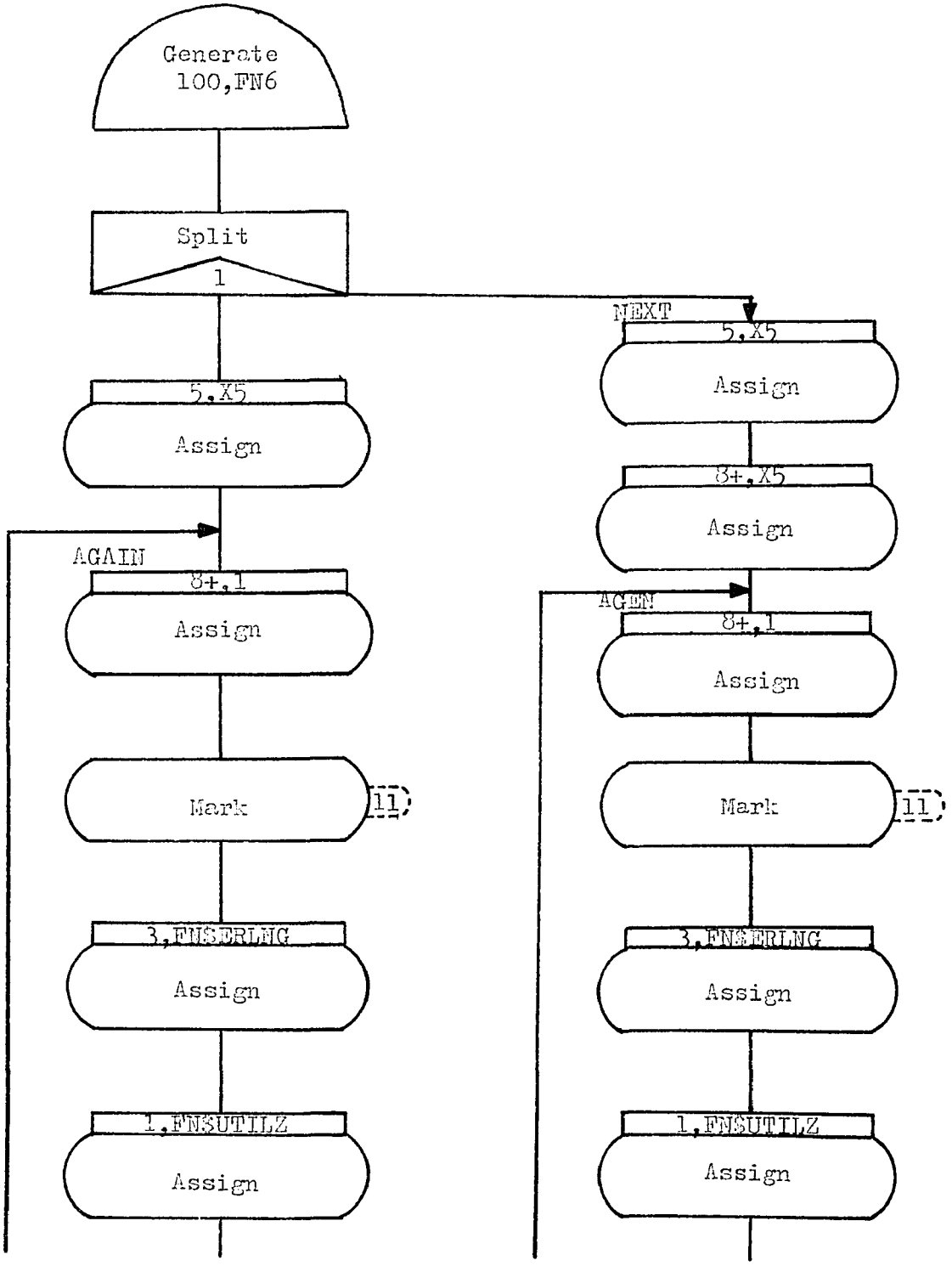
The development of this approach is novel. No previous work had attempted to find isomorphs of series queueing systems with respect to total waiting time. There is a vast amount of research that remains to be done, much of which might answer some of the open questions. The investigation of $M/E_k/1$ and $E_j/E_k/1$ systems as isomorphs must be conducted. This demands further study into the mathematical and graphical forms of the waiting time

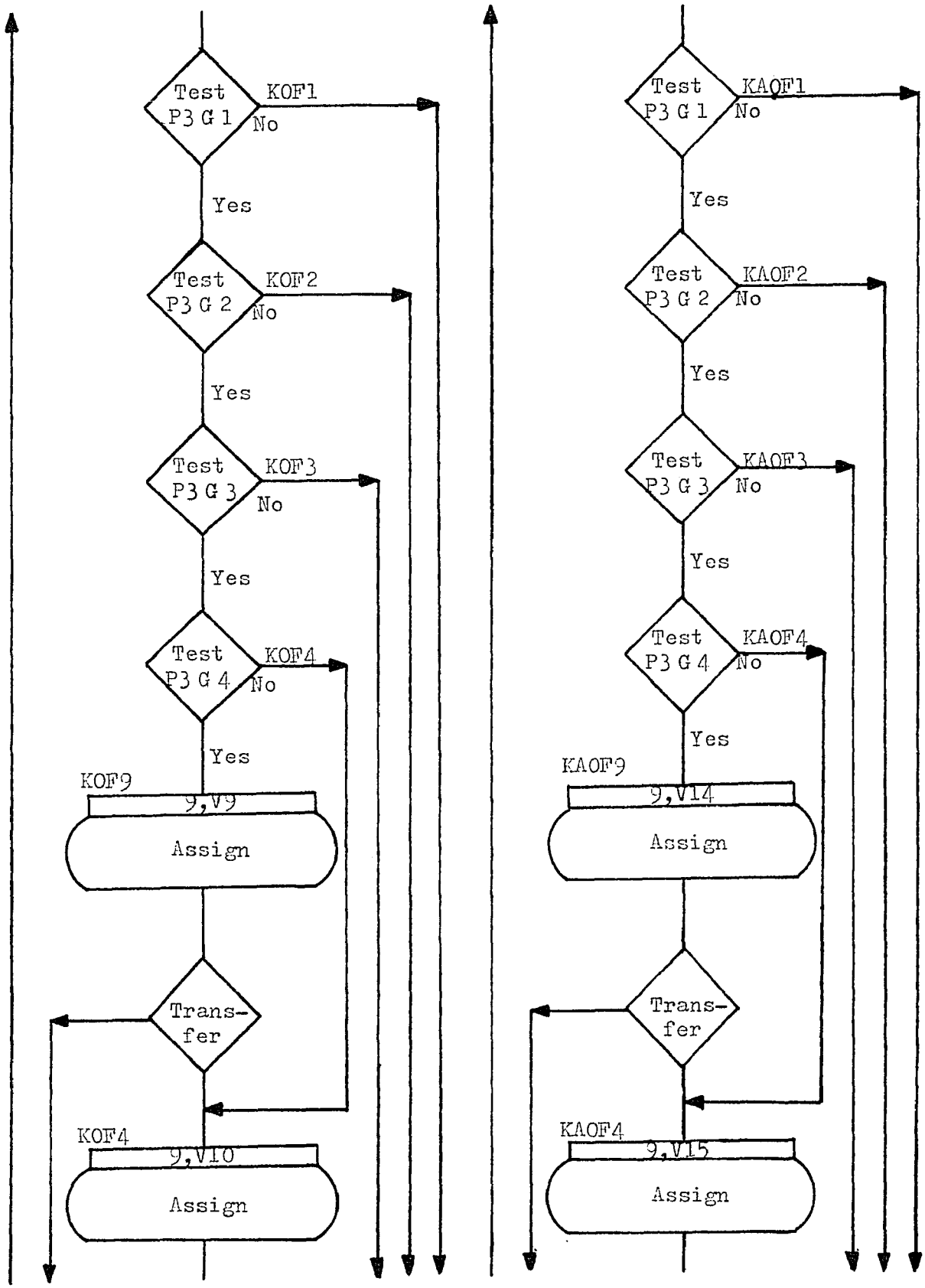
distributions of these systems. Implied in this research is the use of two single server isomorphs in predicting the waiting time distribution in the series queueing system. One may be applicable in studying smaller waiting times, the other for larger waiting times. Analysis of systems of GI/G/1 queues in series and determination of isomorphs with respect to output characteristics other than total waiting time are important areas to be examined. The potential for further study and extensions seems unlimited.

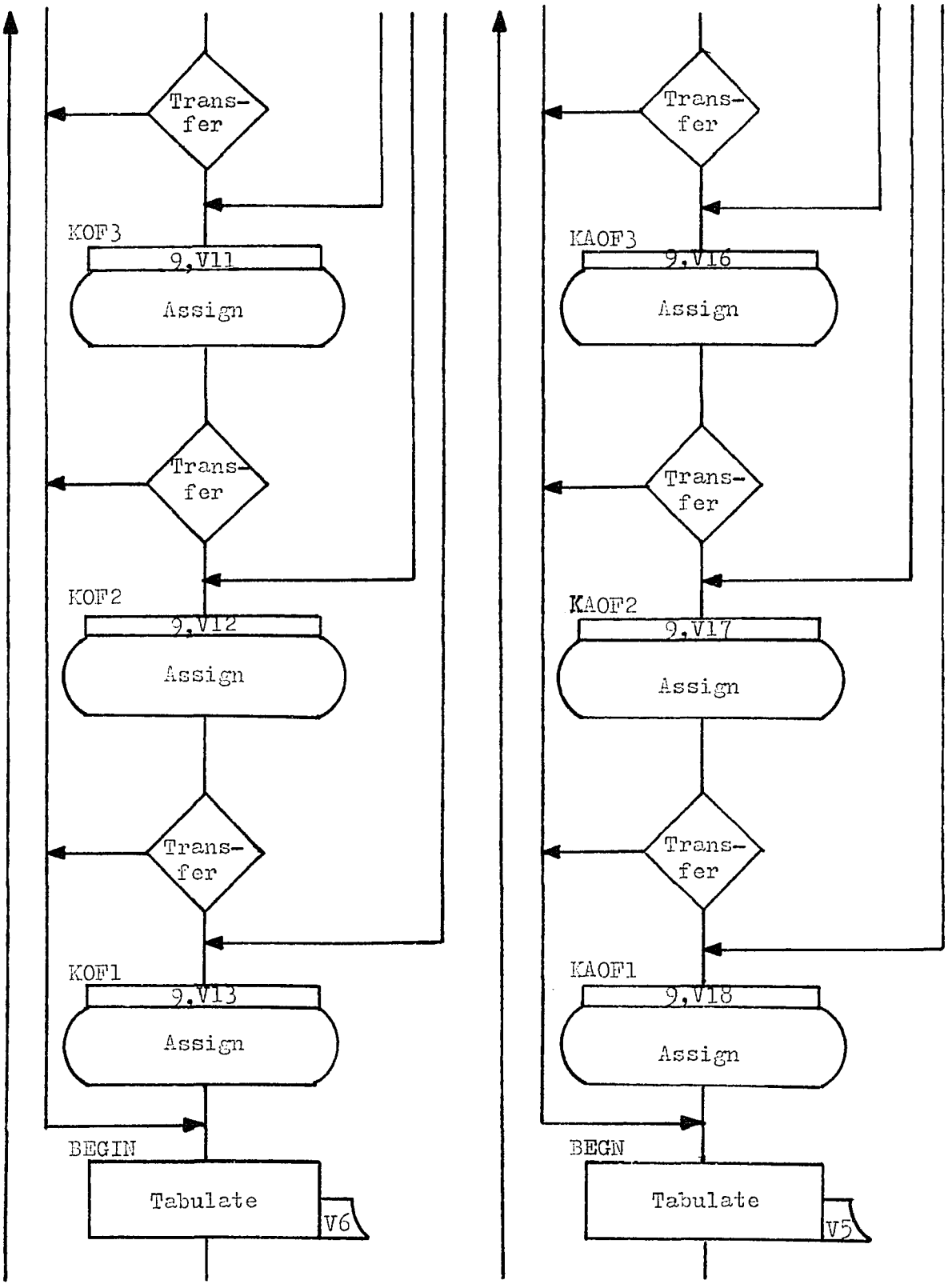
Although the model used in this study has been limited to two stations in series, the extension to n single channel service stations in series, though difficult to work with, is not an impossible task. Study of the two server model expresses all the techniques involved in the analysis of sequencing and isomorphism, and is the most useful system to investigate. This represents many real-life situations and, very often, the more complex problems can be reduced in stages to evaluating two servers in series. Also, further studies in isomorphism will allow the n server system to be reduced through the determination of single server isomorphic queues. Then the methods for evaluating the sequence of stations as developed in this paper can be applied. Together, the study of sequencing and isomorphism will enable analysis of series queueing systems to become a simple and routine task.

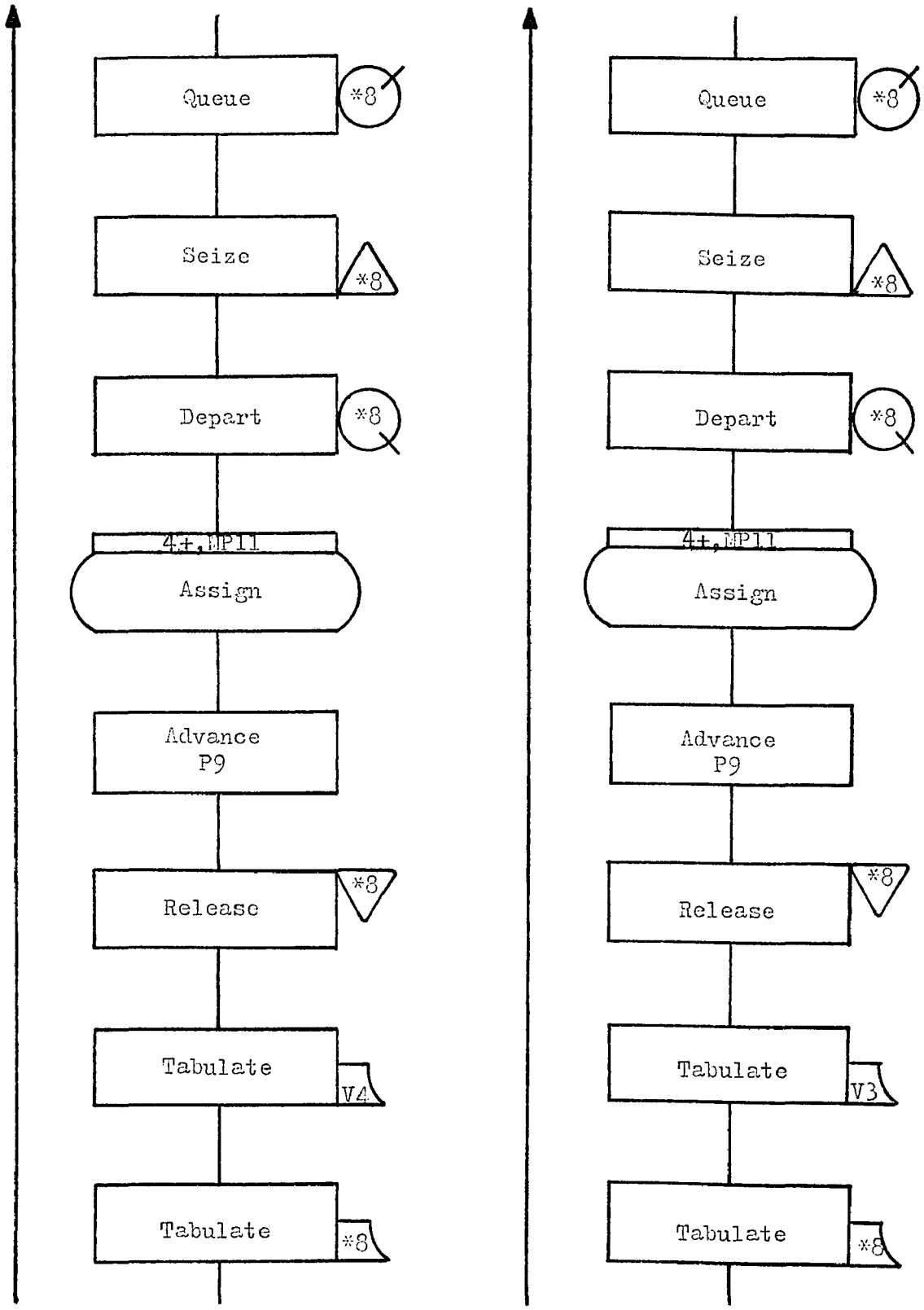
APPENDIX A

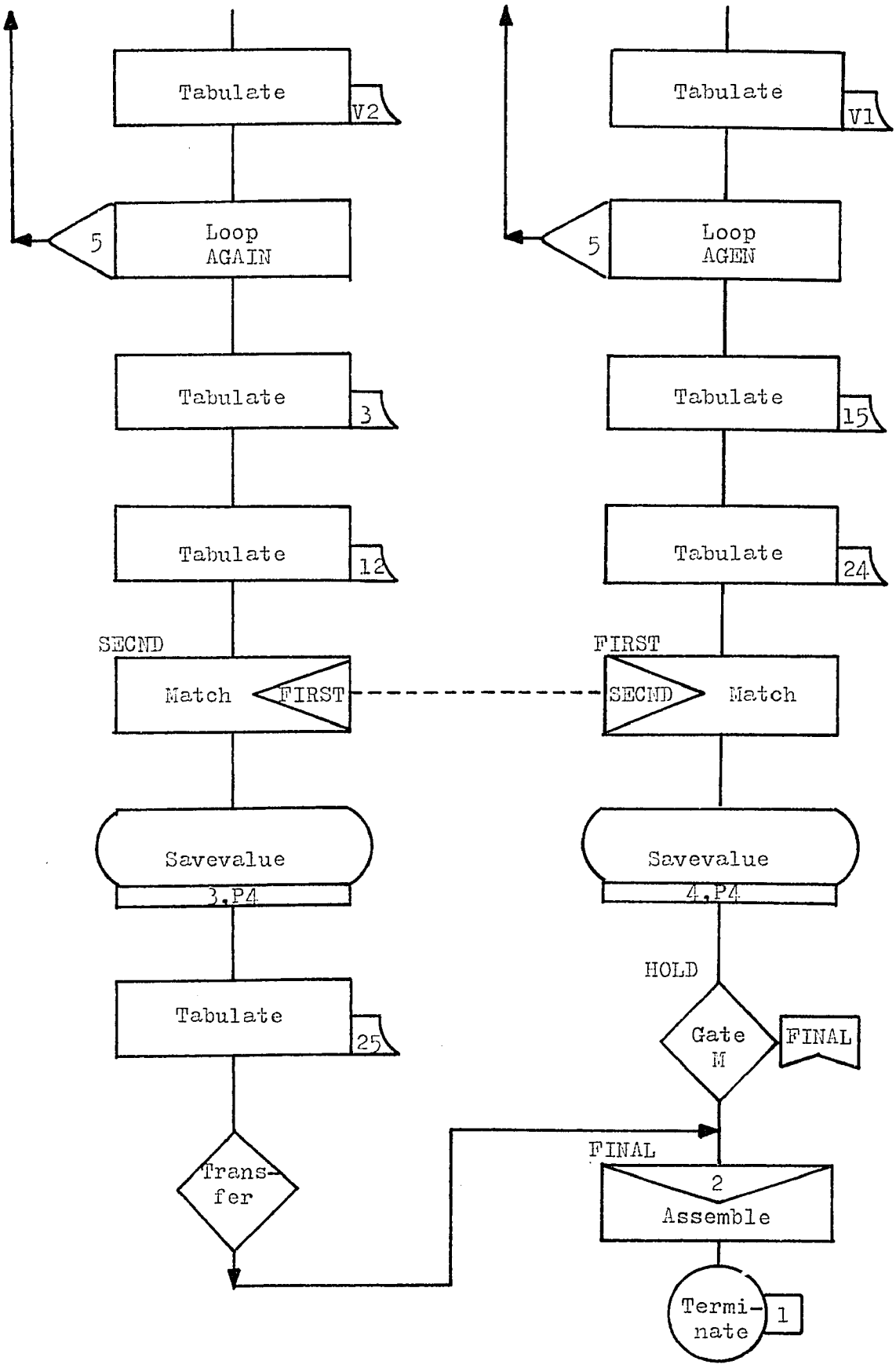
GPSS BLOCK FORMAT OF COMPUTER PROGRAM


Generate
100,FN6
Split
1
NEXT
5,X5
Assign
8+,X5
Assign
AGEN
8+,1
Assign
Mark
11
3,FN$ERLNG
Assign
1,FN$UTILZ
Assign
5,X5
Assign
AGAIN
8+,1
Assign
Mark
11
3,FN$ERLNG
Assign
1,FN$UTILZ
Assign


Test P3 G 1
KOF1
No
Yes
Test P3 G 2
KOF2
No
Yes
Test P3 G 3
KOF3
No
Yes
Test P3 G 4
KOF4
No
Yes
KOF9
9,V9
Assign
Trans-fer
KOF4
9,V10
Assign
Test P3 G 1
KAOF1
No
Yes
Test P3 G 2
KAOF2
No
Yes
Test P3 G 3
KAOF3
No
Yes
Test P3 G 4
KAOF4
No
Yes
KAOF9
9,V14
Assign
Trans-fer
KAOF4
9,V15
Assign


Trans-
fer
KOF3
9,V11
Assign
Trans-
fer
KOF2
9,V12
Assign
Trans-
fer
KOF1
9,V13
Assign
BEGIN
Tabulate
V6
Trans-
fer
KAOF3
9,V16
Assign
Trans-
fer
KAOF2
9,V17
Assign
Trans-
fer
KAOF1
9,V18
Assign
BEGN
Tabulate
V5


Queue
*8
Seize
*8
Depart
*8
4+,MP11
Assign
Advance
P9
Release
*8
Tabulate
V4
Tabulate
*8
Queue
*8
Seize
*8
Depart
*8
4+,MP11
Assign
Advance
P9
Release
*8
Tabulate
V3
Tabulate
*8


Tabulate
V2
5
Loop
AGAIN
Tabulate
3
Tabulate
12
SECND
Match
FIRST
Savevalue
3,P4
Tabulate
25
Trans-
fer
Tabulate
V1
5
Loop
AGEN
Tabulate
15
Tabulate
24
FIRST
SECND
Match
Savevalue
4,P4
HOLD
Gate
M
FINAL
FINAL
2
Assemble
Termi-
nate
1

APPENDIX B

COMPARISON OF CUMULATIVE DISTRIBUTIONS UNDER SEQUENCES A AND B FOR SERIES QUEUEING SYSTEMS 1 THROUGH 10

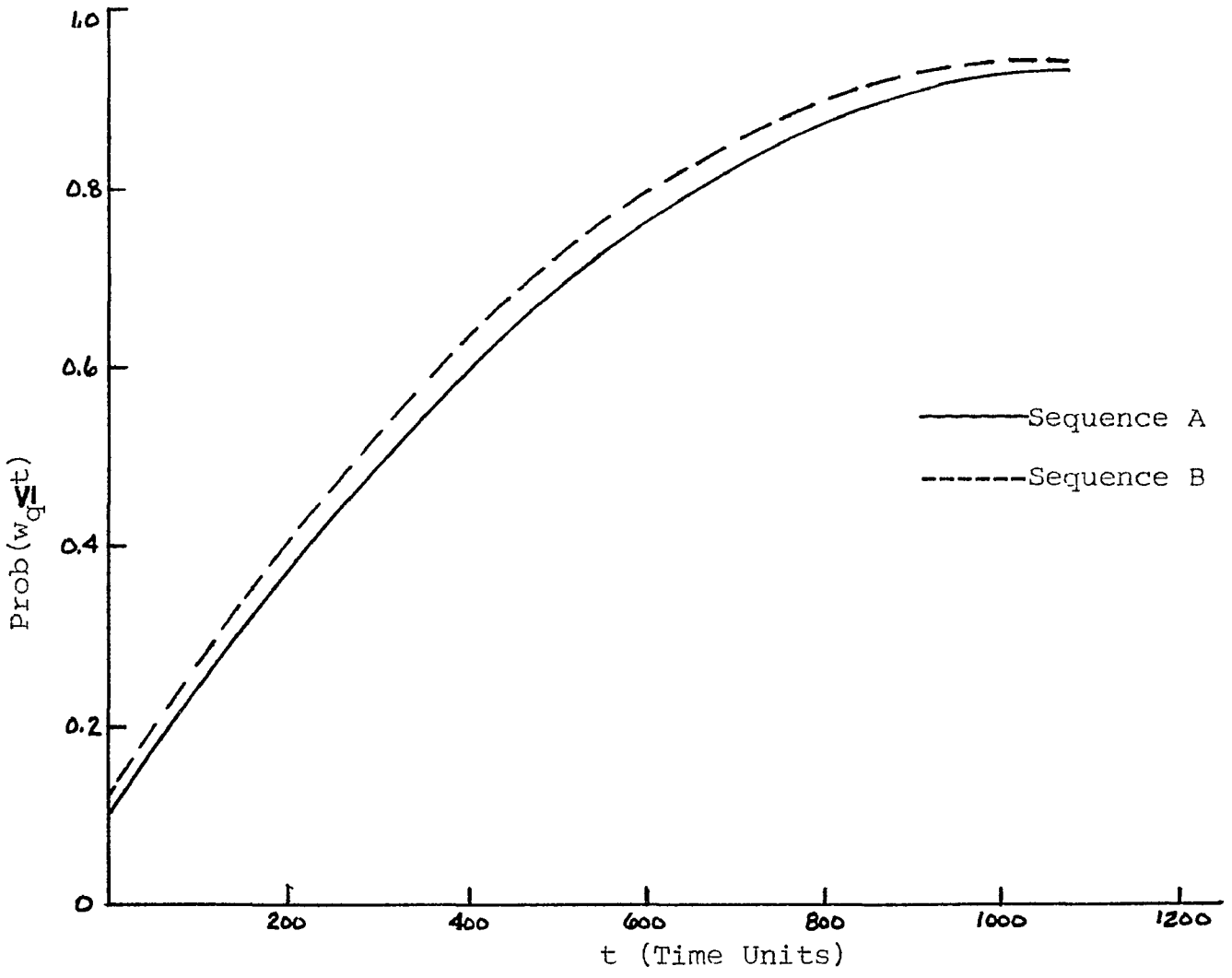


Fig. 30. A Comparison of Cumulative Distributions under Sequences A and B for Series System 1

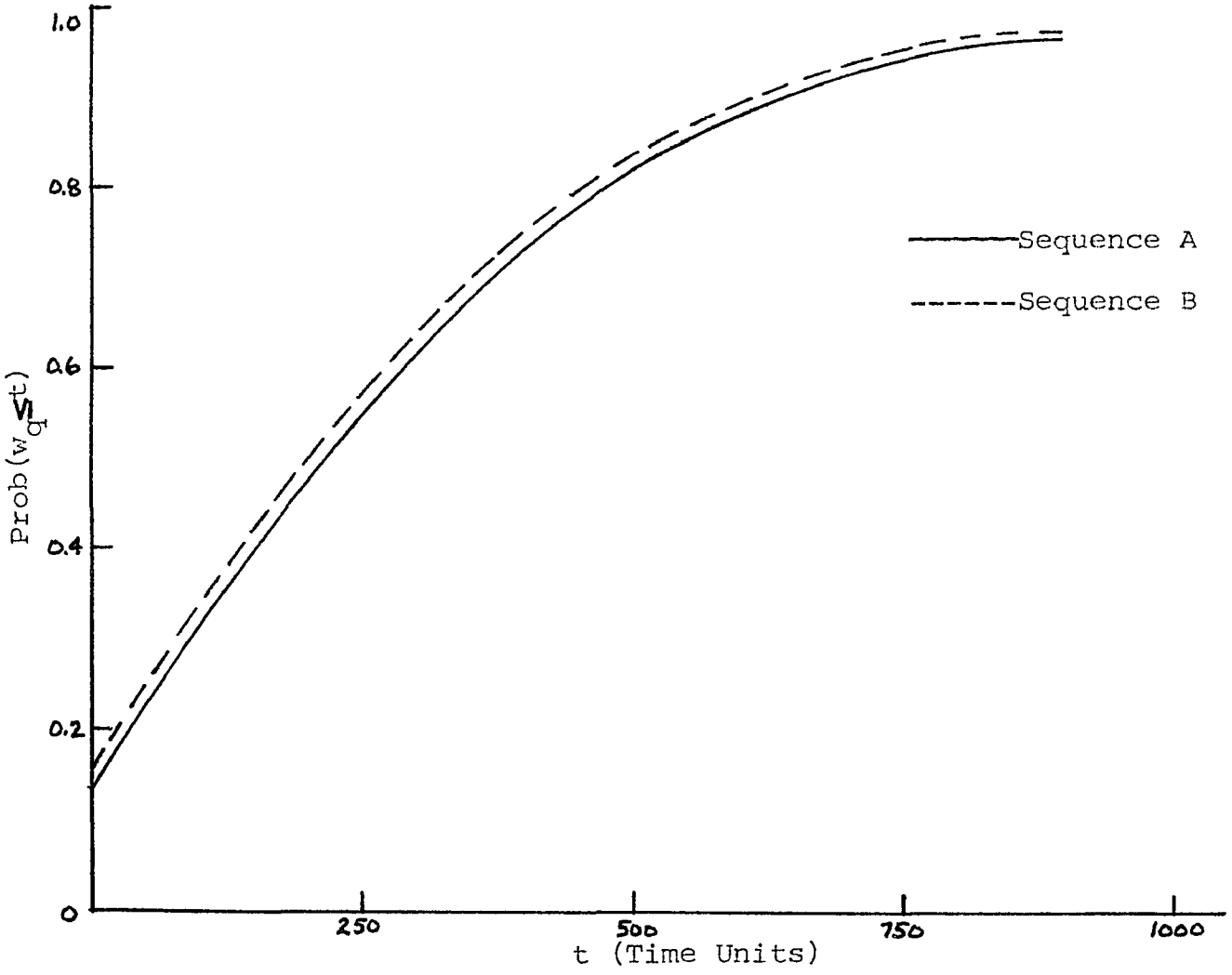


Fig. 31. A Comparison of Cumulative Distributions under Sequences A and B for Series System 2

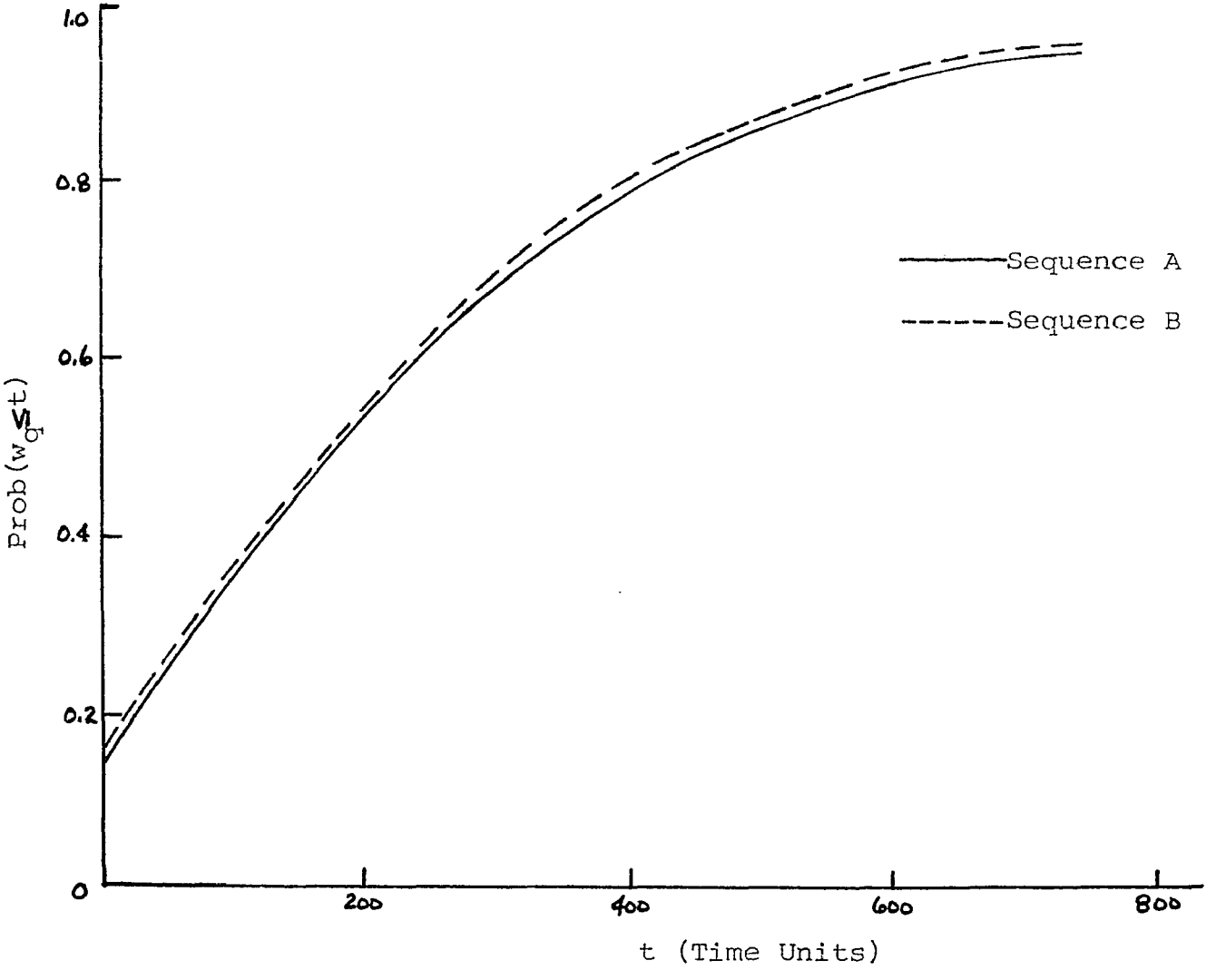


Fig. 32. A Comparison of Cumulative Distributions under Sequences A and B for Series System 3

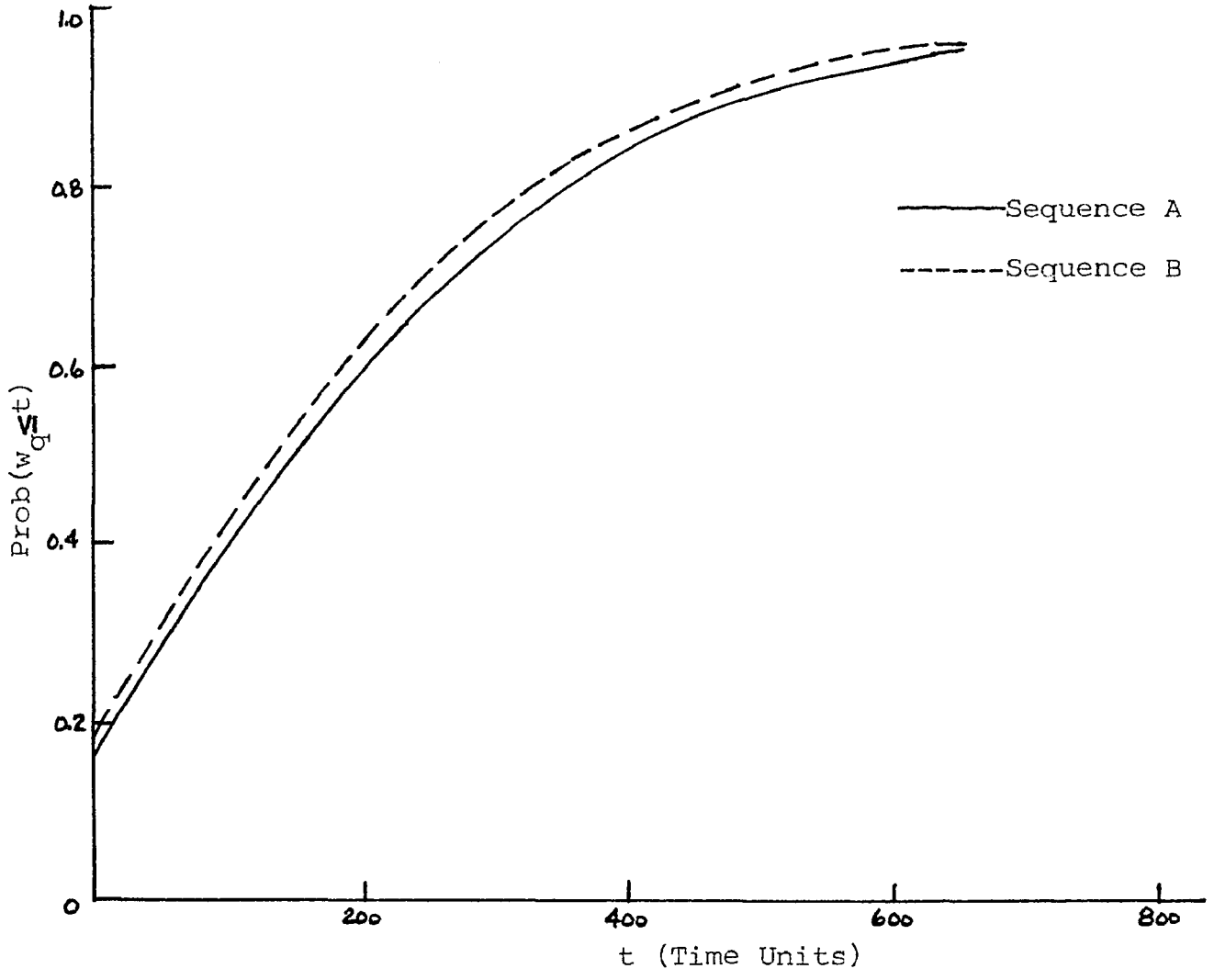


Fig. 33. A Comparison of Cumulative Distributions under Sequences A and B for Series System 4

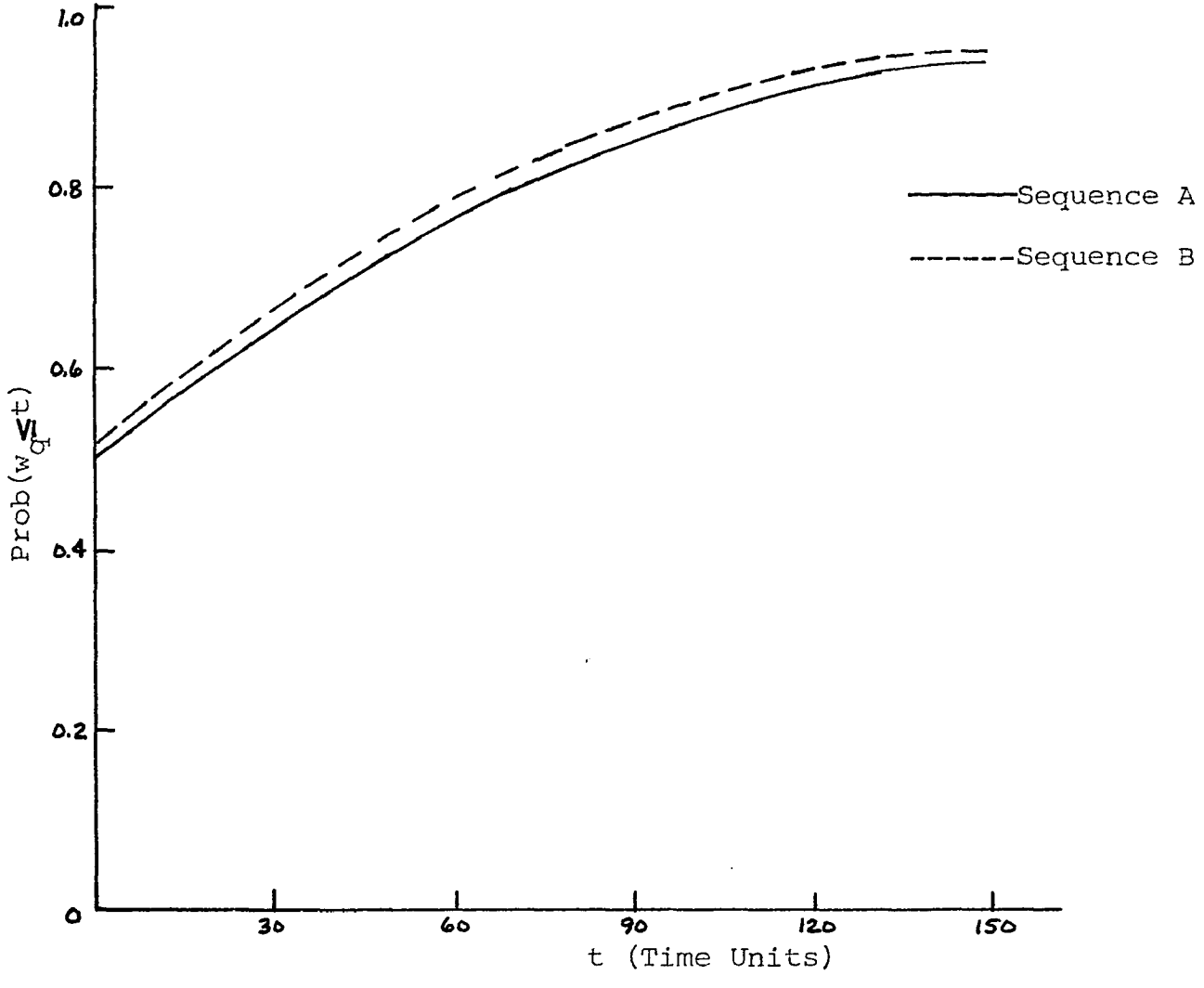


Fig. 34. A Comparison of Cumulative Distributions under Sequences A and B for Series System 5

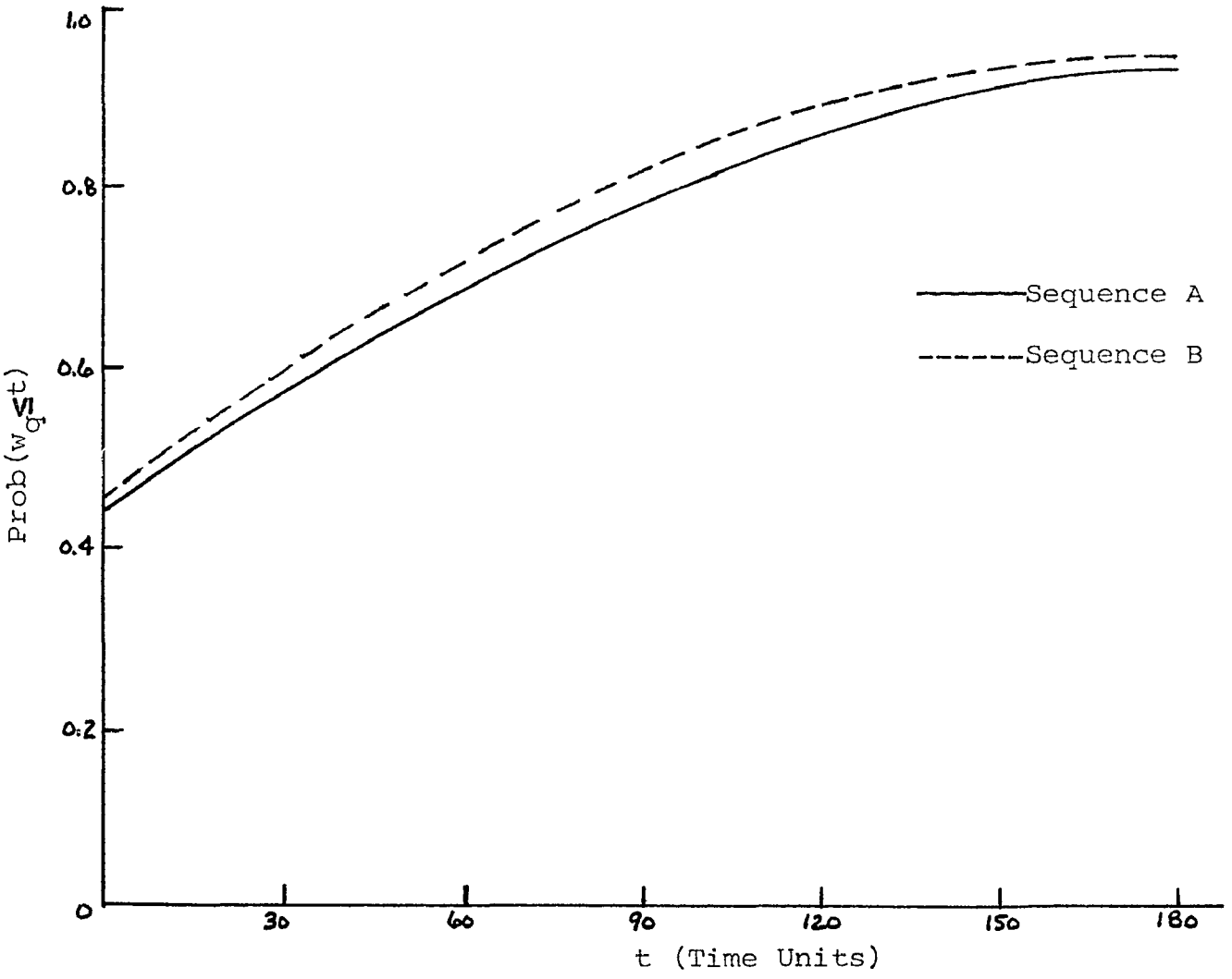


Fig. 35. A Comparison of Cumulative Distributions under Sequences A and B for Series System 6

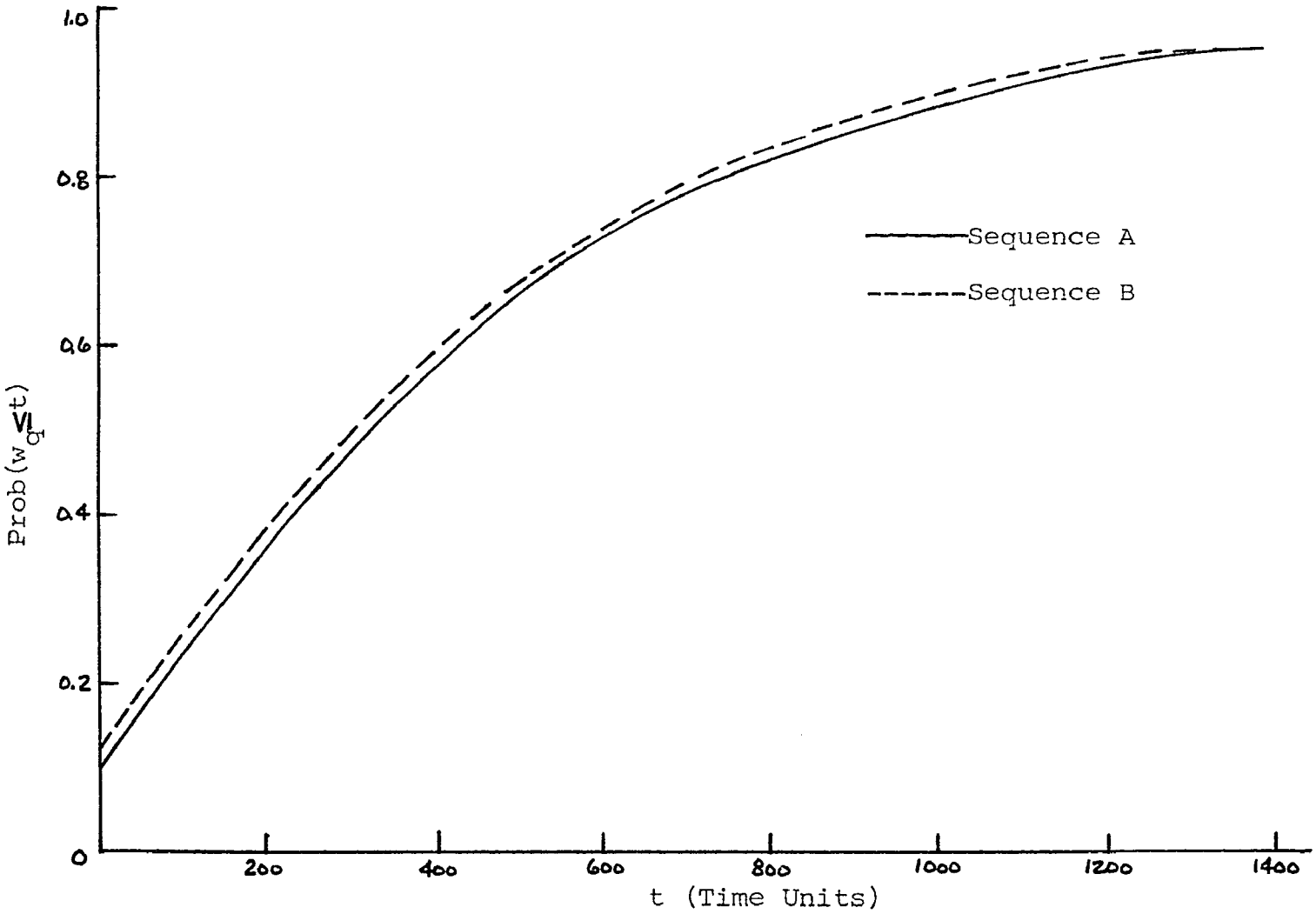


Fig. 36. A Comparison of Cumulative Distributions under Sequences A and B for Series System 7

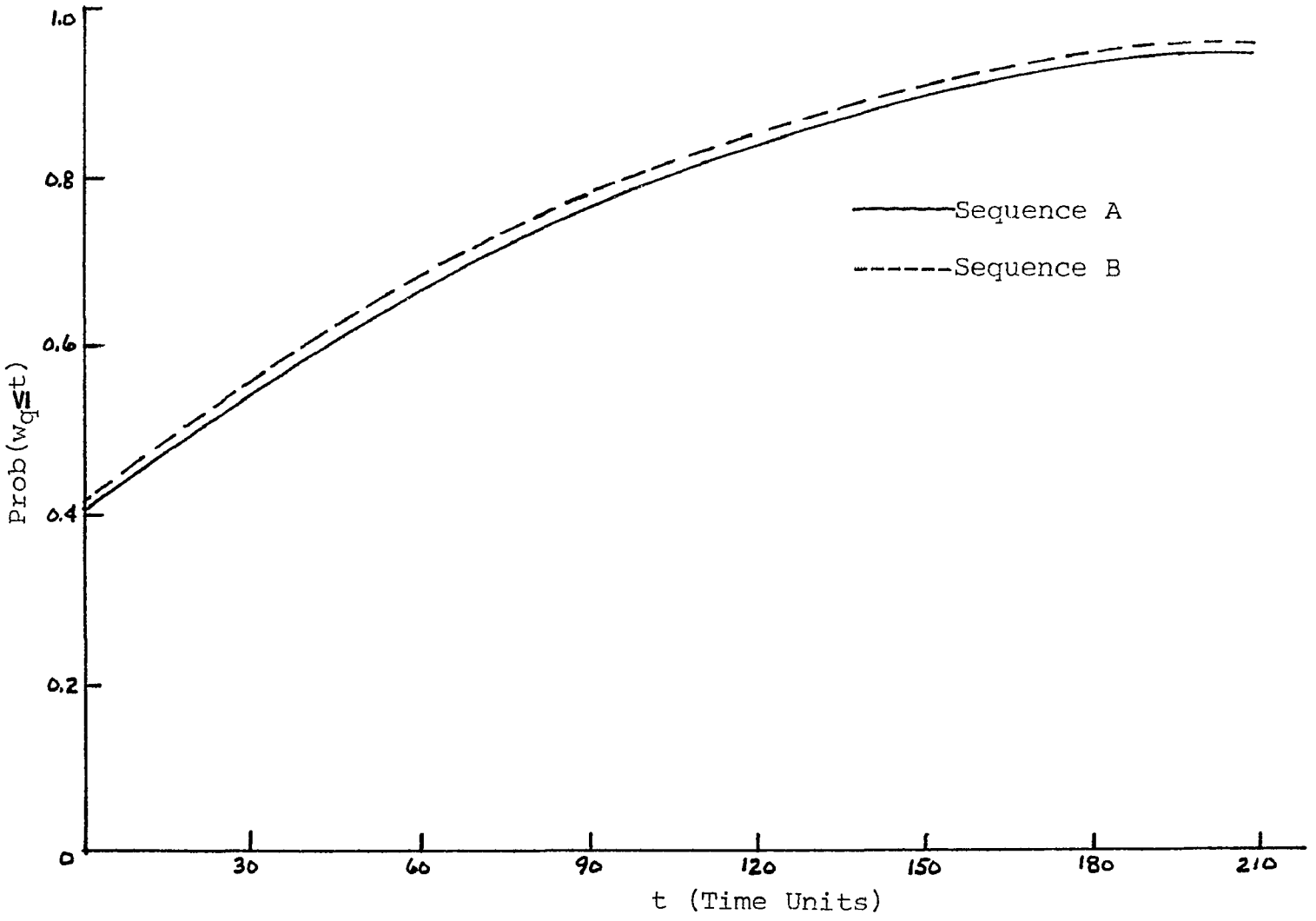


Fig. 37. A Comparison of Cumulative Distributions under Sequences A and B for Series System 8

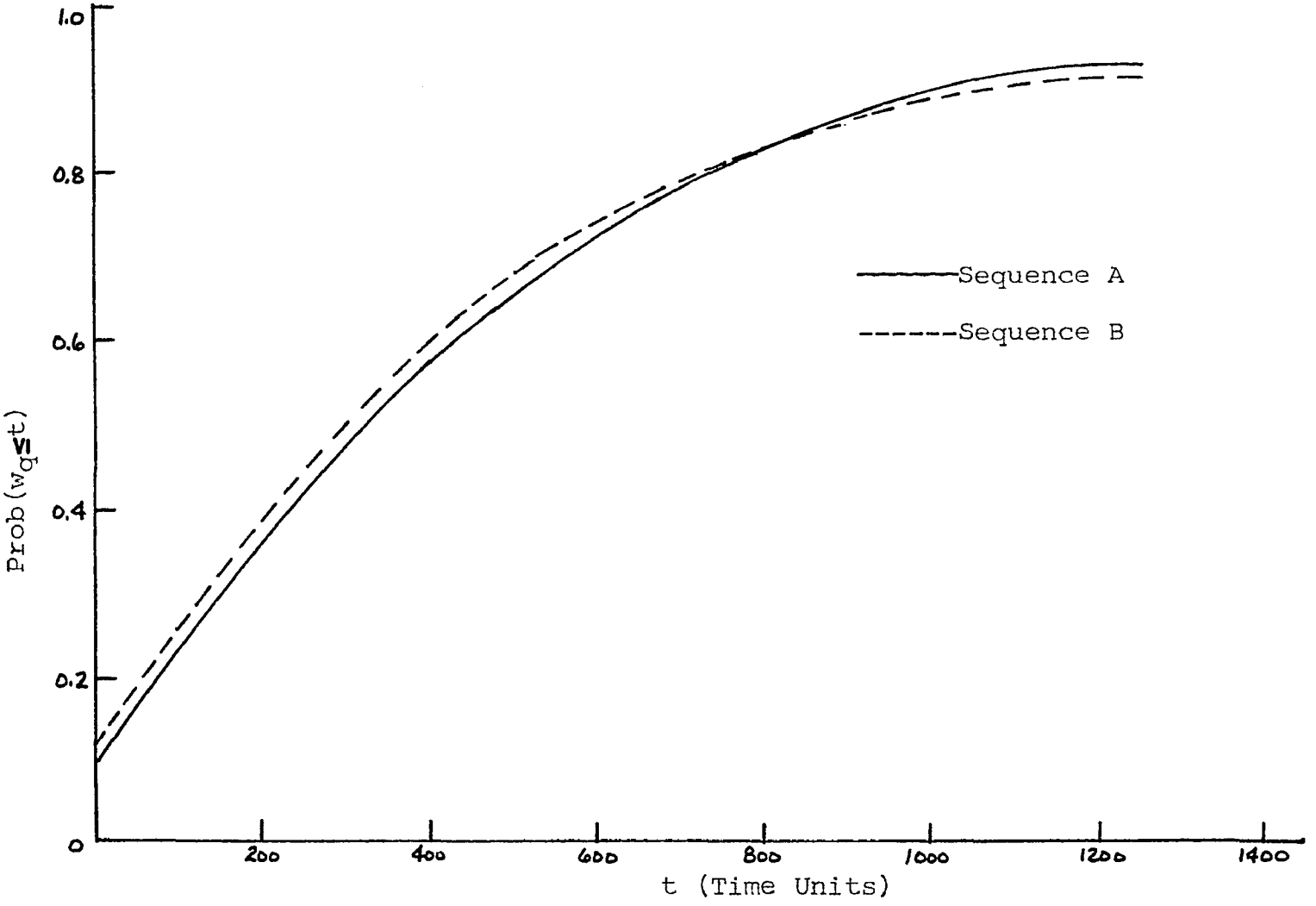


Fig. 38. A Comparison of Cumulative Distributions under Sequences A and B for Series System 9

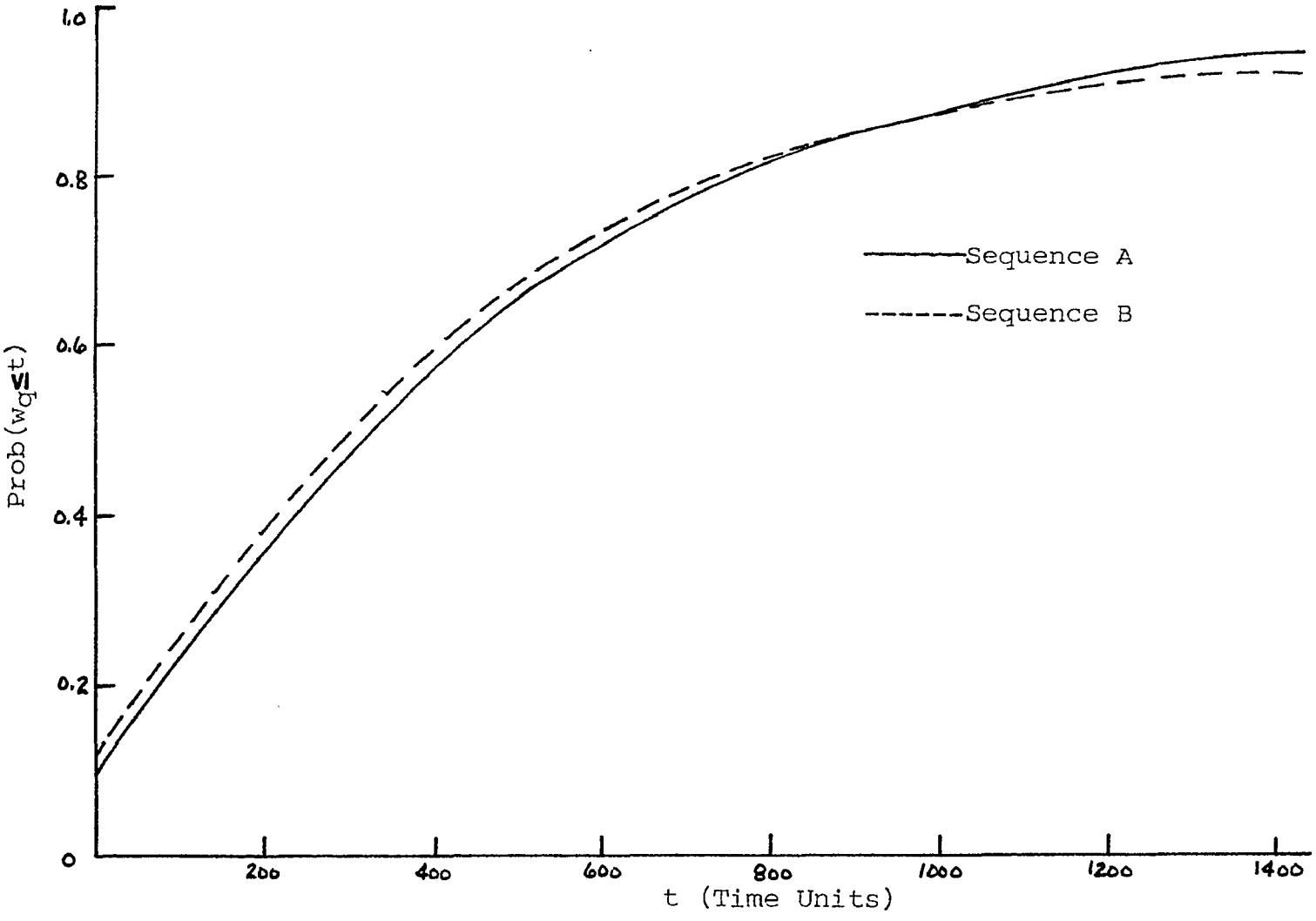


Fig. 39. A Comparison of Cumulative Distributions under Sequences A and B for Series System 10

APPENDIX C

CUMULATIVE DISTRIBUTIONS FOR THE DIFFERENCE IN CUSTOMER WAITING TIMES UNDER SEQUENCES A AND B FOR SYSTEMS 1 THROUGH 10

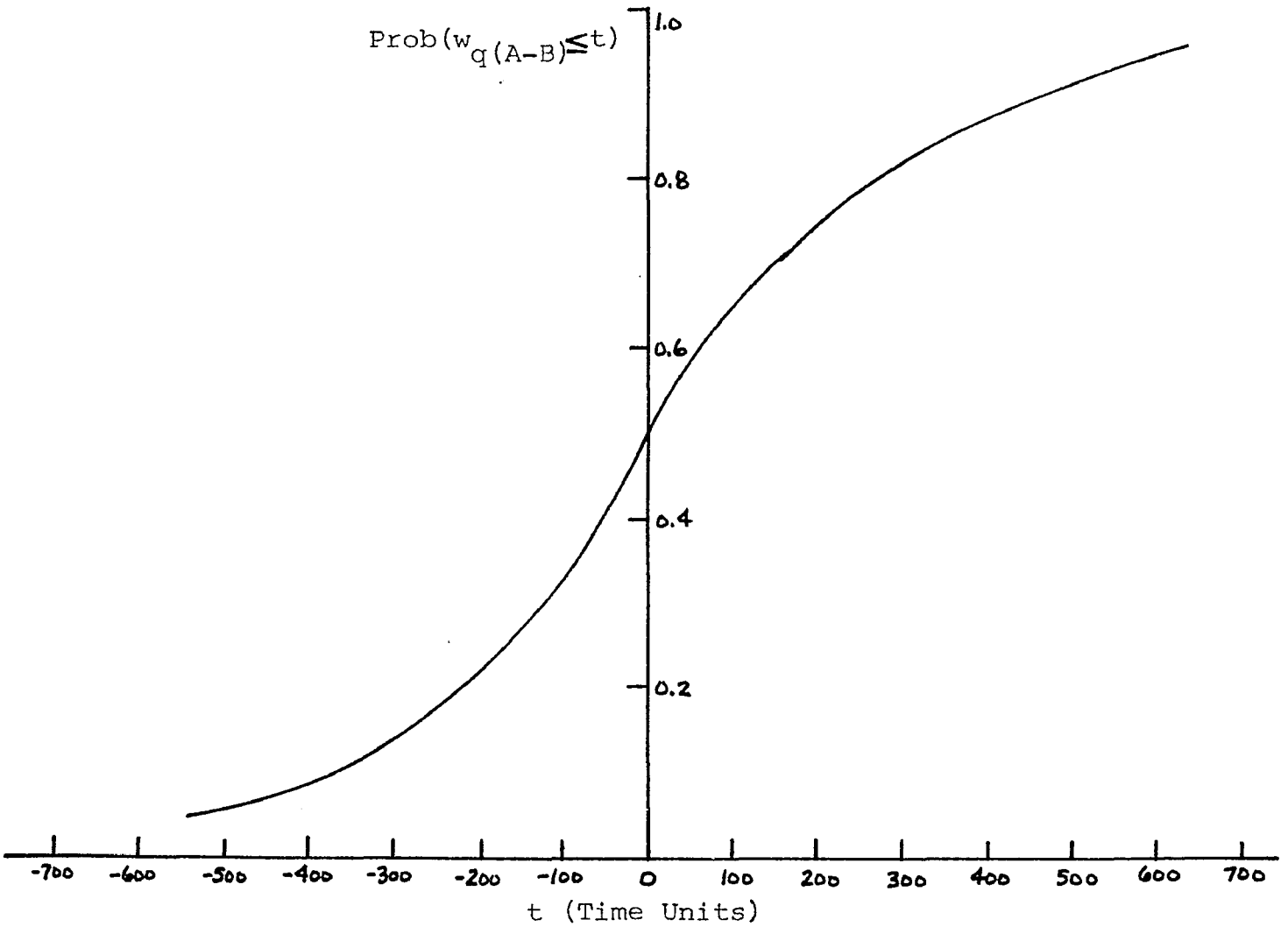


Fig. 40. The Cumulative Distribution for the Difference in Customer Waiting Times under Sequences A and B for System 1

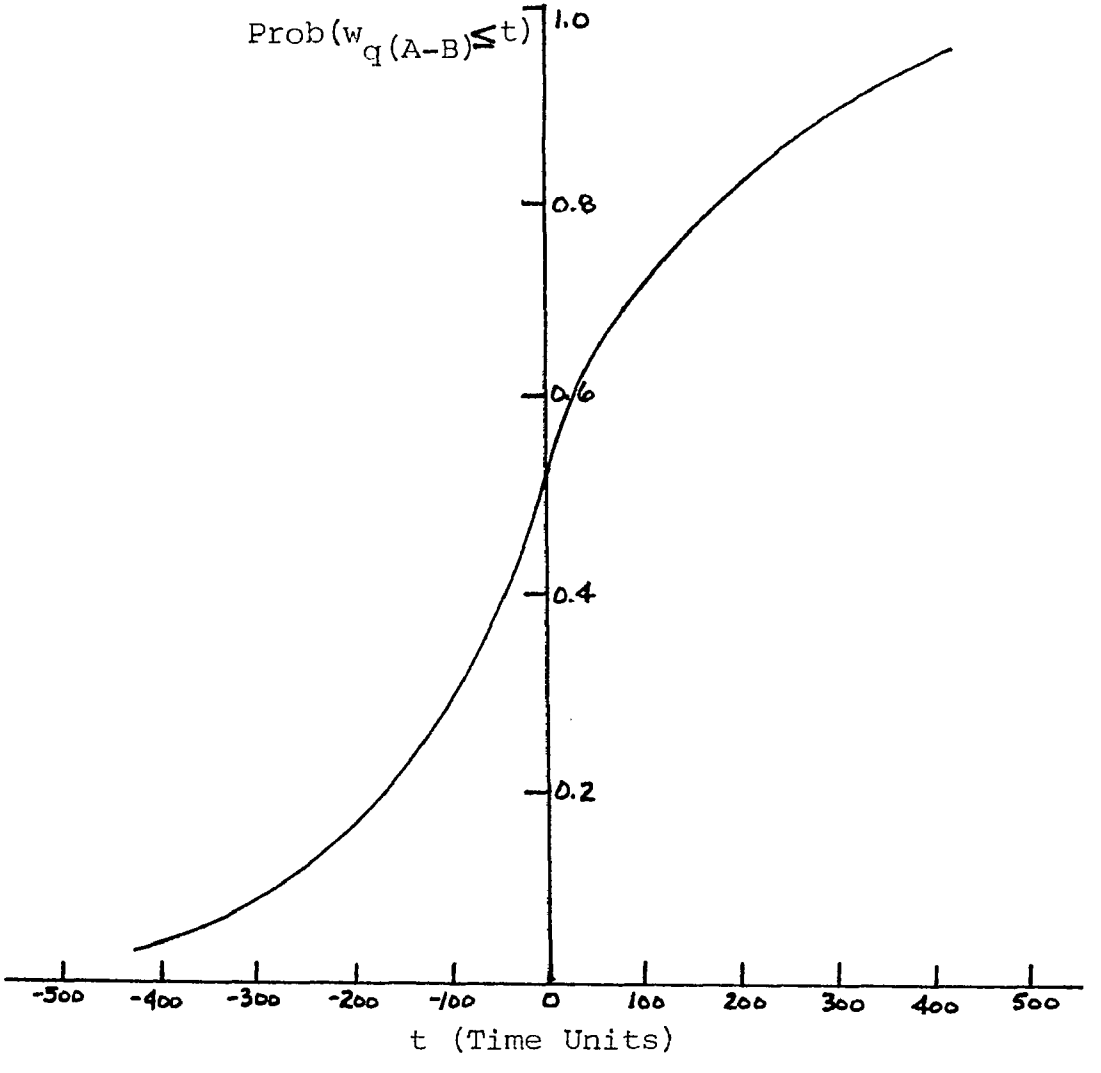


Fig. 41. The Cumulative Distribution for the Difference in Customer Waiting Times under Sequences A and B for System 2

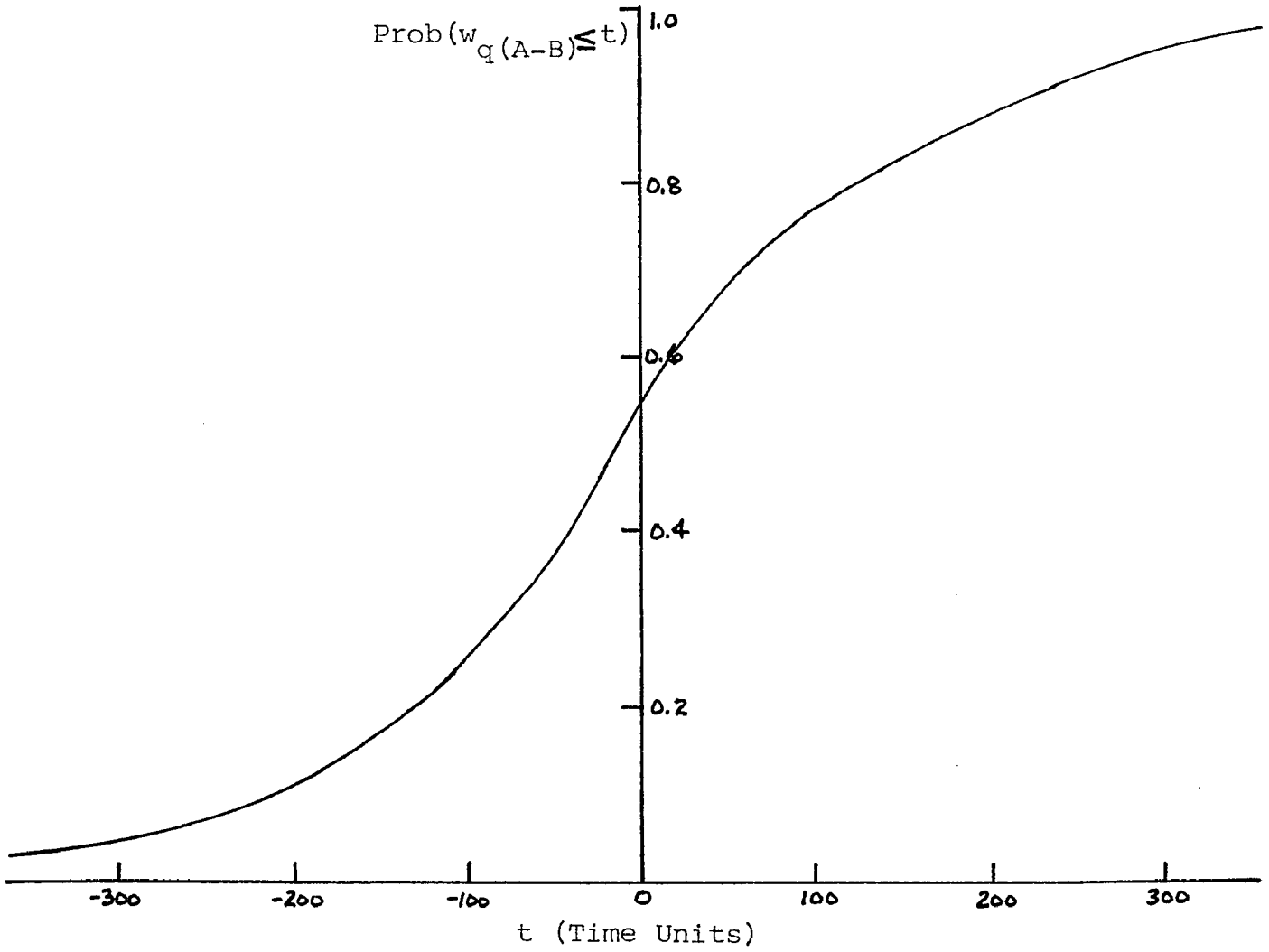


Fig. 42. The Cumulative Distribution for the Difference in Customer Waiting Times under Sequences A and B for System 3

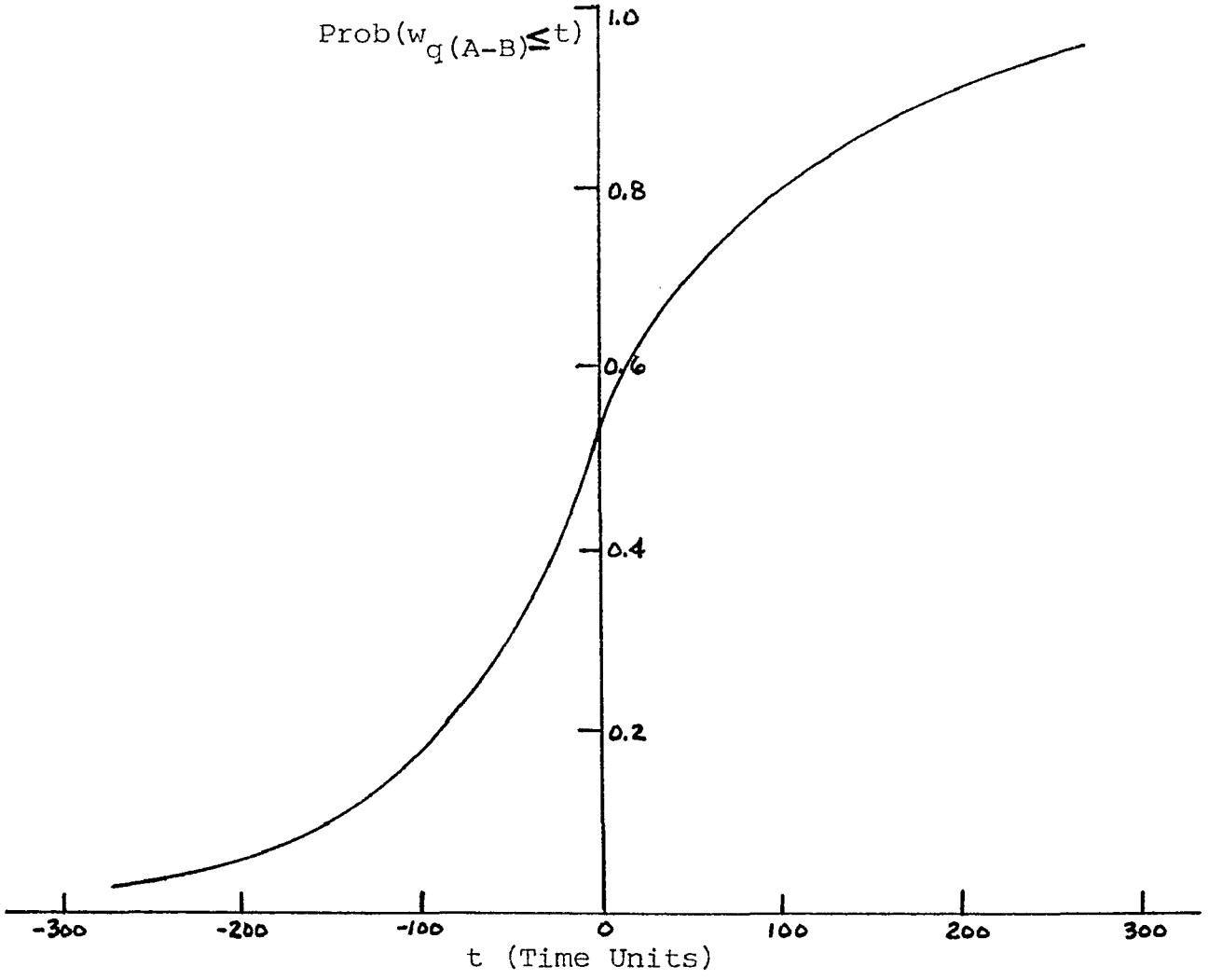


Fig. 43. The Cumulative Distribution for the Difference in Customer Waiting Times under Sequences A and B for System 4

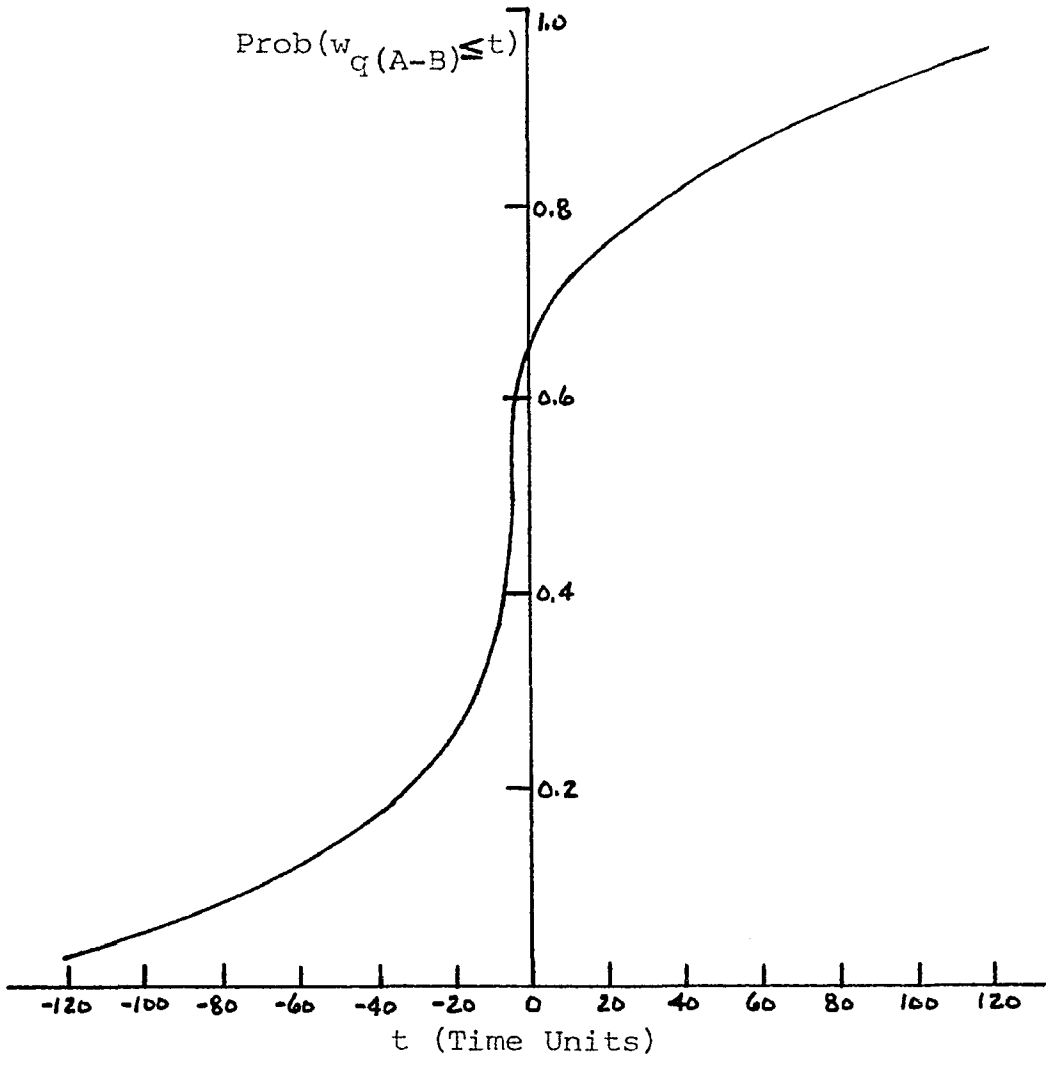


Fig. 44. The Cumulative Distribution for the Difference in Customer Waiting Times under Sequences A and B for System 5

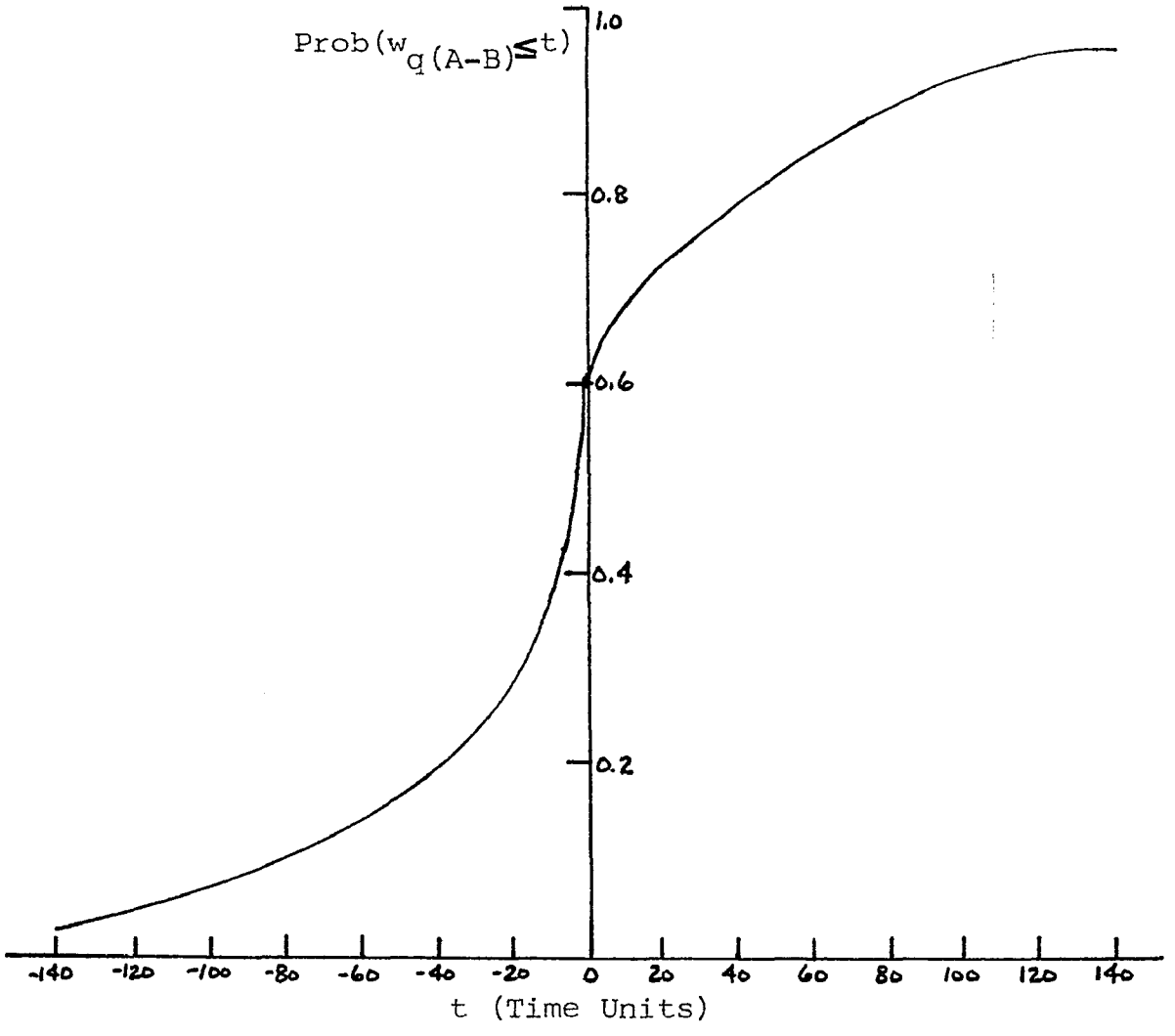


Fig. 45. The Cumulative Distribution for the Difference in Customer Waiting Times under Sequences A and B for System 6

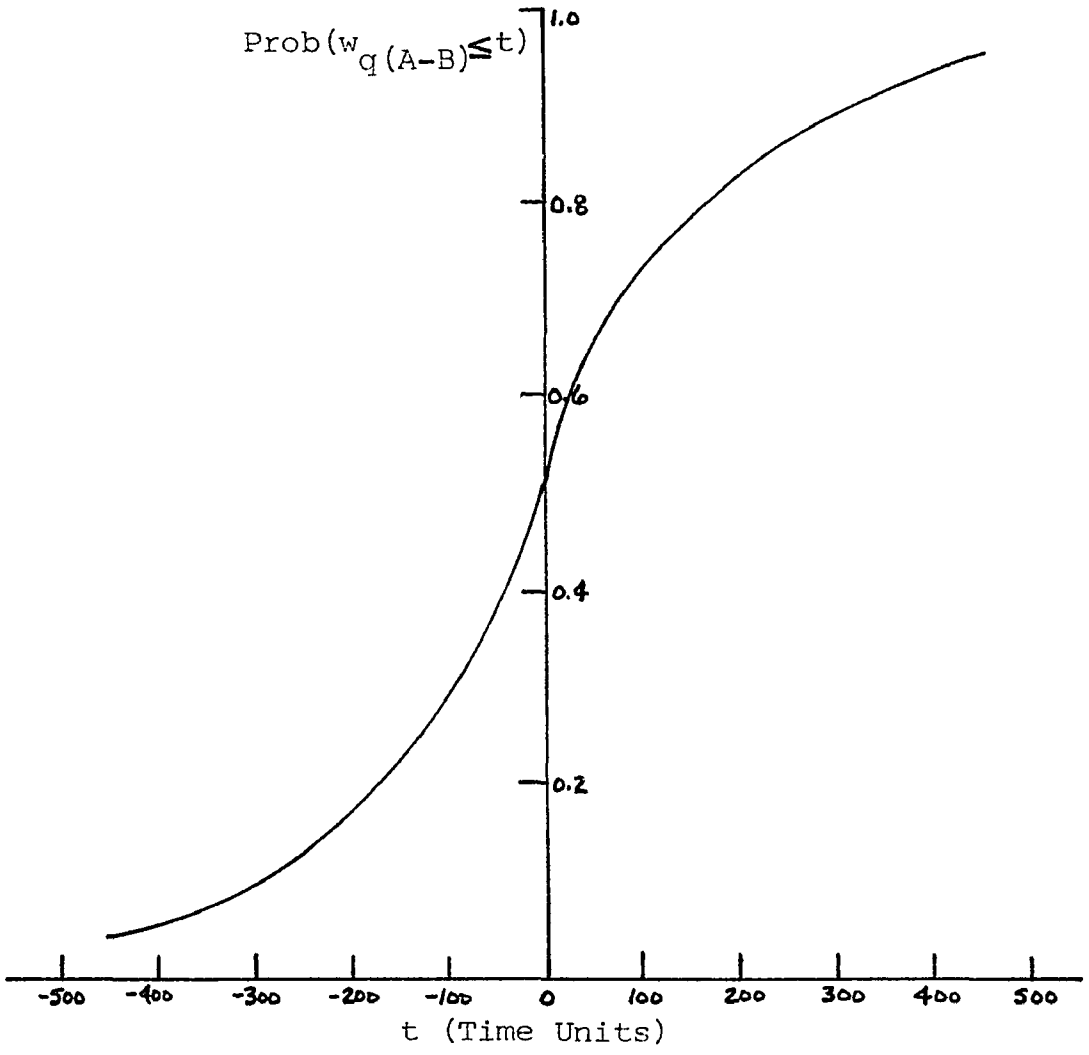


Fig. 46. The Cumulative Distribution for the Difference in Customer Waiting Times under Sequences A and B for System 7

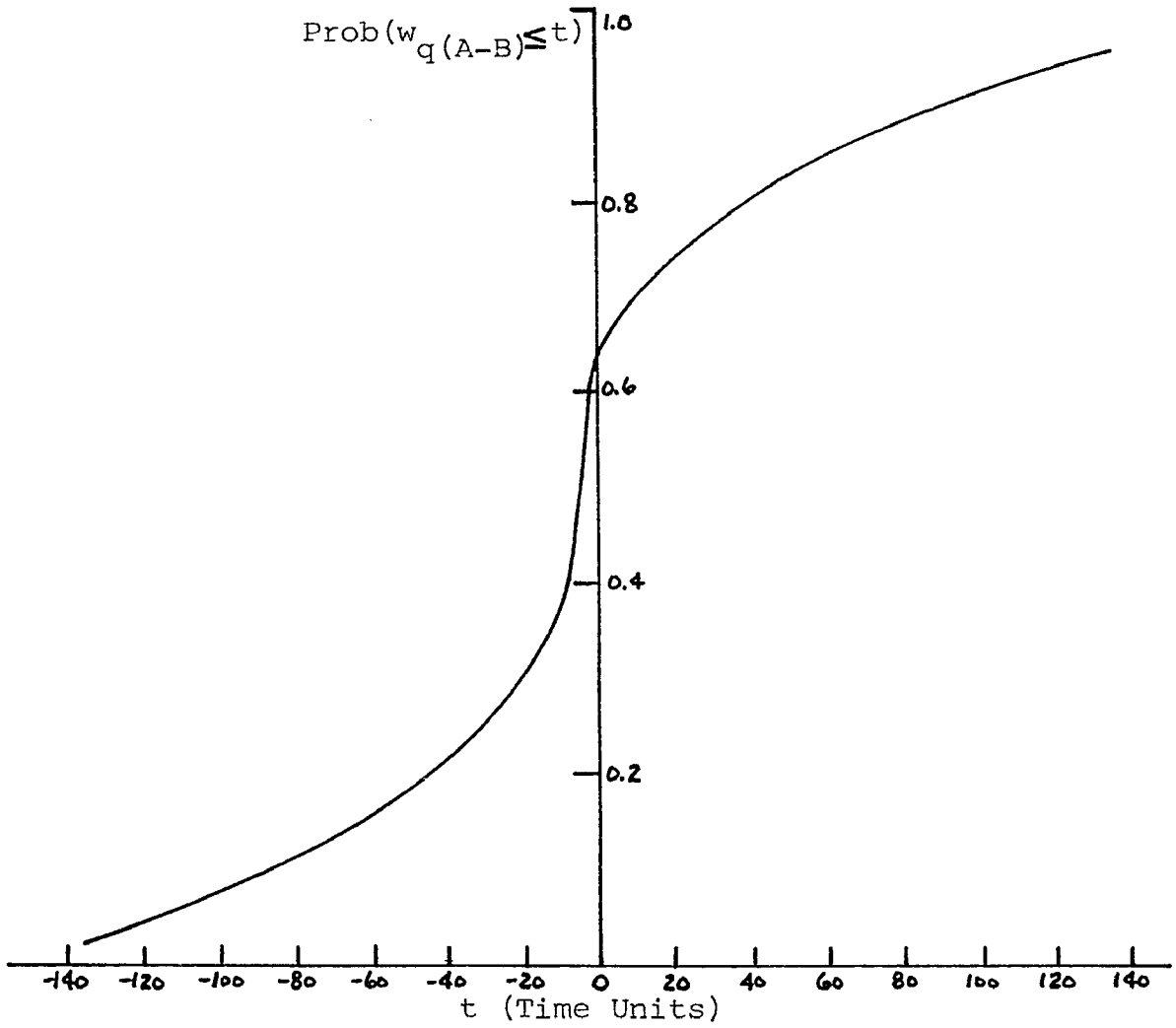


Fig. 47. The Cumulative Distribution for the Difference in Customer Waiting Times under Sequences A and B for System 8

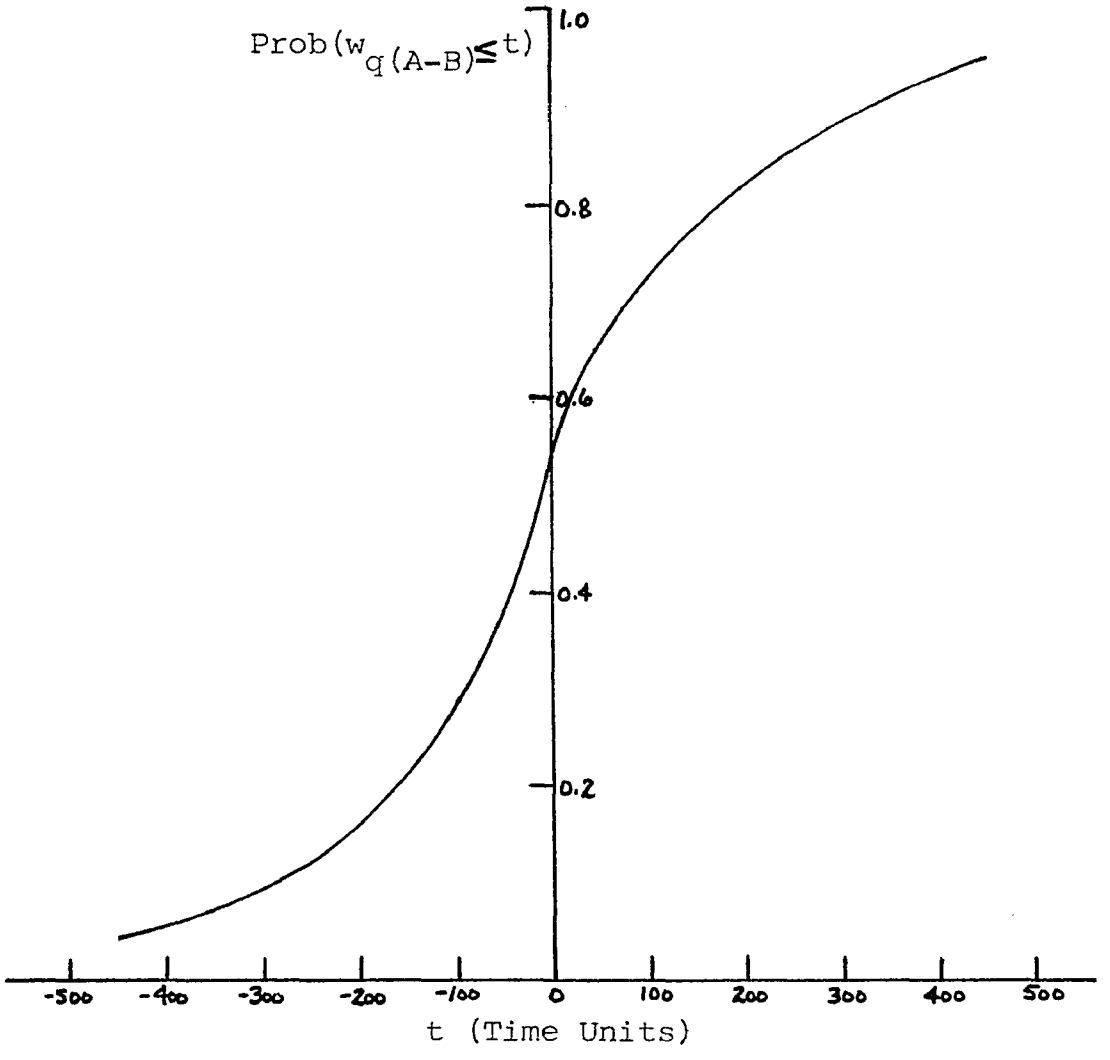


Fig. 48. The Cumulative Distribution for the Difference in Customer Waiting Times under Sequences A and B for System 9

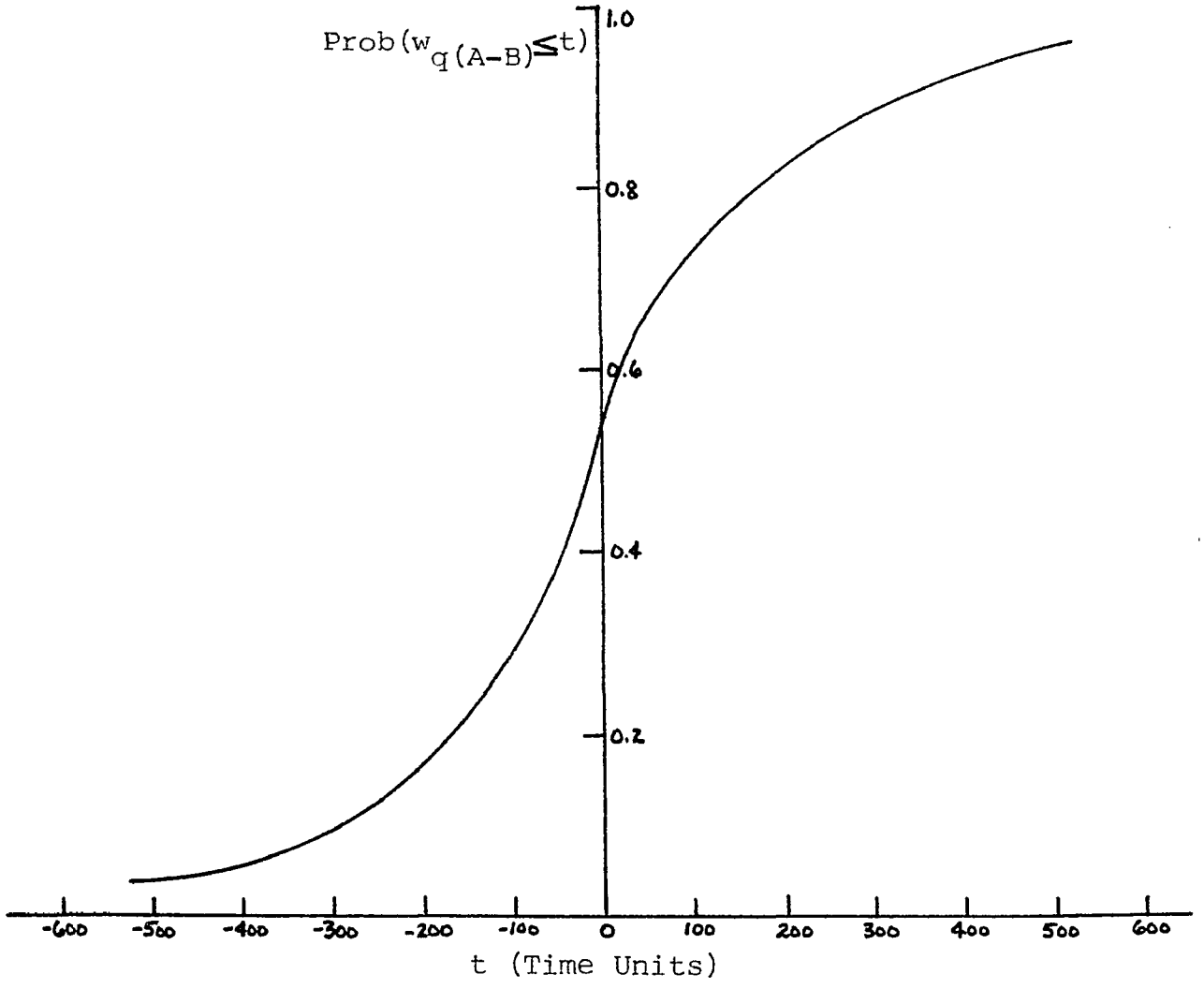


Fig. 49. The Cumulative Distribution for the Difference in Customer Waiting Times under Sequences A and B for System 10

APPENDIX D

ANALYSIS OF THE THIRD ORDER INDIFFERENCE EQUATION FOR ERLANG SERVICE DISTRIBUTIONS

ANALYSIS OF THE THIRD ORDER INDIFFERENCE EQUATION FOR ERLANG SERVICE DISTRIBUTIONS

For Erlang service distributions, the general equation for indifference between sequences on the basis of mean waiting time, from observation, appears to be

$$s(1-\rho_2) - t(1-\rho_1) + (u+v\rho_1^2-w\rho_2^2)(1-\rho_1)(1-\rho_2) = 0 \quad (A1)$$

where $s,t,v,w>0$. This is expressed by equations (4.3), (4.4), and (4.5).

Letting $Y=1-\rho_1$, $Z=1-\rho_2$, and $r=u+v-w$, then equation (A1) can be rewritten as

$$sZ - tY + YZ(r-2vY+vY^2+2wZ-wZ^2) = 0$$

or

$$wZ^3 - 2wZ^2 + \left[-(s/Y)-r+2vY-vY^2\right]Z + t = 0. \quad (A2)$$

Differentiating with respect to Y gives

$$3wZ^2dZ/dY - 4wZdZ/dY + \left[-(s/Y)-r+2vY-vY^2\right]dZ/dY + \left[(s/Y^2)+2v-2vY\right]Z = 0.$$

Solving for dZ/dY yields

$$dZ/dY = -\left[(s/Y^2)+2v(1-Y)\right]Z/\left\{3wZ^2-4wZ+\left[-(s/Y)-r+2vY-vY^2\right]\right\}$$
$$= -\left[(s/Y^2)+2v(1-Y)\right]Z^2/\left\{3wZ^3-4wZ^2+\left[-(s/Y)-r+2vY-vY^2\right]Z\right\}.$$

From equation (A2)

$$\left[-(s/Y)-r+2vY-vY^2\right]Z = -(t+wZ^3-2wZ^2).$$

Hence

$$dZ/dY = -\left[(s/Y^2)+2v(1-Y)\right] Z^2/-(t-2wZ^3+2wZ^2)$$
$$= \left[(s/Y^2)+2v(1-Y)\right]/\left[(t/Z^2)+2w(1-Z)\right]. \qquad (A3)$$

Since $0<Y<1$, and $0<Z<1$, and s,t,v, and w are all positive, the numerator and denominator in equation (A3) are always positive. Therefore $dZ/dY>0$. Also

$$dZ/dY = d(1-\rho_2)/d(1-\rho_1) = -d\rho_2/-d\rho_1 = d\rho_2/d\rho_1>0.$$

The third order indifference equation has a positive slope everywhere and the curve is constantly increasing.

APPENDIX E

DERIVATION OF BOUNDS ON WAITING TIME VARIABILITY IN SOME QUEUEING SYSTEMS

DERIVATION OF BOUNDS ON WAITING TIME VARIABILITY IN SOME QUEUEING SYSTEMS

Sphicas and Shimshak studied waiting time variability in queueing systems. Their results were based on an evaluation of the square of the coefficient of variation of the waiting time in the $M/E_k/1$ and M/H/1 Systems.

For system $M/E_k/1$, the square of the coefficient of variation of the waiting time is given in equation (5.1) as

$$C_{w_q}^2 = (8+4k-5\rho-\rho k)/3\rho(k+1),$$

which can be rewritten as

$$C_{w_q}^2 = 1 + (4/3)\left[(k+2)/(k+1)\right]\left[(1-\rho)/\rho\right].$$

Since k is restricted to values greater than or equal to one, limits on $C_{w_q}^2$ are found. When k=1, the fraction (k+2)/(k+1) is equal to 3/2 and the upper bound on $C_{w_q}^2$ develops. When k approaches infinity, the fraction (k+2)/(k+1) approaches 1 and determines the lower bound on $C_{w_q}^2$. These bounds are given in equation (5.9) as

$$1 + \tfrac{4}{3}\left[(1-\rho)/\rho\right] \leq C_{w_q}^2 \leq 1 + 2\left[(1-\rho)/\rho\right].$$

$C_{w_q}^2$ in the M/H/1 system is shown in equation (5.2) as

$$C_{w_q}^2 = \left[\rho+4(1-\rho)(1-2\gamma+2\gamma^2)\right]/\rho.$$

This can be rewritten as

$$C_{w_q}^2 = 1 + 4(1-2\gamma+2\gamma^2)\left[(1-\rho)/\rho\right].$$

The parameter γ is restricted to values between 0 and $\frac{1}{2}$.

When $\gamma=\frac{1}{2}$, the expression $(1-2\gamma+2\gamma^2)$ is equal to $\frac{1}{2}$ and determines the lower bound on $C^2_{w_q}$. As γ approaches 0, the expression $(1-2\gamma+2\gamma^2)$ approaches 1 and the upper bound on $C^2_{w_q}$ develops. These are expressed in equation (5.10) as

$$1+ 2\left[(1-\rho)/\rho\right] \leq C^2_{w_q} \leq 1 + 4\left[(1-\rho)/\rho\right].$$

Note that the upper bound in the $M/E_K/1$ system (when k=1) is equal to the lower bound in the M/H/1 system (when $\gamma=\frac{1}{2}$). These are both cases where the service distribution is exponential. Thus the M/M/1 system is the boundary case between $M/E_K/1$ and M/H/1.

BIBLIOGRAPHY

Ashour, S. Sequencing Theory. Lecture Notes in Economics and Mathematical Systems, vol. 69. Heidelberg and New York: Springer-Verlag, 1972.

Avi-Itzhak, B. "A Sequence of Service Stations with Arbitrary Input and Regular Service Times." Management Science 11 (March 1965):553-64.

Avi-Itzhak, B., and Yadin, M. "A Sequence of Two Servers with No Intermediate Queue." Management Science 11 (March 1965):565-71.

Barten, Kenneth. "A Queueing Simulator for Determining Optimum Inventory Levels in a Sequential Process." Journal of Industrial Engineering 13 (July-August 1962):245-52.

Burke, Paul J. "The Output of a Queueing System." Operations Research 6 (December 1956):699-704.

_____. "The Dependence of Delays in Tandem Queues." Annals of Mathematical Statistics 35 (June 1964): 874-75.

_____. "The Input process of a Stationary M/M/s Queueing System." Annals of Mathematical Statistics 39 (August 1968):1144-52.

_____. "The Dependence of Sojourn Times in Tandem M/M/s Queues." Operations Research 17 (July-August 1969):754-55.

Cochran, William G. Sampling Techniques. 2d ed. New York: John Wiley & Sons, 1963.

Conway, R. W. "Some Tactical Problems in Digital Simulation." Management Science 10 (October 1963):47-61.

Conway, R. W.; Johnson, B. M.; and Maxwell, W. L. "Some Problems of Digital Systems Simulation." Management Science 6 (October 1959):92-110.

Cox, D. R., and Smith, Walter L. Queues. London: Methuen & Co., 1961.

Crane, Michael A., and Iglehart, Donald L. "Simulating Stable Stochastic Systems, I: General Multiserver Queues." Journal of the Association of Computing Machinery 21 (January 1974):103-13.

DeBaum, R. M., and Katz, S. "An Approximation to Distributions of Summed Waiting Times." Operations Research 7 (November-December 1959):811-13.

Finch, P. D. "The Output Process of the Queueing System M/G/1." Journal of the Royal Statistical Society, ser. B, 21 (1959):375-80.

Fishman, George S. "Statistical Analysis for Queueing Simulations." Management Science 20 (November 1973): 363-69.

Fraker, John R. "Approximate Techniques for the Analysis of Tandem Queueing Systems." Ph.D. dissertation, Clemson University, 1971.

Freeman, Michael C. "The Effects of Breakdowns and Interstage Storage on Production Line Capacity." Journal of Industrial Engineering 15 (July-August 1964): 194-200.

Friedman, Henry D. "Reduction Methods for Tandem Queueing Systems." Operations Research 13 (January-February 1965):121-31.

Gafarian, A. V., and Ancker, C. J., Jr. "Mean Value Estimation from Digital Computer Simulation." Operations Research 14 (January-February 1966):25-44.

Ghosal, A. "Queues in Series." Journal of the Royal Statistical Society, ser. B, 24 (1962):359-63.

_____. Some Aspects of Queueing and Storage Systems. Lecture Notes in Operations Research and Mathematical Systems, vol. 23. Heidelberg and New York: Springer-Verlag, 1970.

_____. "Some Problems in Applied Cybernetics." SCIMA 2 (1973):35-50.

_____. "Isomorphic Queueing Systems." Paper presented at the International Conference on Stochastic Processes at the University of Maryland, 1975. (Mimeographed.)

Ghosal, A., ed. "Isomorphic Queueing Systems and Related Problems." Working Paper G 1/76, Graduate Center of Management, Baruch College, New York, 1976.

Goode, Henry P., and Saltzman, S. "Estimating Inventory Limits in a Station Grouped Production Line." Journal of Industrial Engineering 13 (November-December 1962):484-90.

Greenberg, Stanley. GPSS Primer. New York: John Wiley & Sons, 1972.

Hadar, Josef, and Russell, William R. "Rules for Ordering Uncertain Prospects." American Economic Review 59 (March 1969):25-34.

Hays, William L., and Winkler, Robert L. Statistics. Vol. II. New York: Holt, Rinehart & Winston, 1970.

Hillier, Frederick S., and Lieberman, Gerald J. Introduction to Operations Research. San Francisco: Holden-Day, 1967.

_____. Operations Research. 2d ed. San Francisco: Holden-Day, 1974.

Jackson, James R. "Networks of Waiting Lines." Operations Research 5 (August 1957):518-21.

Jackson, R. R. P. "Queueing Systems with Phase Type Service." Operational Research Quarterly 5 (December 1954):109-20.

_____. "Random Queueing Processes with Phase Type Service." Journal of the Royal Statistical Society, ser. B, 18 (1956):129-32.

Kabak, Irwin W. "Stopping Rules for Queueing Simulations." Operations Research 16 (March-April 1968):431-37.

Kendall, Maurice G., and Stuart, Alan. The Advanced Theory of Statistics. 5th ed. New York: Hafner Publishing, 1952.

Kiviat, Philip J. "Simulation Languages." In Thomas H. Naylor, ed. Computer Simulation Experiments with Models of Economic Systems. New York: John Wiley & Sons, 1971.

Kleijnen, Jack P. C. Statistical Techniques in Simulation. Part 1. New York: Marcel Dekker, 1974.

Koenigsberg, Ernest. "Cyclic Queues." Operations Research Quarterly 9 (March 1958):22-35.

_____. "Production Lines and Internal Storage--A Review." Management Science 5 (July 1959):410-33.

Law, Averill M. "Efficient Estimators for Simulated Queueing Systems." Management Science 22 (September 1975):30-41.

Marshall, K. T. "Some Inequalities in Queueing." Operations Research 16 (May-June 1968):651-65.

Meier, Robert C.; Newell, William T.; and Pazer, Harold L. Simulation in Business and Economics. Englewood Cliffs, New Jersey: Prentice-Hall, 1969.

Mihram, G. A. Simulation: Statistical Foundations and Methodology. New York: Academic Press, 1972.

Morse, Philip M. Queues, Inventories and Maintenance. New York: John Wiley & Sons, 1958.

Nelson, Rosser T. "Waiting-Time Distributions for Application to a Series of Service Centers." Operations Research 6 (November-December 1958):856-62.

_____. "A Simulation Study and Analysis of a Two Station, Waiting-Line Network Model." Ph.D. dissertation, UCLA, 1965.

Page, E. Queueing Theory in OR. New York: Crane Russak & Co., 1972.

Reich, Edgar. "Waiting Times When Queues Are in Tandem." Annals of Mathematical Statistics 28 (September 1957):768-73.

Richman, Eugene, and Elmaghraby, Salah. "The Design of In-Process Storage Facilities." Journal of Industrial Engineering 8 (January-February 1957):7-9.

Riordan, John. Stochastic Service Systems. New York: John Wiley & Sons, 1962.

Rolski, Tomasz, and Stoyan, Dietrich. "On the Comparison of Waiting Times in GI/G/1 Queues." Operations Research 24 (January-February 1976):197-200.

Rosenshine, Matthew, and Chandra, M. Jeya. "Approximate Solutions for Some Two-Stage Tandem Queues, Part 1: Individual Arrivals at the Second Stage." Operations Research 23 (November-December 1975):1155-66.

Saaty, Thomas L. Elements of Queueing Theory. New York: McGraw-Hill, 1961.

Silverman, Fred N. "The Effects of Stochastic Work Times on the Assembly Line Balancing Problem." Ph.D. dissertation, Columbia University, 1974.

Sphicas, Georghios P., and Shimshak, Daniel G. "Waiting Time Variability in Some Simple Queueing Systems." Working Paper, Graduate Center of Management, Baruch College, New York, 1976.

Tembe, Shantanu V., and Wolff, Ronald W. "The Optimal Order of Service in Tandem Queues." Operations Research 22 (July-August 1974):824-32.

Tonge, Fred M. "Assembly Line Balancing Using Probabilistic Combinations of Heuristics." Management Science 11 (May 1965):727-35.

Whitmore, G. A. "Third-Degree Stochastic Dominance." American Economic Review 60 (June 1970):457-59.