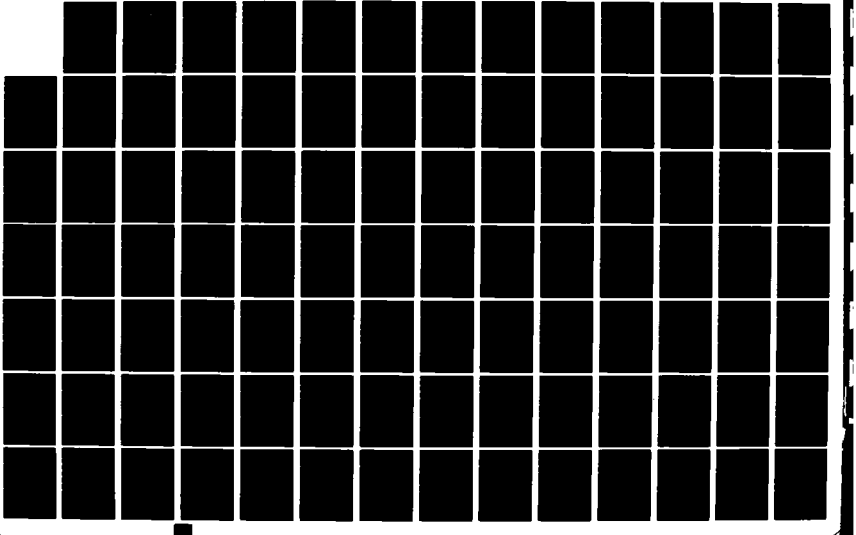
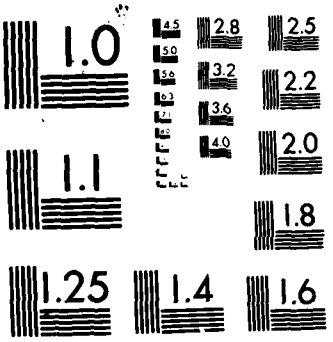
AD-A124 877 A NEW APPROACH FOR SOLVING SIMULTANEOUS RESOURCE POSSESSION PROBLEMS IN C.. (U) AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOOL OF ENGI.. D J FREUND

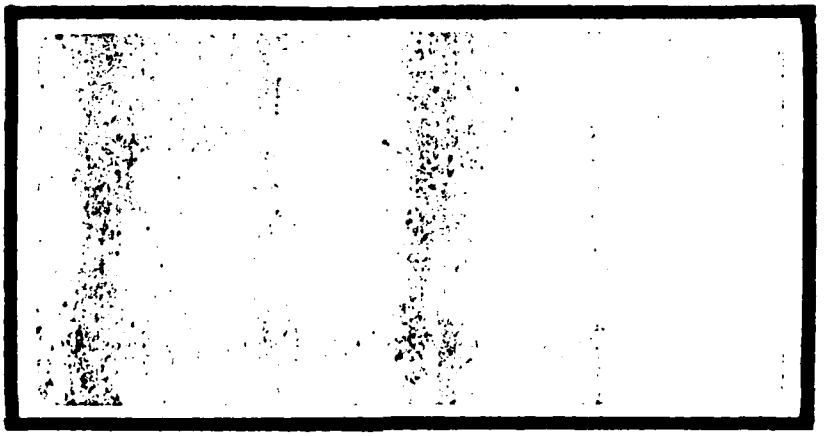
1/3

UNCLASSIFIED DEC 82 AFIT/GCS/MA/82D-2 F/G 12/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY (ATC)

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

(1)

A NEW APPROACH FOR SOLVING SIMULTANEOUS RESOURCE POSSESSION PROBLEMS IN CLOSED QUEUEING NETWORKS THESIS

Darral J. Freund

AFIT/GCS/MA/82D-2

Capt USAF

AFIT/GCS/MA/82D-2

A NEW APPROACH FOR SOLVING SIMULTANEOUS RESOURCE POSSESSION PROBLEMS IN CLOSED QUEUEING NETWORKS

THESIS

Presented to the Faculty of the School of Engineering

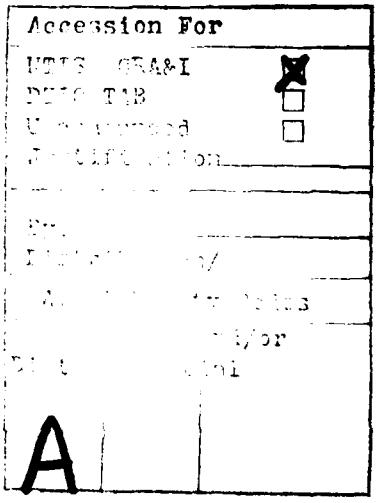
of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the

Requirements for the Degree of

Master of Science



by

Darral J. Freund

Capt USAF

Graduate Computer Science

December 1982

Approved for public release; distribution unlimited.

Preface

This report is the result of an effort to develop a new analytically oriented technique to approximately solve a class of queueing networks that have computer applications. Such a procedure was developed and is presented in this document. It is my hope that this procedure will be beneficial in the performance evaluation of computer systems.

One surprising aspect of this thesis was the magnitude of the job to prepare the manuscript. With 84 figures and 43 tables, the preparation of each iteration for review and rework was a major and time consuming necessity. I would especially like to thank my parents who contributed both their time and their graphics, clerical and word processing facilities in helping to solve that problem.

Finally, I would like to thank my thesis advisor, James N. Bexfield, for his support and the many hours he spend tutoring me in closed queueing network theory, and Dr. W. Ericksen who generously contributed his time in helping to find the symbolic solution of a set of homogeneous linear equations.

Contents

List of Figures

List of Tables

Abstract

A new approximate technique is presented for analyzing closed queueing networks with simultaneous resource possession. It is an analytic non-iterative solution procedure that is suitable for multiple entry systems such as I/O models. It relies on solving a closed queueing network consisting only of the subsystem with simultaneous resource possession, where queues can have service rates that are a function of the utilization of all the queues (number of busy servers). The submodel is solved using a newly discovered product form solution. Results are compared with simulation models and other available analytic techniques.

A NEW APPROACH FOR SOLVING SIMULTANEOUS RESOURCE POSSESSION PROBLEMS IN CLOSED QUEUEING NETWORKS

I. Introduction

As computers have become more and more complex, it has become much more difficult to determine their capabilities, performance, and limitations. Yet the need for this has greatly increased with the wide spread use of computer systems.

The easiest and most direct approach to computer performance evaluation has been to simply measure the desired quantities of interest. However, this has become more and more difficult, and impractical or infeasible in many cases. In these cases performance estimates must be predicted through modeling or other techniques.

Most modeling of computer systems has been either simulation or analytically oriented. Simulation modeling has the advantage that the computer system can be modeled to arbitrary accuracy. However, simulation programs tend to be specific to a particular application, take time to develop, and use large amounts of computer time. Analytic methods on the other hand can be developed quickly and tend to be usable in a variety of different situations. However, analytic models of reasonable complexity are intractable unless certain restrictions are imposed. Only within the the last ten to fifteen years has the analytic solution to a certain class of nontrivial closed queueing networks become tractable. This was made possible by the discovery of a closed form solution to the state prob-

abilities. Analytic solutions to most queueing networks was possible, however, it required the enumeration and solution of a large set of homogeneous linear equations with properly chosen boundary conditions.

Much research has been devoted to expanding the types of networks that have closed or product form solutions, but the conditions required for closed form are still sometimes overly restrictive. Furthermore, many realistic computer models contain networks that have properties which usually violate the product form. As a result research in the area of approximation has proceeded rapidly. However, the techniques developed have either been developed for a specific application, or have been somewhat unpredictable in their results. Most of the methods are still in the test and evaluation stages and need more error analysis before confidence can be gained in order to use them in a stand along basis.

One class of networks that has been difficult to model analytically is a class of closed queueing networks that have simultaneous resource possession. In networks with simultaneous resource possession, customers obtain and use more than one resource simultaneously. For example, consider a line of customers in a bank that has a limited number of tellers available. Since many banking operations are computerized the tellers may have to enter transactions in a terminal. If there are less terminals than tellers, then it is possible that the tellers might have to queue for the use of the terminal before gaining access to the required records. While the teller is using the terminal, the process can be considered as the customer being serviced simultaneously by the teller and the terminal.

It is this concept that is called simultaneous resource possession. This thesis will:

1. Classify all possible types of simultaneous resource possession that can occur in closed queueing networks,

2. Briefly review the current procedures available that can solve the general classes of simultaneous resource possession problems,

3. Present a new approximate procedure that solves a class of simultaneous resource possession problems that have been difficult to solve previously, and

4. Identify the error in the procedure and when it will be prone to occur.

To accomplish this, Chapter Two will outline notation, symbology, and terminology used in the subsequent chapters. Chapter Three will define and categorize the types of simultaneous resource possible as well as review two currently available procedures for solving them. Several examples will be given of their use. Chapter Four will develop and illustrate the new procedure and give several examples of its use. Chapter Five will discuss the potential sources of error, and Chapter Six will discuss conclusions and recommendations for subsequent research.

It is assumed that the reader is familiar with the theory and solution of a closed queueing networks.

II. Product Form Closed Queueing Networks

In this chapter the basic notation, symbolism and conventions that are used in the remainder of the thesis will be described. Some attempt is made to use standard symbolism, however many symbols are used to simplify the typing and representation on paper. Generally the network diagrams are similar to those used by Chandy and Sauer (Ref 1), and the terminology to that used by Jacobson and Lazowska (Ref 2).

Consider a closed single class queueing network with N customers and k queues.

Let SD_i be the mean service requirement or service demand at queue i

$Y_i(n)$ be the service rate called the unnormalized service rates of queue i with n customers at the queue

$[P_{ij}]$ Matrix of transition probabilities of a customer currently at queue i proceeding to queue j.

Let $\theta = (\theta_1,\ldots,\theta_k)$

where

$$\theta = \theta P.$$

The θ_i are the relative throughputs (also called loadings or visit ratios) and are only solvable to within a multiplicative constant. In this thesis the solution is made unique by selecting the relative throughput, θ_i,

of some conveniently chosen queue i to 1.0.

Let $s = (n_1,\ldots,n_k)$ be the state of the network where

n_i = number of customers at queue i, for each $i = 1,\ldots,k$.

Suppose that all states $s = (n_1,\ldots,n_k)$ are feasible where

$$n_1 + \ldots + n_k = N,$$

then the network has $\binom{N+k-1}{k-1}$ states

Further suppose that the probability transition matrix P is independent of the state of the network.

Then the state probabilities are given by

$$P(n_1,\ldots,n_k) = \frac{\prod_{i=1}^{k} f_i(n_i)}{G}, \qquad (2\text{-}1)$$

where

G = a normalizing constant to ensure conservation of probability,

and

$$f_i(n_i) = \frac{\theta_i^{n_i}}{\prod_{j=i}^{n_i} Y_i(j)}, \qquad (2\text{-}2)$$

provided that

1. If queue i has a first come first served service discipline (FCFS), then the service time distribution is exponential, or

2. If queue i has a last come first served service discipline (LCFS) or, processor sharing (PS), then the service time distribution has a rational Laplace transform, or

3. If queue i is an infinite server queue, then the service time distribution has a rational Laplace transform.

Networks that satisfy the above conditions are called product form network (Ref 3).

Let

$$U_i(n) = \frac{Y_i(n)}{SD_i} \quad . \tag{2-3}$$

The set of values $U_i(n)$, for $n = 1,\ldots,N$, will be called the normalized service rate function for queue i. Note that

$$U_i(1) = 1.0, \text{ for all queues } \quad i = 1,\ldots,k$$

Let

$P_i(n)$ = marginal probability of n customers residing at queue i.

Then the utilization of queue i, X_i, is given by

$$X_i = \sum_{n=1}^{N} \frac{U_i(n)}{U_i(N)} P_i(n), \tag{2-4}$$

the throughput at queue i is given by

$$T_i = \sum_{n=1}^{N} U_i(n) P_i(n), \tag{2-5}$$

the mean number of customers at queue i (including the customer in service), L_i

$$L_i = \sum_{n=1}^{N} n\, P_i(n) \tag{2-6}$$

and finally,

the residency or waiting time at queue i, R_i1, is given by

$$R_i = \frac{L_i}{T_i} \tag{2-7}$$

Suppose $\theta_j = 1.0$, then the cycle time, CYC, relative to queue j is given by

$$CYC = \sum_{i=1}^{k} \theta_i R_i. \tag{2-8}$$

The cycle time is the mean time that it takes between visits to queue j.

Note that

$$CYC = N/T_j \tag{2-9}$$

and can change depending on the reference queue.

If the normalized service rate function for queue i has the form

$$U_i(n) = 1.0, \text{ for } n = 1,\ldots,N,$$

then queue i is called a single server queue and will be represented in a network as illustrated in Figure 2-1.

Figure 2-1. Single Sever Queue

If the normalized service rate function for queue i has the form

$$U_i(n) = \begin{cases} n & n = 1,\ldots,c \\ c & n = c+1,\ldots,N \end{cases}$$

for some integer c, then queue i is called a multiple server (c server) queue and will be represented as shown in Figure 2-2.

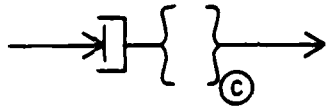


Figure 2-2. Multiple Sever Queue

If the normalized service rate function for queue i has the form

$$U_i(n) = n, \text{ for } n = 1,\ldots,N,$$

then queue i is called as infinite server queue and is represented as illustrated in Figure 2-3.

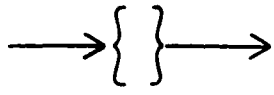


Figure 2-3. Infinite Server Queue

If the normalized service rate function for queue i contains noninteger values, then queue i is called a variable rate queue and is represented as illustrated in Figure 2-4

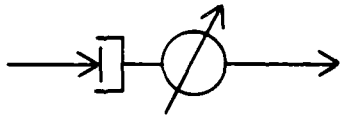


Figure 2-4. Variable Rate Queue

A queue that is nonproduct form because it has a FCFS service discipline and does not have an exponential service time distribution will be considered a block box queue. If it is a single server queue, it will be represented as illustrated in Figure 2-5.

Figure 2-5. Black Box Queue

If it is not a single server queue, it will be represented as shown in Figure 2-6.

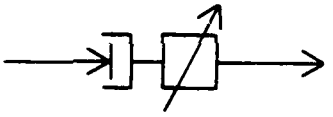


Figure 2-6. Variable Rate Black Box Queue

A subnetwork will be defined as a portion of the network such that

1. There is a single entry path into the subnetwork and out of the network, and

2. The relative throughput into the subnetwork equals the relative throughput out of the network.

The remaining portion of the network will be called the remainder network. These will be represented as illustrated in Figure 2-7

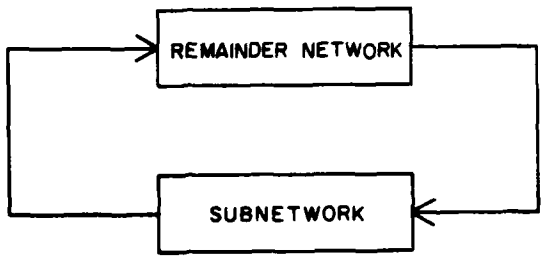


Figure 2-7. Original Network

The representation illustrated will not be limited to product form networks. For example the subnetwork may represent part of the original network that has properties that destroy product form such as blocking, state dependent routing, balking, etc. The variable rate black box queue illustrated in Figure 2-6 may exactly represent the subnetwork in the original network (Figure 2-7), as illustrated in Figure 2-8.

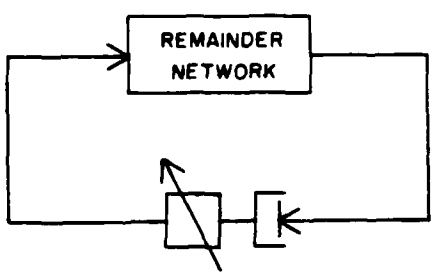


Figure 2-8. Original Network With the Subnetwork Replaced by a Variable Rate Black Box Queue

Blocking and population constraints imposed on part of the network will be represented by the use of resources. Customers will be allocated a resource at a resource allocation queue, sometimes called entry queue, as illustrated in Figure 2-9, for a queue that allocates a single resource, and Figure 2-10 that allocates multiple, say c, resources. No service requirement is associated with the resource allocation.

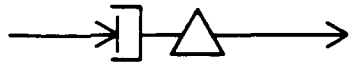


Figure 2-9. Single Resource Allocation Queue

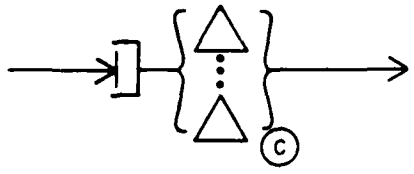


Figure 2-10. Multiple Resource Allocation Queue

Customers proceed through an arbitrary subnetwork after allocation, and are finally deallocated. This is illustrated as shown in Figure 2-11.

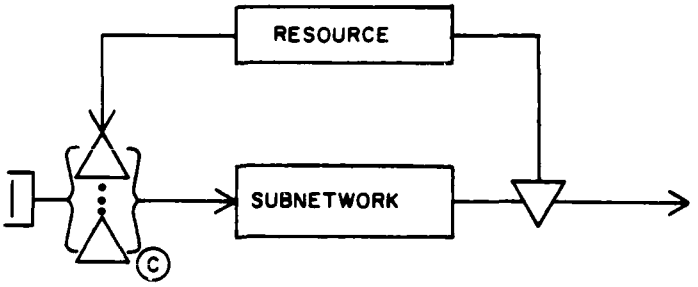


Figure 2-11. Single Queue Allocation Path

The reallocation path is shown when needed for clarity. Resources may be allocated from separate locations. If the resources are fixed to a certain entry queue, the representation in Figure 2-12 will be used.

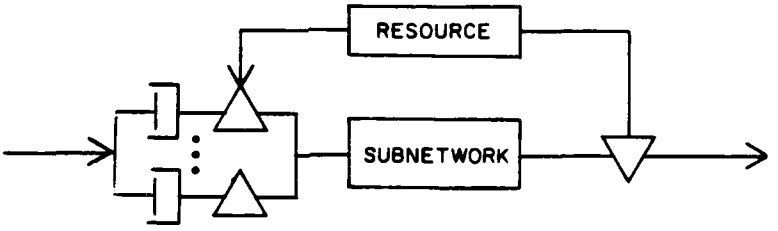


Figure 2-12. Single Allocation Queues in Parallel With a Common Deallocation Point

If there is no difference between the resources, then the representation in Figure 2-13 will be used.

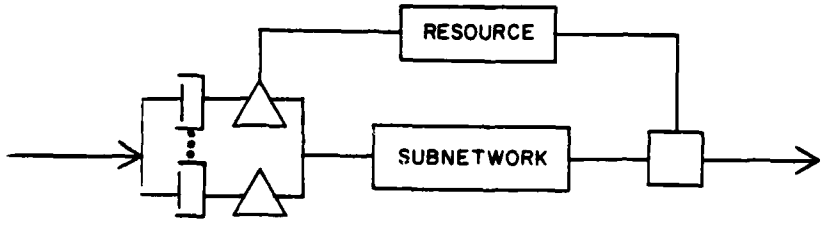


Figure 2-13. Multiple Resources Allocated From Separate Queues

It will be assumed that the reader is familiar with product form networks as defined by Baskett Chandy (Ref 3) and Norton's Theorem. Norton's Theorem sometimes called the decomposition or aggregation theorem (Ref 4) states that given a product form network, a subnetwork may be replaced by a variable rate queue with an appropriately chose.. set of service rates.

The service rates are chosen by evaluating the subnetwork with the remainder network removed. This is illustrated in Figure 2-14, and is called evaluating the subnetwork in isolation. The evaluation must be per-

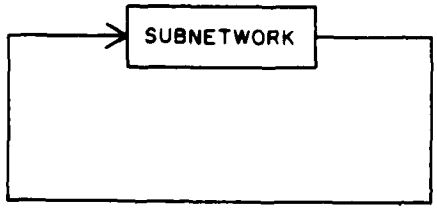


Figure 2-14. Subnetwork Evaluated in Isolation

formed for each feasible subnetwork population, n, and the throughput through the subnetwork, $T_{sn}(n)$, obtained. The unnormalized service rate of the variable rate queue, $Y_v(n)$, is determined by

$$Y_v(n) = T_{sn}(n) \quad \text{for } n = 1,\ldots,N.$$

The resulting network, illustrated in Figure 2-15, is equivalent to the original network, Figure 2-15, with the mean subnetwork population and residency time aggregated into the variable rate queue. Values of performance parameters in the subnetwork can be estimated by a procedure described in Chandy (Ref 4) and Sauer (Ref 5). It is assumed that the reader is familiar with this procedure as it will be used later in the thesis.

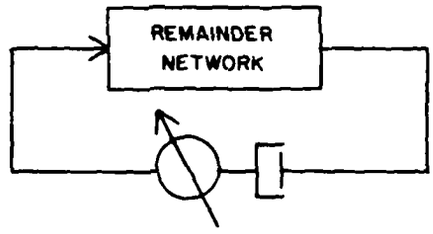


Figure 2-15. Original Network With the Subnetwork Replaced With a Variable Rate Queue (Final Network)

III. Queueing Methods In Simultaneous Resource Possession

Introduction

It frequently occurs that a job running in a computer will have control over multiple resources simultaneously. This is common in two cases: first, when the number of jobs running is limited by the multiprogramming level or memory constraints, and second, in I/O subsystems where part of an I/O operation requires use of a physical channel, and part does not. In the first case, which will be called the memory problem, the multiple resources are the memory, CPU and disks. The memory serves mainly as a population constraint limiting the number of jobs allowed in execution and contending for the CPU and disks. In the second case, which will be called the I/O Problem, the I/O subsystem consists of a number of disks and one or more channels (usually less than the number of disks) for transferring the data. Typically, part of the I/O operation (i.e. part of the rotational latency time and the total transfer time), requires use of the channel, and another part does not (i.e., the seek). Hence there is simultaneous possession of the disk and the channel for part of each I/O operation (and a population constraint imposed on the channel, equal to the number of disks). These instances of simultaneous resource possession are difficult to model analytically as they usually introduce blocking or state dependent routing into the queueing network, and hence invalidate the product form solution. As a result, direct solution techniques are computationally intractable for most nontrivial cases. However, ignoring the existence of simultaneous resource possession in a queueing network can cause highly inaccurate results.

The need for considering simultaneous resource possession can best be illustrated by an example. Consider the network in Figure 3-1, which illustrates a system with one CPU and four disks. A single channel is available for the transfer of the data and its use is required during both rotational latency and transfer phases. However, the disks are capable of seeking without the use of the channel. The multiprogramming level is limited to ten. Suppose that each customer is equally likely to require service at any of the disks and that the mean service demand at the CPU is 0.1, and at the disks 0.4 for the seek and 0.1 for the rotational latency and transfer. All service times are exponential. If the presence of the physical channel is ignored, the system can be solved easily by analytical methods as a central server queueing network with a mean disk service time of 0.5. The results are given in Table 3-1. When a disk I/O is begun a

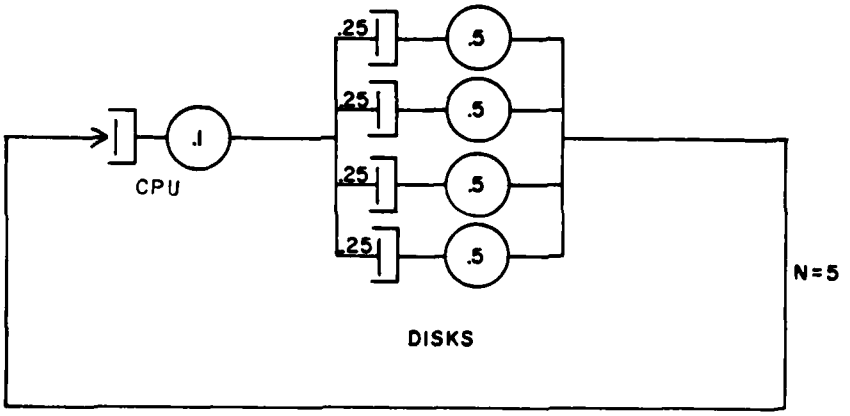


Figure 3-1. Central Sever Model

Table 3-1

Performance Results

	Analytic	Simulation	
Parameters (mean values)	Channel not considered	Channel Considered	97% Confidence Interval
Throughput	5.90	5.63	(5.51-5.75)
CPU Queue length	1.30	1.08	(0.99-1.16)
Total I/O time	1.48	1.58	(1.54-1.62)
Cycle time	1.70	1.77	(1.74-1.81)

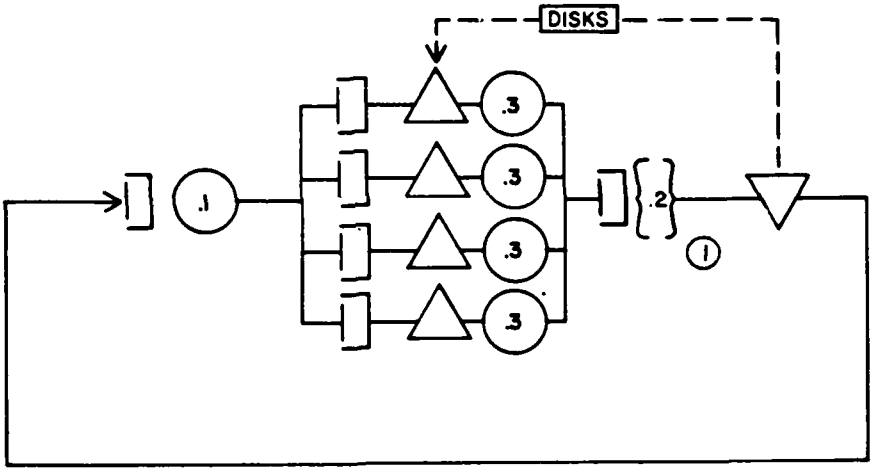


Figure 3-2. Central Sever Model With Expanded I/O Representation

seek is performed at one of the disks and can be performed simultaneously with seeks at the other disks. However, when the seek is completed it must wait for the channel to become available before beginning the rotational latency and transfer. During this time, all further seeks are blocked at that disk. This blocking destroys the product form solution of the network in Figure 3-2 and hence any solution techniques for product form networks will not provide exact results. It is possible to solve the network

analytically, but its solution would be nontrivial, and therefore simulation was used. The performance results obtained using simulation are also given in Table 3-1. Notice that there is a significant difference between the throughput of the two models (5.63 as opposed to 5.90), and between the other parameters as well.

The error caused by ignoring instances of simultaneous resource possession can be high, as in the above example, and has led to the devlopment of several approximation methods for dealing with it. In the reminder of this chapter the effects of simultaneous resource possession on closed queueing networks will be discussed, and the different possibilities classified into two categories. Three currently available methods will be presented and illustrated with several examples.

Types of Simultaneous Resource Possession

Simultaneous resource possession problems present themselves mostly in the form of the memory and I/O problems just described. However, other categories are becoming evident in computer systems as well as in completely different technical fields. By classifying instances of simultaneous resource possession into categories that are conceptually similar, the available methods can be studied and identified as being more or less suited to a particular type of problem. The solution procedure for a typical problem would then be to identify the class to which the problem belongs, and then to choose from available methods the best procedure for solving that problem.

The notation and terminology presented in Jacobson (Ref 2) will be used as much as possible in describing networks with simultaneous resource possession. In queueing networks that have simultaneous resource possession a customer will obtain a primary resource from a pool of available

resources, such as a memory partition or disk, and hold it for a preliminary, and possibly zero service time. This service time is called the nonoverlapped service time. Then the customer will proceed to obtain services from other, possibly more than one, resources, the secondary resources or subsystem, while still holding the primary resource. The total service time while the secondary resources are held is called overlapped service time. In the memory problem, the secondary resources are the CPU and disks, and in the I/O problem, the channel. Finally, every instance of simultaneous resource possession has an associated primary resource, secondary resource(s), and nonoverlapped and overlapped service times. If the nonoverlapped service time is always zero (there is no preliminary service requirement prior to the request for services from the secondary resources), then the primary resource is called passive. In the memory problem, the primary resource (memory) is passive because it does not have its own service requirement. However, in the I/O problem the disk has a preliminary service requirement (the seek) before requesting access to the secondary resource and is therefore a nonpassive or active primary resource. The degree of overlap, d, is a ratio which measures the extent to which the customer holds both resources simultaneously. It is given by

(3-1)

$$d = \frac{\text{mean overlapped service time}}{\text{mean overlapped service time + mean nonoverlapped service time}}$$

When d = 1.000, the primary resource is passive and there is complete overlap. When d = 0.0, there is no simultaneous resource possession. A specific type of queueing, called external contention, is present at the primary resource allocation queues whenever a customer must wait for a primary resource even though other resources are available. For example external contention is present at the entry queues of the I/O model because customers must queue for a specific disk.

Simultaneous resource possession can be classified as one of two types which will be called type one and type two. Type one simultaneous resource possession occurs whenever external contention at the primary resource allocation queues is possible. In both types the primary resources may be active or passive.

Subnetworks that have type one simultaneous resource possession can take on many different shapes. One such shape that is of interest in computer applications is illustrated in Figure 3-3. Customers arrive at the subnetwork and enter a common queue that allocates all resources. Once the customer is allocated a resource, he proceeds through the subnetwork, finally returning his resource and making it available for other customers.

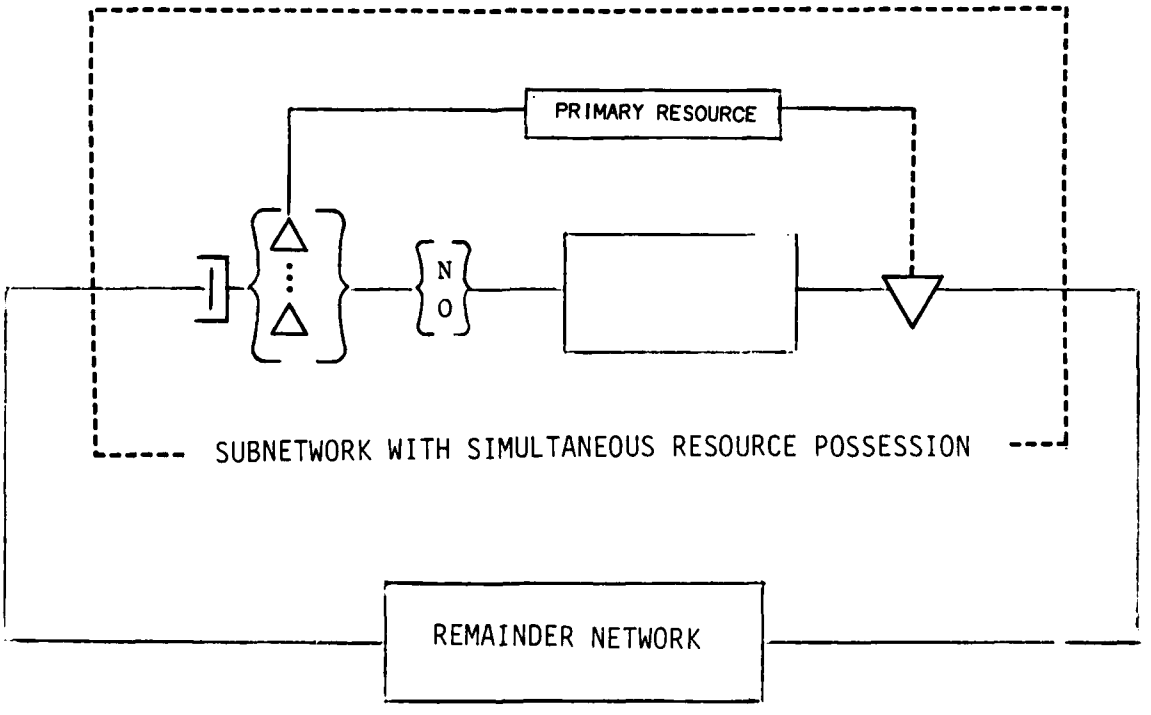


Figure 3-3. Type One Simultaneous Resource Possession

As an example of a type one simultaneous resource possession problem consider the interactive system problem presented by Sauer (Ref 6) and illustrated in Figure 3-4. Interactive terminal users input queries to the system where the queries wait for memory, execute, return memory, and finally respond to the terminal. The terminal queue is an infinite server queue and functions as a delay (i.e., the think time) before the user inputs the next query. The primary resource (i.e. the memory) acts as a population constraint limiting the maximum number of customers in the secondary subsystem to the number of available memory partitions. The impact of the population constraint can considerably change the performance of the system depending on the utilization of the memory partitions.

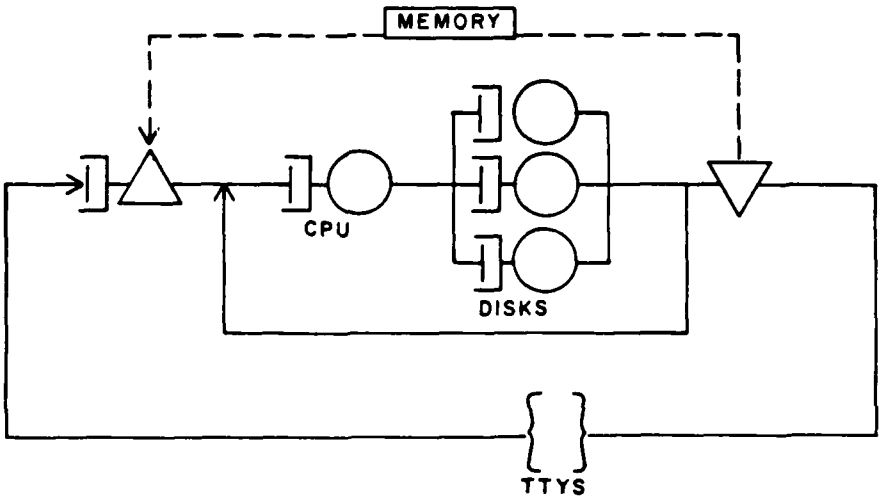


Figure 3-4. Sauer Interactive System Model

Subnetworks that have type two simultaneous resource possession can have many different shapes also. However, subnetworks with type two simultaneous resource possession usually have more than one primary resource

allocation queue. It does not matter how the resources are allocated by the allocation queues as long as external contention is possible. A subnetwork with type two simultaneous resource possession with a shape that is of interest in computer applications is illustrated in Figure 3-5. Customers arrive at the subnnetwork and make a choice as to which resource to use. Once the choice is made, the customer must wait until that resource is available.

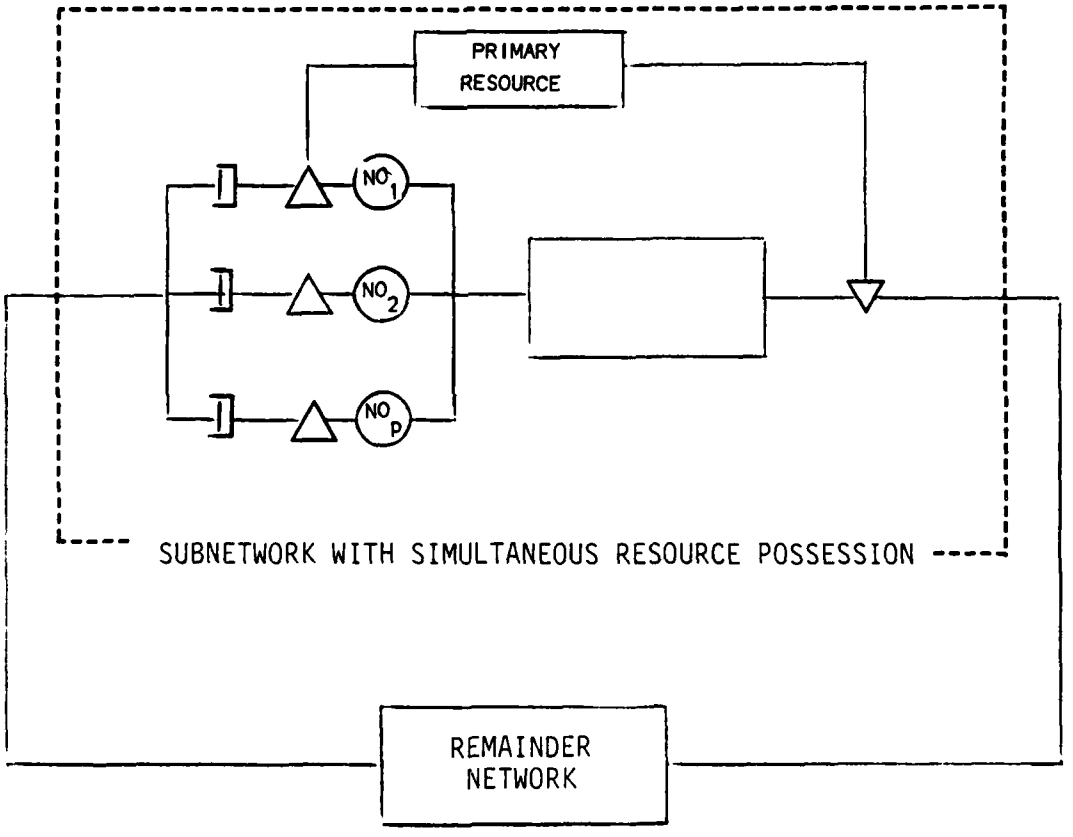


Figure 3-5. Example of Type Two Simultaneous Resource Possession

An example of two simultaneous resource possession with passive primary resources is given by Jacobson (Ref 2, Example 4-1) where a loosely coupled multiprocessor system accesses shared memories. This system is illustrated in Figure 3-6.

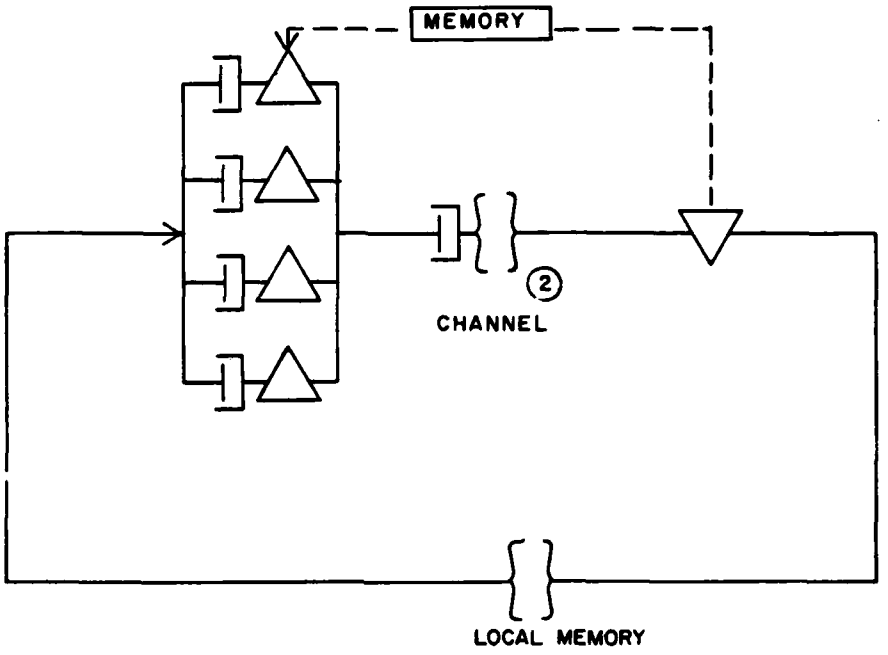


Figure 3-6. Jacobson's Loosely Coupled Multiprocessor System

There are P processors, the customers, that execute the following cycle:

a. Reference their own private local memory for service time lm.

b. Queue for a specific shared memory (one of M).

c. Once shared memory is obtained, queue for any channel (one of B).

d. Once the channel is obtained, use it for service time sm.

The primary resource is passive because there is no service requirement prior to obtaining the secondary resource. External contention is present because the processors must queue for a specific resourse.

The categories of simultaneous resource possession are heirarchical in the sense that a passive type one case is a simplification of active type one, and a type one is a simplification of a type two. Since the only difference between type one and type two is the added element of external contention that is present in type two, that presence must be compensated for in the procedures used. Of the available methods used for type two, all can be simplified to solve type one as a trivial subcase of type two. These relationships between type one and type two and their impacts will become clear through several solved examples presented later in the chapter.

Survey of Available Methods

There are three general approaches that have been used to model simultaneous resource possession problems, namely, Norton's approximation (Ref 6), the method of External Contention Modification (ECM) (Ref 7), and the method of surrogates (Ref 2). The method of surrogates and the ECM method are applicable to both type one and type two problems, Norton's approximation is applicable only to passive type one problems. The ECM method which is a generalization of Norton's approximation, can handle both type one and type two problems and therefore Norton's method will not be discussed separately. Many methods have been developed to model specific problems, such as Bard's interactive virtual memory model (Ref 8), Chen's swapping model (Ref 9), Wilhelm's generalized disk model(Ref 10), and Brown's memory management and response time model (Ref 11) to name a

few. As these methods are not oriented toward general types of simultaneous resource possession problems they will not be discussed further.

The following sections will present a detailed description and procedure of the ECM method and the method of surrogate delays. Examples of their application to type two problems will be illustrated. Both methods restructure the problem so that it can be modeled using product form queueing networks. The accuracy of these two methods and the method presented in Chapter 4, will be discussed in Chapter 5 with the idea of determining what parameters impact the accuracy of the methods.

Method of External Contention Modification (ECM)

The ECM method was first presented by J. N. Bexfield in an unpublished manuscript (Ref 7). It is a generalization of the Norton's approximation approach described in Chandy, Herzog, Woo (Ref 4), and applied in Sauer (Ref 6). The ECM method involves solving two product form models in isolation, a primary resource contention model, and an augmented secondary subsystem model to produce a final variable rate queue. This queue is used in place of the subnetwork with simultaneous resource possession in the original network.

The purpose of the primary resource contention model is to measure the degree of external contention present in the subnetwork. Measures are obtained for each possible network population. The purpose of the augmented secondary subsystem model is to obtain flow equivalent throughputs for each possible customer population in the augmented secondary subsystem while external contention at the primary resources is ignored. The flow equivalents are then adjusted according to the amount of external contention present in order to obtain a final variable rate queue.

Procedure

Consider the part of a closed network which has simultaneous resource possession and call it the subnetwork. Define:

N	Network population
k	Maximum population allowed in the augmented secondary subsystem (also equal to the number of primary resources)
P	Set of primary resource allocation queues
S	Set of nodes in the secondary subsystem
NO_i	Mean nonoverlapped service demand for all primary resources
OL	Mean overlapped service demand for the secondary subsystem
D_i	Mean service demand for queue i
θ_i	Relative throughput of entry queue (or resource) i
θ_{sn}	Relative throughput of the subnetwork
$T_a(n)$	Throughput of the augmented secondary subsystem at population n
$U_p(n)$	Normalized service rate of the primary contention model at population n
$T_p(n)$	Throughput of the primary contention model at population n
$U_f(n)$	Normalized throughput of the final composite queue at population n
SD_f	Mean service demand of the final composite queue

The subnetwork should be analyzed to determine the type of simultaneous resource possession present, what elements comprise the set of primary resources and whether they are active or passive, and what nodes make up the secondary subsystem. The secondary subsystem should be a product form

subnetwork. If it exhibits blocking, variable routing, balking, or any other characteristic that will destroy the product form, it must be analyzed by itself as a system with simultaneous resource possession and replaced with a composite queue prior to continuing. Once this is done the following five step procedure should be used:

1. Calculate the following values:

a. The mean nonoverlapped service demand for all primary resources. This is given by

$$NO = \sum_{i \epsilon P} \frac{NO_i \theta_i}{\theta_{sn}} \qquad (3\text{-}2)$$

θ_{sn} is the relative throughput or loading of the entire subnetwork as if it were a single composite queue. Note that the boundaries of the subnetwork must be chosen such that the relative throughput into the subnetwork is equal to the relative throughput out of the subnetwork. θ_i is the relative throughput through the resource allocation center for primary resource i, and NO_i is its nonoverlapped service requirement which may be zero.

b. The mean overlapped service demand for the secondary subsystem. This is given by

$$OL = \sum_{i \varepsilon S} \frac{D_i \theta_i}{\theta_{sn}} \qquad (3\text{-}3)$$

D_i is the mean service demand and θ_i is the relative throughput at entry queue i.

2. Construct and evaluate the augmented secondary subsystem model. Recall that the purpose of the augmented secondary subsystem is to obtain the throughput capability of the subnetwork while ignoring the presence of contention and congestion at the primary resources. This is done by analyzing the secondary subsystem along with an infinite server queue with service demand NO in isolation. The infinite server queue represents the nonoverlapped service demand. If NO is zero, there is no nonoverlapped service requirement for any primary resource, and the infinite server queue need not be included in the model. The form of the augmented secondary subsystem model is illustrated in Figure 3-7.

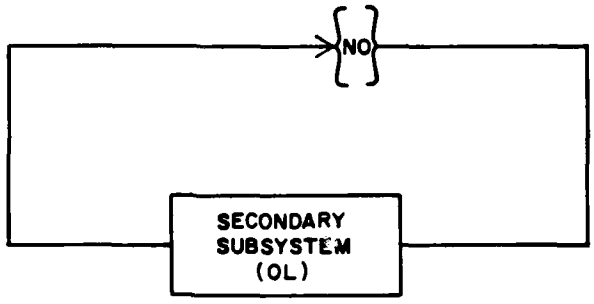


Figure 3-7. Augmented Secondary Subsystem Model

The augmented secondary subsystem model must be evaluated and the throughputs, $T_a(n)$ obtained for each customer population n, for n = 1,...k. Set

$$SD = \frac{1}{T_a(1)} \tag{3-4}$$

SD_f will be the service demand of the final composite node. Note that if the primary resources are passive, this step is equivalent to obtaining the population dependent throughputs of the secondary subsystem as specified by Norton's Theorem for populations 1 to k.

3. Construct and evaluate the primary contention model. The purpose of the primary contention model is to measure the degree of external contention present in the subnetwork. The primary contention model is constructed by replacing each primary resource allocation center i, for all i ε P, with a queue having service demand

$$D_i = NO_i + OL, \tag{3-5}$$

removing the secondary subsystem, and evaluating the resulting network in isolation for customer populations 1 to N. Figure 3-8 illustrates the primary contention model for the simultaneous resource possession problem illustrated in Figure 3-5.

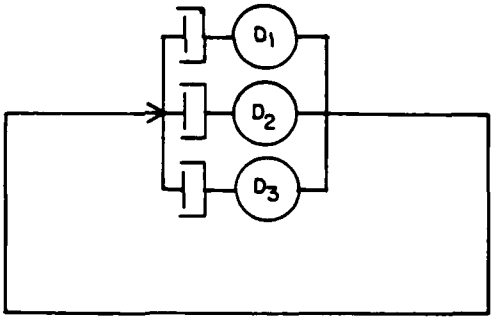


Figure 3-8. Primary Contention Model

Throughputs $T_p(n)$ should be obtained for customer populations 1 to N. The normalized service rates $U_{(n)}$, should be calculated using the formula

$$U_p(n) = \frac{T_p(n)}{T_p(1)}, \quad \text{for each } n,\ n = 1,\ldots,N. \tag{3-6}$$

The normalized service rates are a measure of the external contention present in the subnetwork. They approximate the mean number of customers allowed past the primary resource allocation queues, for network population n, after the effects of external contention are considered. The service rates $U_p(n)$ are a function of the θ_i's and the ratio of the D_i's. For example, if all D_i are equal, the specific values of D_i do not affect the results, as long as all D_i are nonzero and equal.

If the subnetwork exhibits type one simultaneous resource possession, the primary resource contention model will consist of one multiple queue with P servers in isolation. In this case it is simple to show that

$$U_p(n) = \begin{cases} n, & \text{for } n = 1,\ldots,K \\ k, & \text{for } n > K \end{cases} \tag{3-7}$$

and as a result it is not necessary to evaluate the primary contention model for type one problems.

4. Determine the normalized service rates, $U_f(n)$, for n = 1...,N, for the final variable rate queue. Note that the mean service demand, SD_f, was already determined in step 2. The normalized service rates are given by:

$$U_f(n) = \frac{T_a[U_p(n)]}{T_a(1)}, \qquad \text{for } n = 1,\ldots,N \tag{3-8}$$

The $U_f(n)$ are adjusted such that when there are n in the subnetwork, only $U_p(n)$ may enter the augmented secondary subsystem due to the affect of external contention . The values of U_p will not all be integer valued, the throughputs determined from the augmented secondary subsystem model with fractional customer populations must be estimated using interpolation. Simple linear interpolation has been used with good results.

For type one problems equation 3-8 reduces to

$$U_f(n) = \begin{cases} \dfrac{T_a(N)}{T_a(1)}, & n = 1,\ldots,K \\ \dfrac{T_a(L)}{T_a(1)}, & n > K \end{cases} \qquad (3\text{-}9)$$

and when the primary resource is passive this procedure is identical to Norton's approximation.

5. Replace the subnetwork with simultaneous resource possession with a variable rate queue with service demand SD_f and service rate function U_f, and evaluate the resulting network. The performance parameters for the other nodes in the network can be obtained directly from the analysis of the final model. The mean waiting time and mean number of customers at the variable rate queue reflect the total time and the mean number of customers at the subsystem.

Examples

The following examples demonstrate the application of the ECM procedure to type two simultaneous resource possession problems. Examples of applications to type one problems will not be provided as they are simplifications of the type two problem and are well illustrated in the literature (Ref 1, 5, 6, 12).

Example 3-1, I/O Model. Consider a batch processing system with one CPU and one four disk I/O subsystem. All four disks operate through the same controller which has a single channel available for data transfer. All four disks can seek with the use of the channel, but its use is required during the rotational latency and data transfer phases. Suppose that there are ten jobs that alternately execute and perform I/O. All I/O requests have the same probability of accessing any particular disk. The mean CPU service demand per visit is .1, all disks have the same mean seek time of 0.4, and combined rotational latency and transfer time of 0.1. It is desired to know the maximum throughput of the system in terms of CPU-I/O cycles per unit time.

The system is illustrated in Figure 3-9.

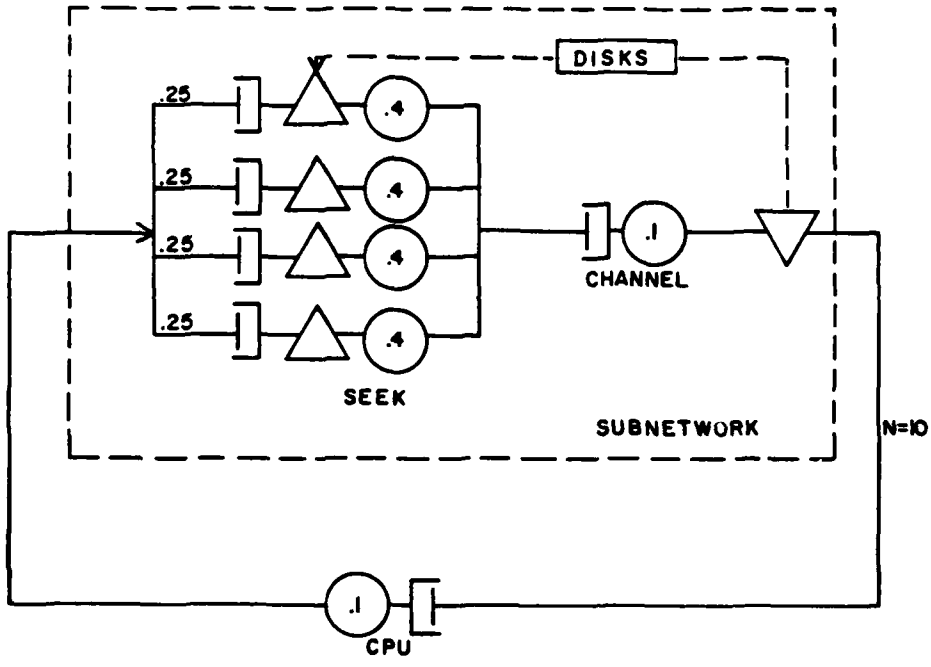


Figure 3-9. CPU - I/O Subsystem Model

This network exhibits simultaneous resource possession of type two. The primary resources are active and consist of the four disks. The secondary resource is the channel. The subnetwork is shown within the dashed box. The procedure follows:

1. The nonoverlapped service time consists of the seek, and makes the primary resource active. The mean nonoverlapped service demand, NO, is given by equation 3-2 and results in the value NO = .4. Similarly, the mean overlapped service demand, OL, is given by equation 3-3, and OL = 0.1. The degree of overlap, d = 0.2.

2. The augmented secondary subsystem model is shown in Figure 3-10, and must be evaluated for populations 1 to 4. The augmented secondary subsystem throughputs, T_a, are given in Table 3-2. The mean service demand of the final composite node is given by equation 3-4, and SD_f = 0.5.

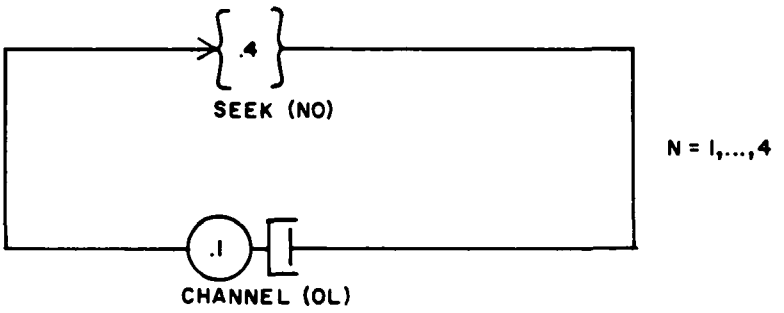


Figure 3-10. Example 3-1 Augmented Secondary Subsystem Model

Table 3-2

Example 3-1 Augmented Secondary Subsystem Throughputs

Population	Throughput [$T_a(n)$]
1	2.000
2	3.846
3	5.493
4	6.893

3. The primary contention model is shown in Figure 3-11 and must be evaluated for customer populations of 1 to 10. The throughputs, T_p, and the normalized service rates, U_p, are given in Table 3-3. The normalized service rates were computed from the throughputs, T_p, using equation 3-6.

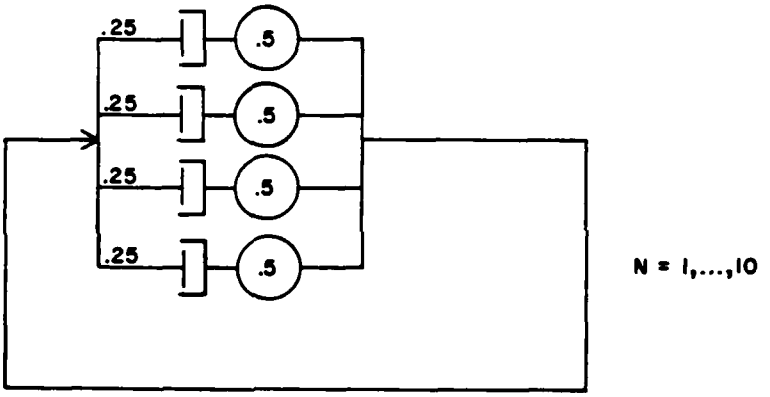


Figure 3-11. Example 3-1 Primary Contention Model

Table 3-3
Example 3-1. Primary Contention Model Throughputs

Number	Throughput [$T_p(n)$]	Normalized Service Rate [$U_p(n)$]
1	2.000	1.000
2	3.200	1.600
3	4.000	2.000
4	4.571	2.286
5	5.000	2.500
6	5.333	2.667
7	5.600	2.800
8	5.818	2.909
9	6.000	3.000
10	6.154	3.077

4. The normalized service rates for the final composite node were obtained using equation 3-8. Interpolation was required to obtain most values and linear interpolation was used in these cases. The final variable rate throughputs are listed in Table 3-4.

Table 3-4

Example 3-1 Final Subnetwork Service Rates

Population	Normalized Service Rate $[U_f(n)]$
1	1.000
2	1.554
3	1.923
4	2.158
5	2.335
6	2.472
7	2.582
8	2.672
9	2.746
10	2.800

5. The final normalized service rates and mean service demand were used in the variable rate node in the network in Figure 3-12. The final results are given in Table 3-5 and will be compared for accuracy in Chapter 5.

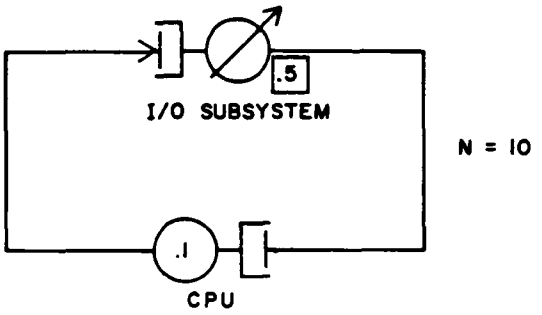


Figure 3-12. Example 3-1 Final Network

Table 3-5

Example 3-1 Performance Parameters

Parameter	Value
Throughput	5.437
Mean I/O residence	1.635
Mean CPU residence	.204
Mean cycle time	1.839

Example 3-2. Loosely coupled multiprocessor system. The following example is a duplication of Example 4-1 in (Ref 2), which illustrates passive type two simultaneous resource possession. Consider a loosely coupled multiprocessor system with eight processors, two shared buses, and four shared memories. In addition, each processor additionally has its own local (private) memory. A processor cycle consists of the following sequence:

1. Reference to local memory for mean time 2.0
2. Queue for specific shared memory (all are equally probable)
3. When memory is obtained queue for any shared bus
4. When bus is obtained access shared memory for mean time 1.0

The system can be modeled as shown in Figure 3-13 where the customers are the eight processors. The desired performance parameters are the throughput in terms of cycles completed per unit time, the mean cycle time, and the local memory and shared memory residency times (including queueing time). The procedure follows:

1. Since the primary resources, the memories, are passive the mean nonoverlapped service demand is zero, or NO = 0.0. The mean overlapped service demand is OL = .1, and the degree of overlap, d = 1.0.

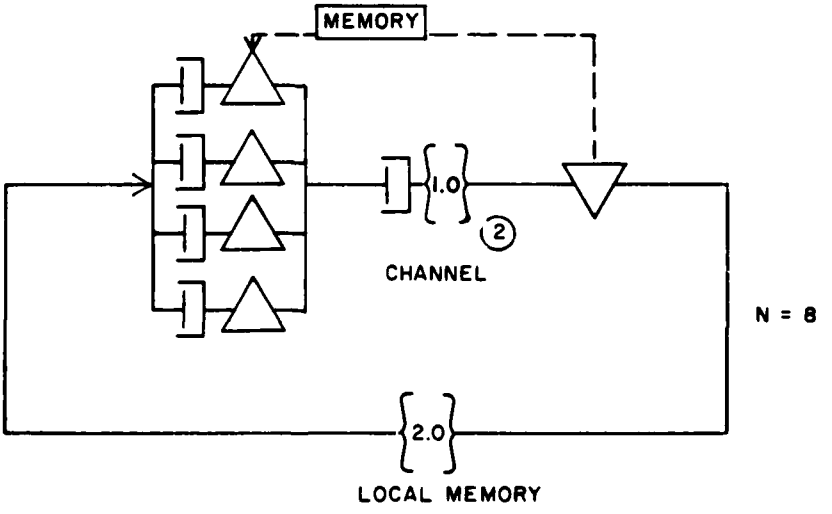


Figure 3-13. Example 3-2 Loosely Coupled Multiprocessor System

2. The augmented secondary subsystem is illustrated in Figure 3-14 and consists of a two server queue which must be evaluated for populations of 1 to 4. In this case it is not necessary to evaluate the model, and the throughputs are given in Table 3-6. $SD_f = 1.0$.

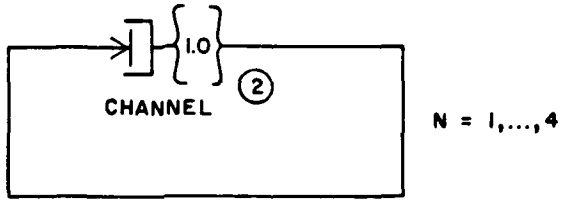


Figure 3-14. Example 3-2 Augmented Secondary Subsystem Model

Table 3-6

Example 3-2 Augmented Secondary Subsystem Throughputs

Population	Throughput $[T_s(n)]$
1	1.0
2	2.0
3	2.0
4	2.0

3. The primary contention model is identical to the one illustrated in Figure 3-11. All the service demands in the primary contention model, D_i, are identical as all $NO_i = 0.0$, for each $i \in P$, and as a result the normalized service rates (not the throughputs) are independent of the service demands D_i. Hence, the normalized service rates already computed in Table 3-3, for populations 1 to 8 may be used directly.

4. The final normalized service rates, U_f, are listed in Table 3-7.

Table 3-7

Example 3-2 Final Subnetwork Service Rates

Population (n)	Normalized Service Rate $U_f(n)$
1	1.000
2	1.600
3	2.000
4	2.000
5	2.000
6	2.000
7	2.000
8	2.000

5. The final model is illustrated in Figure 3-15 and was evaluated with a customer population of 8. The results are given in Table 3-8.

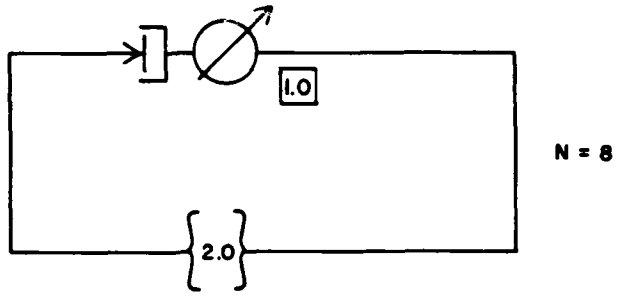


Figure 3-15. Example 3-2 Final Network

Table 3-8

Example 3-2 Performance Parameters

Parameter	Value
Throughput	1.881
Mean shared mem & bus R	2.254
Mean lcl mem R	2.000
Mean cycle time	4.254

Method of Surrogates

The method of surrogates was first presented in 1981 (Ref 13) and later described in (Ref 2). It is an iterative procedure that solves both type one and type two simultaneous resource possession problems, but was developed primarily for type two problems. It involves evaluating an augmented

secondary subsystem model, identical to the one used in the ECM procedure, and iterating between two other models, a primary contention model and a secondary contention model. Performance parameters are obtained from the primary contention model and the secondary contention model after convergence.

The purpose of the augmented secondary subsystem model is the same as in the ECM procedure: to obtain flow equivalent throughputs of the subnetwork in the absence of external contention and congestion at the primary resources. The primary contention model contains an explicit representation of the primary resources and is used to estimate the external contention in terms of queueing delay. A delay server is used to represent the queueing delay due to congestion and contention in the augmented secondary subsystem. The secondary contention model contains an explicit representation of the augmented secondary subsystem and is used to estimate the contention and congestion present in the secondary subsystem model. It contains a delay server to represent the queueing delay due to external contention. Hence, each model estimates a component of the total queueing delay.

The models are evaluated alternately where the component of queueing delay estimated in one model is inserted into the delay server of the other model allowing a more accurate estimate of the original component. Iterations between the two models are performed until convergence is achieved. Convergence, however, is not guaranteed.

Procedure

Consider the part of a closed queueing network which has simultaneous resource possession and call it the subnetwork. Call the other part the 'remainder' network. Define:

N	Network population
K	Maximum population allowed in the secondary subsystem and equal to the number of primary resources
P	Set of primary resource allocation queues
S	Set of nodes in the secondary subsystem
NO_i	Mean nonoverlapped service demand of primary resource i, for each ieP
NO	Mean nonoverlapped service demand for all primary resources
OL	Mean overlapped service demand for the secondary subsystem
D_i	Mean service demand for node i
θ_i	Relative throughput of node (or resource) i
θ_{sp}	Relative throughput of the subnetwork
$T_a^{sp}(n)$	Throughput of the augmented secondary subsystem at population n

The following parameters relate to the primary contention model:

R_{pi}^n	n^{th} iterate of the residency time at node i of the primary nodes
R_p^n	n^{th} iterate of the mean residency time for all primary nodes
D_{pd}^n	n^{th} iterate value of the mean service demand for the primary delay server
T_p^n	n^{th} iterate value of the throughput through the subnetwork in the primary contention model

The following parameters relate to the secondary contention model:

R_s^n	n^{th} iterate of the mean residency time at the variable rate node
D_{sd}^n	n^{th} iterate of the mean service demand of the delay server
T_s^n	n^{th} iterate of the throughput through the subnetwork as modeled in the secondary contention model
D_{sv}	Mean service demand of the variable rate node
$U_{sv}(n)$	Service rate of the variable rate node at population n

The subnetwork should be analyzed to determine the type of simultaneous resource possession present, what elements comprise the set of primary resources and whether they are active or passive, and what nodes make up the secondary subsystem. The following procedure will first describe how the models are constructed (setup procedure) and analyzed prior to the iterative cycle (iterative procedure). Then the iterative procedure will be described.

Setup Procedure

1. Calculate the following values:

a. The mean nonoverlapped service demand for all primary resources, NO, using equation 3-2.

b. The mean overlapped service demand, OL, for the secondary subsystem using equation 3-3.

2. Construct and evaluate the augmented secondary subsystem model. This model is identical to the one constructed in step 2 of the ECM procedure. The throughputs, $T_a(n)$, for n = 1,...,K should be obtained.

3. Create the primary contention model which will be the first model evaluated in the iterative procedure. The purpose of the primary contention model is to estimate the amount of external contention present at the primary resources in terms of queueing delay (called external queueing delay). The primary contention model consists of the remainder network, a set of nodes, called the primary nodes, (that explicitly model the external contention present at the primary resources), and an infinite server node, called the delay server. The explicit representation of the external contention at the primary resources is constructed by replacing each

primary resource allocation center i, for all iEP, with a queue having service demand, D_i, given by equation 3-11, and relative throughput θ_i. The presence of the secondary subsystem is ignored. The form of the model is illustrated in Figure 3-16.

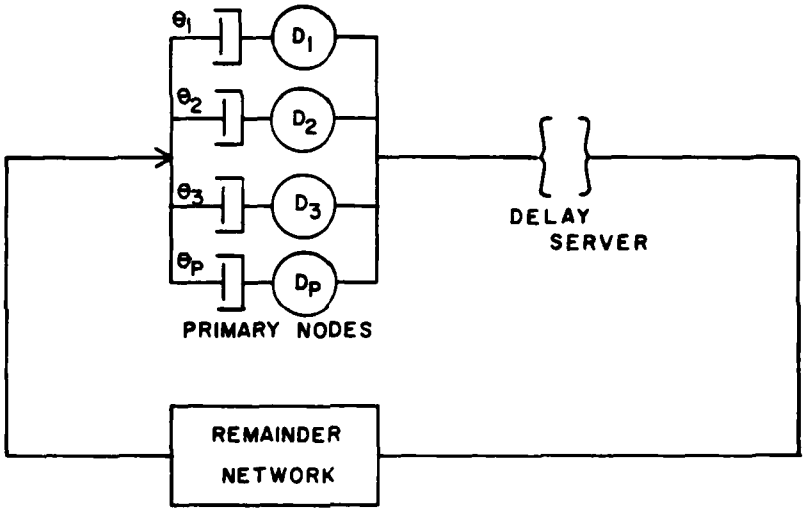


Figure 3-16. Method of Surrogates Primary Contention Model

The primary nodes will usually consist of a set of queues in parallel. The mean service demands and rate functions for all nodes in the remainder network and in the primary nodes will remain constant for each iteration of the model. However, the mean service demand for the delay server, D^n_{pd}, will change at each iteration. The value of the service demand at the n^{th} iteration, D^n_{pd}, will be determined from parameters obtained in the $n-1^{th}$ iteration of the secondary contention model. The primary contention model is evaluated with a network population of N.

4. Create the secondary contention model which will be the second model evaluated in the iterative procedure. The purpose of the secondary contention model is to obtain more accurate estimates of the congestion and contention present in the augmented secondary subsystem. The secondary contention model consists of the remainder network, a variable rate node, SV, that explicitly models the augmented secondary subsystem, and an infinite server queue, called the delay server. The mean service demand, D_{sv}, and rate function, U_{sv}, for the variable rate queue are obtained from the augmented secondary subsystem model and are given by:

$$D_{sv} = \frac{1}{T_a(1)} \tag{3-10}$$

$$U_{sv}(n) = \begin{cases} \dfrac{T_a(n)}{T_a(1)}, & \text{for } n = 1,\ldots,K \\ \dfrac{T_a(K)}{T_a(1)}, & \text{for } n \rangle K \end{cases} \tag{3-11}$$

The above values are fixed for all iterations of the model. The mean service demand of the delay server, D^n_{sd}, varies for each iterate. The value of the n^{th} iterate of the service demand, D^n_{sd}, is obtained from the outputs of the n^{th} iteration of the primary contention model. The second-

ary contention model is always evaluated with a network population of N. The general format of the secondary contention model is illustrated in Figure 3-17.

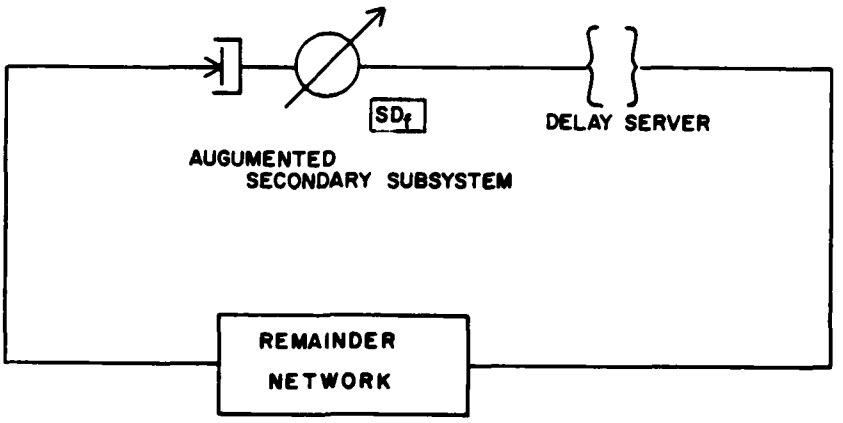


Figure 3-17. Method of Surrogates Secondary Contention Model

Iterative Procedure

The iterative procedure consists of alternate evaluation of first the primary contention model and then the secondary contention model. Evaluation of both models completes one iteration. The following procedures should be used.

1. Set the service demand of the delay server of the primary contention model, D^1_{pd}, to zero for the first iteration. Set N = 1.

2. Evaluate the primary contention model for iteration n using the mean service demand D^n_{pd} obtained from step 1, or 3.c:

a. Record the throughput, T_p^n, through the subnetwork of the primary contention model.

b. Record the residency time, R_{pi}^n for each primary resource i, for all $i \epsilon P$.

c. Calculate the mean residency time, R_p^n, for all primary resources using the following formula:

$$R_p^n = \sum_{i \epsilon P} \frac{R_{pi}^n \theta_i}{\theta_{sn}} \qquad (3\text{-}12)$$

If all R_{pi} are equal, then equivalently

$$R_p^n = R_{pi}^n, \text{ for any } i \epsilon P \qquad (3\text{-}13)$$

d. Calculate the mean service demand of the delay server, D_{sd}^n, for the secondary contention model using the formula:

$$D_{sd}^n = R_p^n (NO + OL) \qquad (3\text{-}14)$$

3. Evaluate the n^{th} iteration of the secondary contention model using the mean service demand D_{sd}^n for the delay server computed in step 2.d above:

a. Record the throughput, T_s^n, through the subnetwork of the secondary contention model.

b. If $T_s^n > T_p^n$, then the iterations will not converge, so stop.

c. Record the mean residency time of the variable rate node, R_s^n, and set the mean service demand of the delay server, D_{pd}^{n+1}, for the $n+1^{st}$ iterate of the primary contention model using the following formula:

$$D_{pd}^{n+1} = R_s^n - (NO + OL). \tag{3-15}$$

4. Set n = n + 1 and go to step 2. Repeat these two steps until:

a. The throughputs T_p^n and T_s^n are equal to within the desired accuracy or

b. The sequences $[R_p^n]$ and $[T_s^n]$ converge to within the desired accuracy.

Convergence

The convergence of the method of surrogates hinges on the premise that the primary contention model evaluated without the effects of contention or congestion in the augmented secondary subsystem will always have a higher throughput than the secondary contention model evaluated with the first estimate of the queueing delay for external contention. Hence, it is assumed that T_p^1 will always be greater than or equal to T_s^1. The successive estimates of the queueing delay caused by external contention would then decrease as the estimates of the queueing delay caused by contention and congestion in the augmented secondary subsystem would increase. As the

increasing queueing delay estimates would be used in the primary contention model, the sequence of primary contention model throughputs, $[T_p^n]$, would monotonically decrease. Similarly, decreasing queueing delay estimates would be used in the secondary contention model and therefore its sequence of throughputs, $[T_s^n]$, would increase monotonically. As a result of these properties of the sequence $[T_p^n]$ and $[T_s^n]$, the algorithm will diverge if- $T_p^1 < T_s^1$.

Another possibility is that the sequences are not bounded, and there will exist an N, such that for all $n > N$, $T_p^n < T_s^n$ is met.

The divergence of the method of surrogates occurs under certain conditions when the congestion and contention present in the augmented secondary subsystem is small. Examples will be given of both types of divergence as well as an analysis of the errors involved in Chapter 5.

Example 3-3

Only one example will be provided of the method of surrogates. Others exist (Ref 2). Since the method was primarily intended for type two simultaneous resource possession problems a type two example will be given.

Consider the I/O model given in Example 3-1 of the ECM procedure and illustrated in Figure 3-9, except let NO_i=.3 for all disks, and OL=.2. The primary resources are active and consist of the four disks. The secondary resource consists of the channel. The procedure follows:

Setup

1. The mean nonoverlapped service demand. NO, is given by equation 3-2, and here NO = .3. Similiarily, OL = .2.

2. The augmented secondary subsystem model is also identical to that in Example 3-1 where the load dependent throughputs are given in Table 3-2.

3. The primary contention model consists of the remainder network (in this case the queue representing the CPU), four primary nodes explicitly modeling the primary resources, and the delay server representing the queueing delay due to contention and congestion in the augmented secondary subsystem. The primary contention model is illustrated in Figure 3-18.

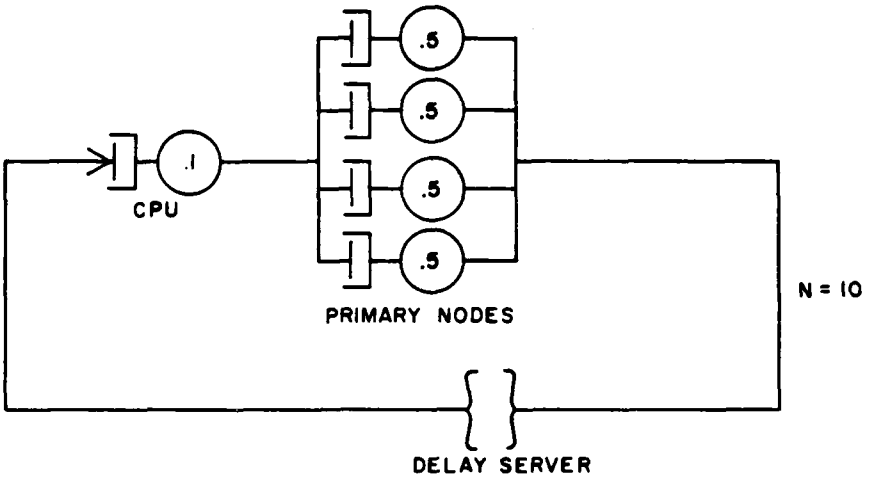


Figure 3-18. Example 3-3 Method of Surrogates Primary Contention Model

The service rates of the primary nodes, D_i, are given by equation 3-5. In this case $D_i = .5$, for all $i \in P$. The model will be evaluated with ten customers during the iterative procedure.

4. The secondary contention model consists of the CPU queue, the delay server which is an infinite server node, and a variable rate queue representing the augmented secondary subsystem. The model is illustrated in Figure 3-19. The mean service demand and service rate function are given by equations 3-10 and 3-11, respectively. The mean service demand, $D_{sv} = .5$, and the service rates are given in Table 3-9. The model will be evaluated with ten customers during each iteration. The service demand of the delay server will be determined by the primary contention model during the iterative procedure.

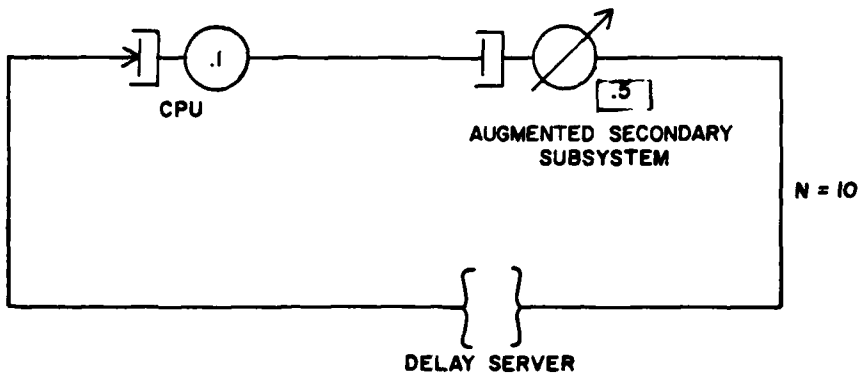


Figure 3-19. Example 3-3 Method of Surrogates Secondary Contention Model

Table 3-9

Example 3-3 Secondary Contention Model Throughputs

Population (n)	Service rate [$U_{sv}(n)$]
1	1.000
2	1.724
3	2.164
4	2.380
5	2.380
6	2.380
7	2.380
8	2.380
9	2.380
10	2.380

Iterative Procedure

The iterative procedure begins by evaluating the first iterate of the primary contention model with an input mean service demand for the delay server set to zero. Table 3-10 displays the iterative solution.

TABLE 3-10

Iterative Solution of Example 3-1

Primary Contention Model				Secondary Contention Model		
Iteration	Input D^n_{pd}	Throughput T^n_p	Mean R^n_p	Input D^n_{sd}	Throughput T^n_s	Mean R^n_s
1	0	5.900	1.4752	.9572	4.445	1.1055
2	.6055	5.159	1.1468	.6468	4.651	1.3245
3	.8245	4.855	1.0590	.5590	4.684	1.3947
4	.8947	4.757	1.0342	.5342	4.692	1.4153
5	.9153	4.728	1.0271	.5271	4.694	1.4212
6	.9212	4.720	1.0251	.5251	4.695	1.4229
7	.9229	4.718	1.0246	.5246	---	---

Each iteration consists of first evaluating the primary contention model, and then evaluating the secondary contention model. One row of the table corresponds to one iteration, hence seven and a half iterations are illustrated. At the stopping point the iterates had not completely converged. The queueing delay D_{sd}^{n} for the n^{th} iterate of the secondary contention model is determined from the n^{th} iterate of the primary contention model where

$$D_{sd}^{n} = R_{p}^{n} - (NO+OL) \qquad (3\text{-}14)$$

repeated

As an example for the second iterate:

$$D_{sd}^{2} = R_{p}^{2} - (NO+OL)$$

$$= 1.1468 - (.3+.2)$$

$$= .6468$$

The queueing delay, D_{pd}^{n+1}, for the $n+1^{st}$ iterate of the primary contention model is determined from the n^{th} iterate of the secondary contention model (or if n=0, then $D_{pd}^{1} = 0$) where

$$D_{pd}^{n+1} = R_{s}^{n} - (NO+OL) \qquad (3\text{-}15)$$

repeated

For the third iterate:

$$D_{pd}^{3} = R_{s}^{2} - (NO+OL)$$

$$= 1.3245 - (.3+2)$$

$$= .8245$$

If this model was evaluated with the nonoverlapped and overlapped service times as given in Example 3.1 (NO = .4,OL=.1) then type two divergence results. The sequence of iterations are listed in Table 3-11.

The procedure was stopped after the third iteration because T^3_{sd}. If the procedure is continued the throughputs will diverge forever.

TABLE 3-11

Example 3-1. Method of Surrogates

Primary Contention Model				Secondary Contention Model		
Iteration	Input D^n_{pd}	Throughput T^n_p	Mean R^n_p	Input D^n_s	Throughput T^n_{sd}	Mean R^n_s
1	0	5.900	1.4752	.9752	5.456	.6690
2	.1690	5.714	1.3712	.8712	5.675	.6945
3	.1945	5.684	1.3563	.8563	5.706	

Remarks

The examples given in the previous pages were all networks containing two simultaneous resource possession. No examples were given to illustrate the procedure for networks with type one simultaneous resource possession because the analysis of such networks using Norton's approximation has been well documented in the literature (Ref 1, 4, 5 6. 12). The two procedures presented were developed primarily to solve type two simultaneous resource possession problems to which Norton's approximation is not applicable.

The problem with the currently available methods for solving networks with type two simultaneous resource possession is two fold. First, the error present in the analytic results can be substantial and unpredictable,

as the causes of error are not well known. Second, it is difficult to obtain estimates of parameters for queues within the subnetwork with simultaneous resource possession. Yet these are often the parameters for which estimates are desired.

In the following chapter, a new procedure will be presented for solving a subset of networks with type two simultaneous resource possession. The subset includes all networks that allocate one primary resource at each entry queue, and where the nonoverlapped service requirement, if any, is equal for all primary resources. The procedure allows easy estimation of any queue parameter within the subnetwork with simultaneous resource possession. In Chapter 5 measures are identified that impact the possibility for error and its magnitude.

IV. New Procedure

Introduction

In the previous chapter, two methods were presented that can approximately solve type two simultaneous resource possession problems. A step by step procedure was presented for each method along with examples illustrating their use. In this chapter a new multi-entrance queue will be defined and a procedure called the multi-entrance queue (MEQ) procedure will be presented that has advantages over both of the previous methods in solving type two problems. It is not applicable to type one problems. These advantages are due to its simplistic and direct solution procedure along with allowing estimation of the performance parameters within the subnetwork.

Similar to the methods in Chapter 3, this new procedure utilizes the throughputs obtained from the augmented secondary subsystem model. These throughputs are used as the service rates in a closed multi-entrance queue model that more accurately estimates the effects of external contention. The output of the model is a mean service demand and set of service rates which are used in a variable rate queue that replaces the subnetwork with simultaneous resource possession in the original network. The closed multi-entrance queue is solved by calculating the probability distribution and then the desired throughputs. The probabilities are calculated using a product form solution that differs from the "product form solution currently known". The development and derivation of the product form solution are given in this chapter. Algorithms that solve the model are also presented later in the chapter.

An advantage of this procedure is that it allows approximation of the performance parameters of the queues within the original subnetwork with simultaneous resource possession. This is not possible with the method of surrogates and difficult at best in the method of ECM.

In the following sections of this chapter a description of the heuristics underlying the procedure is given, a derivation of the solution to the closed multi-entrance queue is provided, and finally, a step by step procedure describing the method with examples is given.

Model Development

In all type two simultaneous resource possession problems external contention is present customers are required to wait for a specific resource even though others are available). It is this type of queueing delay that has been difficult to characterize previously, and in which this method focuses. A special closed multi-entrance queue is evaluated in isolation for all possible customer populations in the network in order to obtain the estimates of queueing delay in terms of throughputs. This is the similar to the idea that is used in the ECM procedure, but the multi-entrance queue procedure uses a different method of determining the throughputs. The following paragraphs will address the model development from the original network with simultaneous resource possession.

Consider the model in Figure 4-1 which illustrates a general active type two simultaneous resource possession problem. Suppose that each primary resource entry queue only allocates one primary resource, that all other customers in the queue are blocked until that customer completes service, and that the mean nonoverlapped service demands for all resources are equal. Under these conditions the model in Figure 4-1 can be replaced by the model in Figure 4-2.

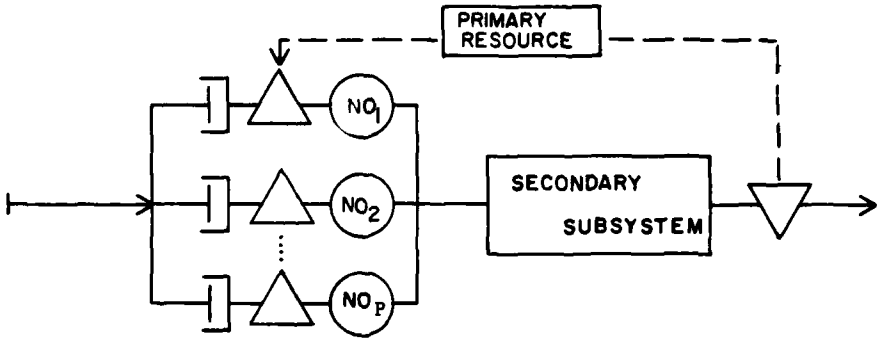


Figure 4-1. Representation of the Subnetwork With Simultaneous Resource Possession

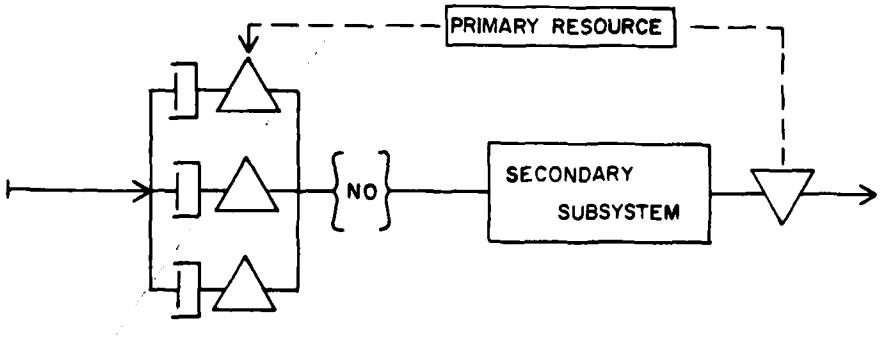


Figure 4-2. Subnetwork With a Common Node for the Nonoverlapped Service Time

In the presence of blocking at the primary resource, it does not matter which resource is allocated as long as the blocking at the primary resource allocation center where the customer arrived is maintained until the customer being serviced departs. The network in Figure 4-2 is not equivalent to the one in Figure 4-1 if the nonoverlapped service times are not all equal. The error introduced by this departure is a subject of further investigation. If the nonoverlapped service queue and the secondary subsystem are analyzed in isolation, as illustrated in Figure 3-7, and are replaced by a variable rate server, the network in Figure 4-3 is obtained. Hence, the purpose of the augmented secondary subsystem model is to transform the network of Figure 4-1 into the network of Figure 4-3. However, the network in Figure 4-1 is not equivalent to the network in

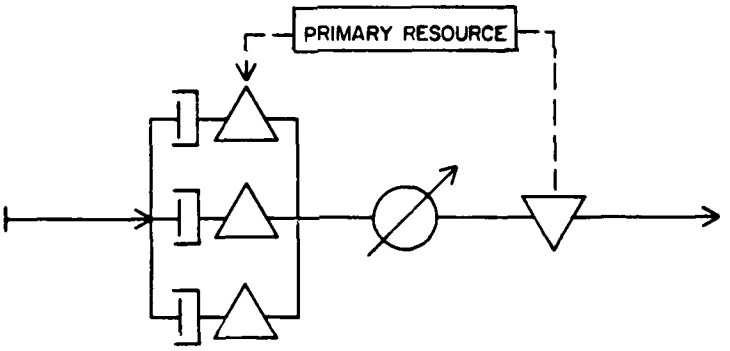


Figure 4-3. Multi-entrance Queue Representation of the Subnetwork. The nonoverlapped queue and Secondary Subsystem are Replaced With a Variable Rate Queue

Figure 4-3, even when all nonoverlapped service times are equal. Error is introduced by the transformation. This error is important because it will be present in all methods that use the augmented secondary subsystem model. The error introduced by the augmented secondary subsystem model is a topic of discussion in Chapter 5.

The next step is to reduce the network of Figure 4-3 to a single variable rate queue that can be inserted into the original network. The method of ECM and the method of surrogates each do this in different ways. The method in the multi-entrance queue procedure is to solve the network of Figure 4-3 in isolation exactly (by making certain assumptions), and as a result obtain flow equivalent throughputs for each possible network population. These throughputs are then used in a variable rate queue replacing the subnetwork.

The multi-entrance queue is illustrated in Figure 4-4. From its solution, the appropriate population dependent throughputs can be obtained. Its queues are serviced in FCFS order and the customers in service subdivide the service capacity using a processor share discipline. The service capacity depends on the number of nonempty or busy queues.

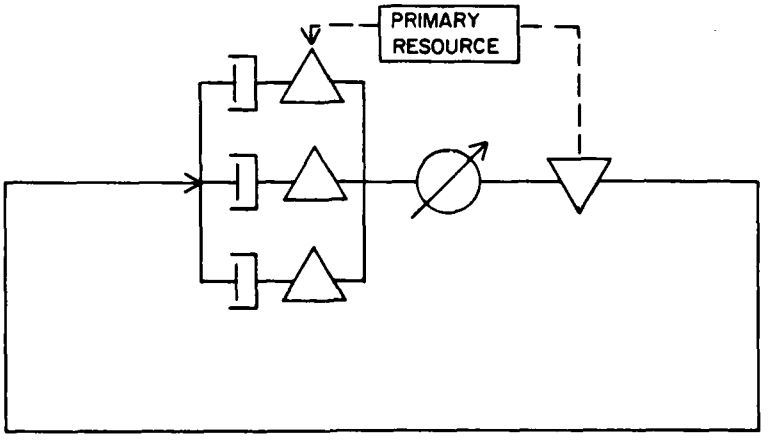


Figure 4-4. Multi-entrance Queue Evaluated in Isolation

Figure 4-4.

There are two aspects to the service discipline of the multi-entrance queue. First, customers are allowed entry into the subnetwork from the each primary resource entry queue using first come first served (FCFS) order. Each entry queue that has customers in line must have one, and only one, customer in service. Since there may not be sufficient service capacity to service all customers without delay, the available service capacity is divided among the customers using a processor sharing service discipline. However, in most systems of interest, the service capacity is divided among the customers using a FCFS discipline. The processor sharing service discipline subdivides the service capacity in the same proportions as the FCFS discipline, but the two disciplines are not equivalent, and the model is difficult to solve if the FCFS discipline is used. The sources of error caused by this assumption are investigated in Chapter 5.

The multi-entrance queue has service rates that depend on the number of nonempty queues because each entry queue can have at most one customer in service. This is different from the population dependent service rates of conventional product form queues. Note that the maximum number of customers that can be in service is the smaller of:

1. The number of queues, or
2. The number of customers in the system.

The throughputs of the augmented secondary subsystem model, T_a, are used as the service rates for the multi-entrance queue. Recall that the throughputs obtained from the augmented secondary subsystem model for each feasible subnetwork population reflect the throughput capability of the subnetwork when that number of customers are in service. Let:

k = number of resources and hence number of queues, and

$U_m(n)$ = exponential service rate with n busy queues for n=1,...,k.

Then

$$U_m(n) = T_a(n) \quad \text{for } n = 1,....,k, \tag{4-1}$$

where

$T_a(n)$ is the throughput through the augmented secondary subsystem model evaluated with n customers.

Note that this service rate function depends on the number of nonempty queues, or equivalently the number of customers in service, in the multi-entrance queue. The object of evaluating the multi-entrance queue is to determine service rates that are a direct function of the number in the subnetwork, and hence can be used in a variable rate queue.

The next step is to evaluate the multi-entrance queue with an arbitrary number of entry queues and a general service rate function that depends on the number of nonempty queues. Solving balance equations is clearly not desirable because of the number and necessity of enumerating each equation. Fortunately, the multi-entrance queue, when evaluated in isolation, has a simple product form solution that will enable easy calculation of the desired performance metrics. The following section will present the product form and its proof.

Product form solution

Consider a closed multi-entrance queue evaluated in isolation where customers are served from each queue in FCFS order. All customers in service must processor share an exponential server whose capacity is a function of the number in service. Suppose that only one customer from each queue can be in service at any one time, even if other queues are empty. Let:

N = number of customers in the multi-entrance queue evaluated in isolation.

k = number of queues.

S be the set of all states, s, of the form $(n_1, \ldots, n_k)$, where

n_i = number of customers in queue, i for i=1,..., N.

There are $\binom{N + k - 1}{k - 1}$ states.

q = number of nonempty queues for state s

θ_i = relative throughput at queue i, for i=1,...,k.

P_i = probability that a customer will proceed to queue i after completing service.

The relative throughputs are only solvable to within a multiplicative constant and are given by

$$\theta_i = \sum_{j=1}^{k} \theta_j P_i \qquad (4\text{-}2)$$

Let

$U_m(q)$ = unnormalized exponential service rate with q nonempty queues for q=1,...,k. This service rate function depends on the number of nonempty or busy queues.

Theorem. Given a multi-entrance queue evaluated in isolation where the entry queues are serviced in FCFS order and the available service capacity of an exponential server is subdivided using a processor share service discipline, the state probabilities are given by

$$P(n_1,\ldots,n_k) = q \frac{\theta_1^{n_1}\ldots\theta_k^{n_k}}{U_m(q)\ G} \qquad \text{for each state } s \in S, \qquad (4\text{-}3)$$

where

G is a normalizing constant ensuring conservation of probability.

Proof. The theorem will be proved by showing that the product form solution satisfies all global balance equations for the queue of which there are

$$\binom{N + k - 1}{k - 1}.$$

The general form of the balance equation for each state $s \in S$, is given by

$$P(n_1,\ldots,n_k)\ U_m(q) = \sum_{\substack{i=1 \\ n_i \neq 0}}^{k} \frac{P(n_1,\ldots,n_k)\ P_i\ U_m(q)}{q}$$

$$+ \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i \\ n_j \neq 0}}^{k} \frac{P(n_1,\ldots,n_i+1,\ldots,n_j-1,\ldots,n_k)\ P_j\ U_m(q^*)}{q^*} \qquad (4\text{-}4)$$

where

q* corresponds to the number of busy queues for state $(n_1,\ldots,n_i+1,\ldots,n_j-1,\ldots,n_k)$.

The term on the left equates the rate of flow out of state s, and, the terms on the right equate the rate of flow into state s. Term (a) represents transitions where a customer will re-enter the same queue he was previously situated, and term (b) represents transitions where a customer enters a different queue. The proof proceeds by substituting the product form solution into the general balance equation 4-4, and showing that all equations are satisfied.

Substituting the product form into 4-4 yields:

$$\left[\frac{q\ \theta_1^{n_1}\ldots\theta_k^{n_k}}{G\ U_m(q)}\right]\frac{U_m(q)}{} = \sum_{\substack{i=1\\ n_i\neq 0}}^{k}\left[\frac{q\ \theta_1^{n_1}\ldots\theta_k^{n_k}}{U_m(q)\ G}\right]\frac{P_i\ U_m(q)}{q} \qquad \text{(a)}$$

(4-5)

$$+\sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq i\\ n_j\neq 0}}^{k}\frac{\theta_i}{\theta_j}\left[\frac{q^*\ \theta_1^{n_1}\ldots\theta_k^{n_k}}{U_m(q^*)\ G}\right]\frac{P_j\ U_m(q^*)}{q^*} \qquad \text{(b)}$$

where

q^* and $U(q^*)$ reflect the number of nonempty queues and service rate for state $(n_i,\ldots,n_i+1,\ldots,n_j-1,\ldots,n_k)$.

Let $\theta = \theta_1^{n_i}\ldots\theta_k^{n_k}$, multiply equation 4-5 by G, and simplify yielding

$$q\ \theta = \sum_{\substack{i=1\\ n_i\neq 0}}^{k}\theta\ p_i + \sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq i\\ n_j\neq 0}}^{k}\frac{\theta_i}{\theta_j}\ \theta\ p_j \qquad (4\text{-}6)$$

Divide equation 4-5 by θ giving

$$q = \sum_{\substack{i=1\\ n_i\neq 0}}^{k} P_i + \sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq i\\ n_j\neq 0}}^{k}\frac{\theta_i}{\theta_j}\ P_j \qquad (4\text{-}7)$$

Recall that $P_j = \dfrac{\theta_j}{\theta_1+\ldots+\theta_k}$ for any j, j=1,...,K

Hence

$$q = \sum_{\substack{i=1 \\ n_i \neq 0}}^{k} P_i + \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i \\ n_j \neq 0}}^{k} \frac{\theta_i}{\theta_j} \frac{\theta_j}{(\theta_1 + \ldots + \theta_k)} \tag{4-8}$$

Collect θ_i in the numerator and $\theta_1+\ldots+\theta_k$ in the denominator, replacing it by P_i, and yielding

$$q = \sum_{\substack{i=1 \\ n_i \neq 0}}^{k} P_i + \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i \\ n_j \neq 0}}^{k} P_i \tag{4-9}$$

Let

$$M_i = \sum_{\substack{j=1 \\ j \neq i \\ n_j \neq 0}}^{k} P_i \tag{4-10}$$

Then

$$q = \sum_{\substack{i=1 \\ n_i = 0}}^{k} P_i + \sum_{i=1}^{k} M_i \tag{4-11}$$

Subdivide the last summation in 4-11 into two summations: 1) those where $n_i \neq 0$, and 2) those where $n_i = 0$.

$$q = \sum_{\substack{i=1 \\ n_i \neq 0}}^{k} P_i + \sum_{\substack{i=1 \\ n_i \neq 0}}^{k} M_i + \sum_{\substack{i=1 \\ n_i = 0}}^{k} M_i \qquad (4\text{-}12)$$

Note that

$$M_i = \begin{cases} q\, P_i & \text{if } n_i = 0 \\ (q-1)\, P_i & \text{if } n_i \neq 0 \end{cases} \qquad (4\text{-}13)$$

Hence

$$q = \sum_{\substack{i=1 \\ n_i \neq 0}}^{k} P_i + (q-1) \sum_{\substack{i=1 \\ n_i \neq 0}}^{k} P_i + q \sum_{\substack{i=1 \\ n_i = 0}}^{k} P_i \qquad (4\text{-}14)$$

Separate the last summation of 4-14 into a (q-1) term and a (1) term

$$q = \sum_{\substack{i=1 \\ n_i \neq 0}}^{k} P_i + (q-1) \sum_{\substack{i=1 \\ n_i \neq 0}}^{k} P_i + (q-1) \sum_{\substack{i=1 \\ n_i = 0}}^{k} P_i + \sum_{\substack{i=1 \\ n_i = 0}}^{k} P_i \qquad (4\text{-}15)$$

Rearranging terms

$$q = \sum_{\substack{i=1 \\ n_i \neq 0}}^{k} P_i + \sum_{\substack{i=1 \\ n_i = 0}}^{k} P_i \qquad (4\text{-}18)$$

$$+ (q-1) \sum_{\substack{i=1 \\ n_i \neq 0}}^{k} P_i + (q-1) \sum_{\substack{i=1 \\ n_i = 0}}^{k} P_i$$

$$q = 1 + q - 1$$

$$q = q$$

Q E D

Hence all equations are satisfied which shows that the product form expressions for the probabilities are a solution to the balance equations.

Calculating Performance Parameters

By utilizing the product form solution, the probabilities are easily calculated. Two algorithms will be provided to show this. Algorithm 4-1 calculates only the population dependent throughputs required for the final variable rate queue. None of the performance parameters for the multi-entrance queue are determined. Algorithm 4-2 calculates the throughputs and

the multi-entrance queue performance parameters, and this algorithm must be used if estimates of performance parameters within the subnetwork are desired.

Algorithm 4-1

Let T = mean throughput
U_i = queue dependent unnormalized service rate for i busy queues
G = normalizing constant
P = probability for state s
n_i = number of customers in queue i including the customer in service
k = number of queues
θ_i = relative throughput for queue i
N = population of original network
q = number of nonempty queues for state s

```
Begin
    Input N,k,θi, Ui  for i=1,...,k
    For n=1,...,N do/* calculate throughputs */
        /* initialization */
        T = 0.0
        G = 0.0
        For i=1,...,k do
            ni=0.0
        end
        For each feasible state (n1,...,nk) where
            n1+...+nk=n  do
            q=0.0
            For i=1,...,k do  /* find number of busy queues */
                If ni≠0 q=q+1
            end
            P=[q*(θ1**n1*...*θk**nk)]/Uq
            G = G+P
            T = T+(P*Uq)
        end
        T = T/G
        Print throughput
    end
end
```

Algorithm 4-2

Let
- G = normalizing constant
- U_i = queue dependent unnormalized service rate for i busy queues
- P = probability for state s
- n_i = number of customers in queue i including the number in service
- Θ_i = relative throughput for queue i
- L = mean queue length for all queues
- L_i = mean queue length for queue i
- W = mean waiting time for all queues
- W_i = mean waiting time for queue i
- T = total mean throughput for all queues
- T_i = mean throughput for queue i
- PB_n = probability of n busy queues
- BQ_i = probability queue i is busy (includes customer in service)
- BQ = mean number of busy queues (also equals number of customers in subsystem)
- C = mean number of customers in subsystem not being served
- MT = mean throughput of all queues
- R = mean subsystem residency time for all queues
- R_i = mean subsystem residency time for queue i
- CYC = mean cycle time for all queues
- CYC_i = mean cycle time for queue i
- BS = mean number of busy servers
- q = number of busy queues
- k = number of queues
- N = population of original network

```
Begin
     Input N, k, Θi, Ui, for i=1,...,k
     For n=1,...,N do /* compute performance parameters */
          /* initialize */
          G=0.0
          L=0.0
          T=0.0
          BQ=0.0
          MT=0.0
```

```
For i=1,...,N do
     n_i=0.0
     L_i=0.0
     T_i=0.0
     PB_i=0.0
     BQ_i=0.0
end
For each feasible state do
     q=0.0
     For i=1,...,k do /* find number of busy queues */
          If (n_i≠0) q=q+1
     end
     P=[q*(θ_1**n_1*...*θ_k**n_k)]/U_q
     G=G+P
     T=T+(P*U_q)
     PB_q=PB_q+P
     For i=1,...,k do /* compute queue related performance
     parameters */
          If (n_i≠0) then
               L_i=L_i+(n_i-1)*P
               T_i=T_i+(U(q)*P)/q
               BQ_i=BQ_i+P
          end
     end
end
T=T/G
L=L/G
BS=T/U(1)
CYC=N/T
For i=1,...,k do
     L_i=L_i/G
     T_i=T_i/G
     W_i=L_i/T_i
     BQ_i=BQ_i/G
     BQ=BQ+BQ_i
     PB_i=PB_i/G
     R_i=BQ_i/T_i
     CYC_i=W_i+R_i
     L=L+(L_i/q)
     MT=MT+(T_i/q)
end
```

```
                C=BQ-BS
                R=BQ/T
                print results
            end
        end
```

Example 4-1

As an example of the procedure consider a closed multi-entrance queue with four queues where an arriving customer will enter any queue with equal probability, and with service rates that depend on the number of nonempty queues as follows:

Number of Nonempty Queues	Throughput
1	2.0
2	4.0
3	6.0
4	8.0

With these throughputs, the mean time a customer requires use of the primary resource will not vary with the congestion level of the queue, and hence there are effectively four servers servicing four queues. As a result the model is equivalent to the ordinary product form network illustrated in Figure 4-5. In this example, the service rates that depend on the number of nonempty queues in the multi-entrance queue model are identical to the service rates that depend on the number of nonempty queues in the network in Figure 4-5. Therefore the population dependent service rates of both models will be equal and the two representations are equivalent. This can also be proved analytically by showing that the two sets of balance equations are identical. Throughput comparisons are given in Table 4-1 for network populations of one to thirty.

There was no difference in the throughputs obtained by either method for any of the thirty populations. This was reassuring as it was felt that the Algorithm 4-1 and Algorithm 4-2 would be susceptible to numerical instabilities. However, the values were calculated in single precision on a CDC Cyber 174/75, which has a sixty bit word size.

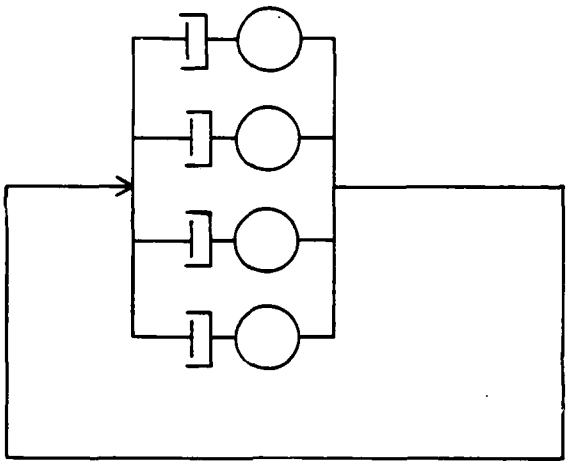


Figure 4-5. Example 4-1 I/O Subnetwork

In the more general case, the service rates at a group of queues can depend on the current utilization of the group. The usefulness of this type of service dependency is illustrated in the following example.

Table 4-1

Example 4-1 Multi-entrance Queue Throughputs

Network Population	Throughput Using Single Server Mean Value Analysis	Throughput Using Multi-entrance Queue Product Form
1	2.00000000	2.00000000
2	3.20000000	3.20000000
3	4.00000000	4.00000000
4	4.57142857	4.57142857
5	5.00000000	5.00000000
6	5.33333333	5.33333333
7	5.60000000	5.60000000
8	5.81818182	5.81818182
9	6.00000000	6.00000000
10	6.15384615	6.15384615
11	6.28571429	6.28571429
12	6.40000000	6.40000000
13	6.50000000	6.50000000
14	6.58823529	6.58823529
15	6.66666667	6.66666667
16	6.73684211	6.73684211
17	6.80000000	6.80000000
18	6.85714286	6.85714286
19	6.90909091	6.90909091
20	6.95652174	6.95652174
21	7.00000000	7.00000000
22	7.04000000	7.04000000
23	7.07692308	7.07692308
24	7.11111111	7.11111111
25	7.14285714	7.14285714
26	7.17241379	7.17241379
27	7.20000000	7.20000000
28	7.22580645	7.22580645
29	7.25000000	7.25000000
30	7.27272727	7.27272727

Example 4-2

Consider five queues in parallel as illustrated in Figure 4-6 where an arriving customer is equally likely to enter any queue. Suppose that a

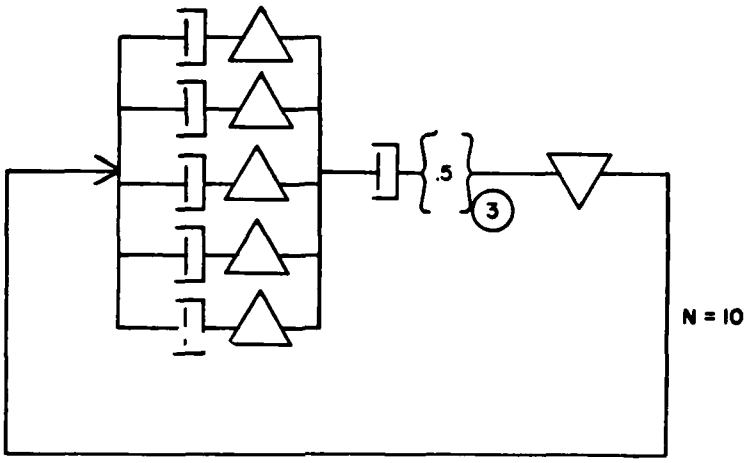


Figure 4-6. Example 4-2 Network to be Solved

customer has an exponentially distributed mean service demand of .5 and there are ten customers in the system. However, there are only three servers which must be distributed among all the queues. An approximate solution can be obtained by using the multi-entrance queue model if it is assumed that customers at the head of each queue processor share the available service capacity. The solution is approximate and not exact because the multi-entrance queue model assumes that the residency time at the server is exponentially distributed. However, this assumption is not true as it is the service time that is exponentially distributed. The difference can also be seen by considering the maximum number of customers that can be in service. Assuming a processor share service division, five customers can be in service whereas only three customers can be in service assuming a FCFS service discipline. The error caused by the differences

and other sources of error will be discussed in Chapter 5. Throughputs are given in Table 4-2 for customer populations one to ten. Note that when the population is three or less, the network can be solved using standard product form techniques.

Table 4-2

Example 4-2 Multi-entrance Queue Throughputs

Network Population	Throughput T
1	2.000
2	3.333
3	4.286
4	4.884
5	5.250
6	5.478
7	5.625
8	5.723
9	5.789
10	5.837

The above examples illustrate the immediate usefulness of the multi-entrance queue model. In the following section the procedure for using the model to solve type two simultaneous resource possession problems is is given and illustrated.

Applicability

Several assumptions were required during the development of the multi-entrance queue. First, it was assumed that the nonoverlapped service requirement for each primary resource was equal. If this is not true then the network in Figure 4-1 will not be equivalent to the one in Figure

4-2, which is the basis of the development of the augmented secondary subsystem model and the multi-entrance queue. Hence, the model may not be representative of the original network.

Second, it was assumed that only one primary resource could be allocated at each entry queue. If this is not true then the multi-entrance queue is not applicable. It can only consider allocation centers with a single primary resource.

Third, the multi-entrance queue was developed using a processor share service discipline to subdivide the capacity of a server whose residency time distribution is exponential. These assumptions are not true in many applications of interest. The error introduced by these assumptions will be a topic of discussion in Chapter 5.

In the following section a procedure is given that uses the multi-entrance queue model to approximately solve type two simultaneous resource possession problems where all primary resource nonoverlapped service requirements are equal and each entry queue only allocates a single resource.

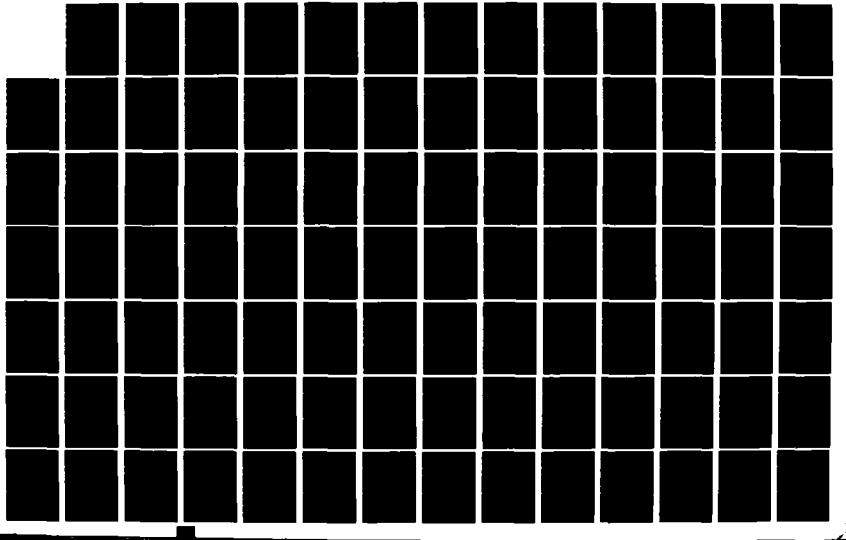
Procedure

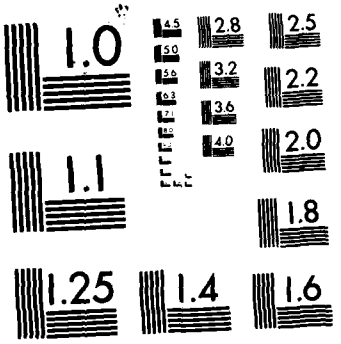
The complete step by step procedure is similar to the ECM procedure except at the step where the multi-entrance queue is evaluated in isolation to determine the population dependent throughputs which are used as service rates in the variable rate queue that replaces the subnetwork. The procedure follows:

1. Calculate the following values:

AD-A124 877 A NEW APPROACH FOR SOLVING SIMULTANEOUS RESOURCE POSSESSION PROBLEMS IN C..(U) AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOOL OF ENGI.. D J FREUND 2/3

UNCLASSIFIED DEC 82 AFIT/GCS/MA/82D-2 F/G 12/1 NL





a. The mean nonoverlapped service demand for all primary resources, NO, using equation 3-2

b. The mean overlapped service demand, OL, for the secondary subsystem using equation 3-3.

These steps are identical to those outlined in the ECM procedure.

2. Construct and evaluate the augmented secondary subsystem model. This model is identical to the one used in both methods presented in the previous chapter. The throughputs, $T_a(n)$, for n=1,...,k, should be obtained.

3. Create the multi-entrance queue model, illustrated in Figure 4-7, with k queues. Recall that only one resource can be allocated at each queue.

a. Determine the relative throughputs θ_i, for i=1,...,k, which are given by

$$\theta_i = \sum_{j=1}^{k} \theta_j P_i \qquad (4\text{-}19)$$

(4-2) repeated

where

P_i = probability that a customer leaving service will next enter queue i

An arbitrary θ_i is chosen constant to provide a unique solution to equation 4-19 (In all cases presented, θ_1=1.0, was used with satisfactory results).

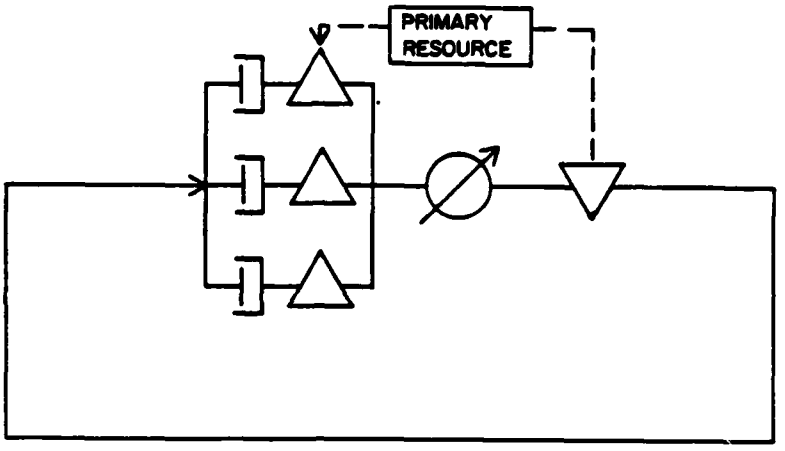


Figure 4-7. Multi-entrance Queue Evaluated in Isolation

Figure 4-7

b. Determine the queue dependent service rates, which are given by

$$U_m(q) = T_a(q), \text{ for } q=1,\ldots,k \qquad (4\text{-}20)$$

The queue dependent service rates are equal to the set of population dependent throughputs obtained from the augmented secondary subsystem. In the program used to solve the multi-entrance queue model, unnormalized service rates are used. This could equivalently be translated into a mean service demand and set of normalized service rates. In this case the mean service demand, SD_m, is given by

$$SD_m = \frac{1.0}{T_a(1)} = (NO+OL), \qquad (4\text{-}21)$$

and the normalized service rates are given by

$$U_m(q) = \qquad \text{for } q=1,\ldots,k \qquad (4\text{-}22)$$

c. Solve the multi-entrance queue model using Algorithm 4-1 or 4-2 for each feasible network population (from one to N). Obtain for each population the following:

Throughput $T_m(n)$

If performance estimates within the subnetwork are desired later, also obtain:

$L_i(n)$ mean queue length for queue i, i=1,...,k
$W_i(n)$ mean waiting time for queue i, i=1,...,k
$PB^n(j)$ probability that k servers are busy when n in model, for j=1,...,k, Note $PB^n(0)=0.0$ as long as $n \neq 0$, and $PB^0(0)=1.0$

4. Replace the subnetwork with simultaneous resource possession in the original network with a variable rate queue as shown in Figure 4-8.

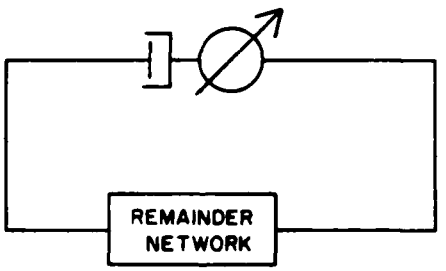


Figure 4-8. Original Network With the Subnetwork Replaced With a Variable Rate Queue

The mean service demand of the variable rate queue is given by

$$SD_f = \frac{1}{T_m(1)} \qquad (4\text{-}23)$$

and the normalized service rate function by

$$U_f(n) = \frac{T_m(n)}{T_m(1)}, \qquad \text{for } n = 1,\ldots,N \tag{4-24}$$

Solve the resulting network using conventional product form methods. If performance values within the subnetwork are needed, obtain the marginal probabilities, $P_f(n)$, of n residing in the variable rate node, for n=1,...,N. This is an output of the solution to the network in Figure 4-8.

5. Obtain the subnetwork performance parameters if desired.

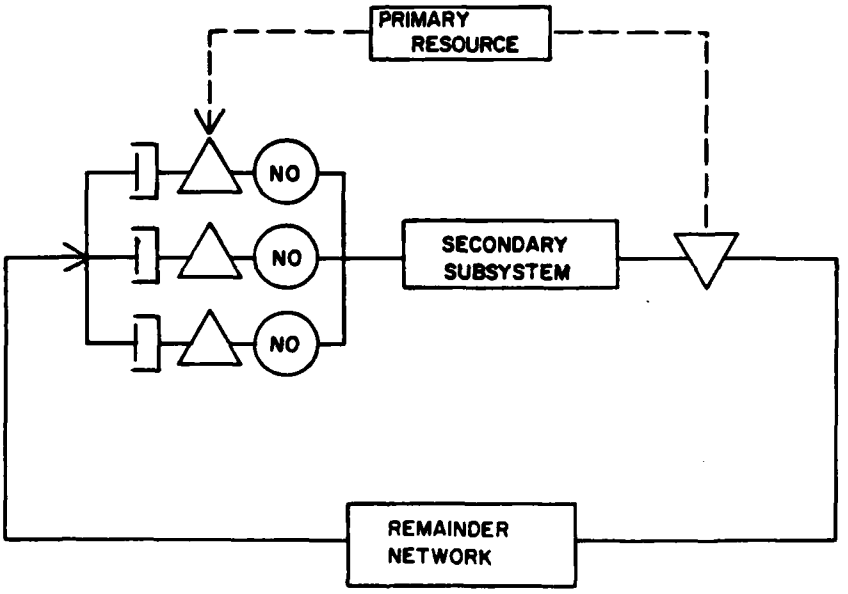


Figure 4-9. Original Network With Expanded Subnetwork Representation

a. The mean primary resource allocation center queue length L_i, as illustrated in Figure 4-10, for queues i=1,...,k is approximated by:

$$L_i = \sum_{n=1}^{N} L_i(n) P_f(n) \tag{4-25}$$

where

$L_i(n)$ = mean queue length of entry queue i when the multi-entrance queue is evaluated with a population of n.

b. For performance parameters within the secondary subsystem let

X = desired mean performance parameter

$X_a(n)$ = mean value of performance parameter obtained by evaluating the augmented secondary subsystem model with a population of n.

Then an estimate for X is given by

$$X = \sum_{n=1}^{N} P_f(n) \sum_{j=1}^{k} PB^n(j) X_a(j) \qquad (4\text{-}26)$$

An example will be given to illustrate the procedure.

Example 4-3

Consider the I/O Model solved in Example 3-1 as illustrated again in Figure 4-10. It is desired to obtain performance estimates from the final model as follows

Throughput
Cycle Time
CPU Utilization
CPU Queue Length
I/O Node Queue Length

Additionally, it is desired to obtain performance estimates within the subnetwork as follows

Mean Entry Queue Length
Mean Channel Queue Length

The procedure follows.

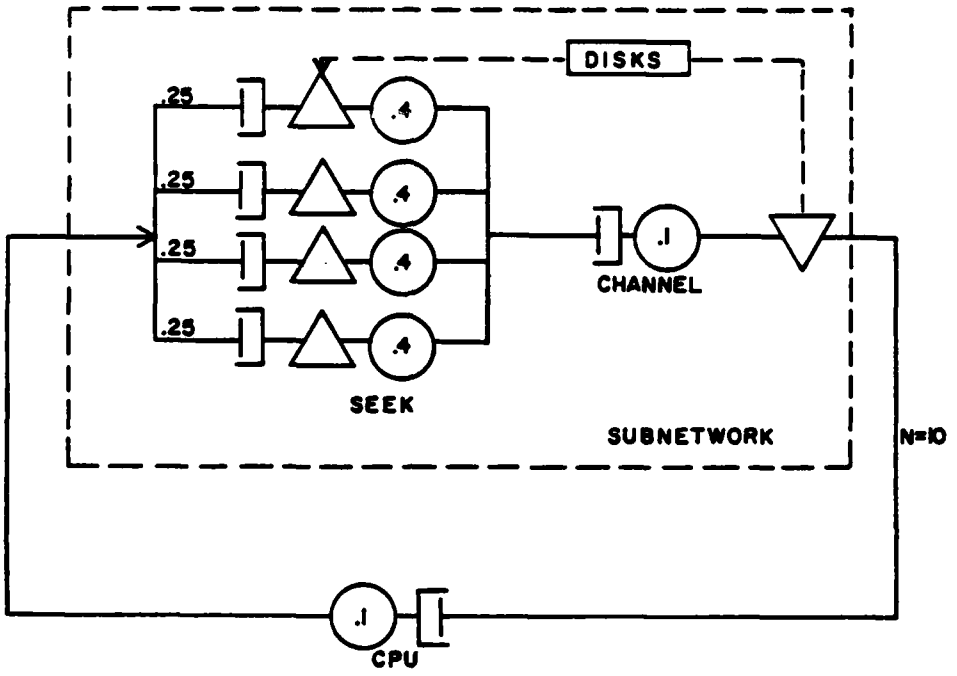


Figure 4-10. Example 4-3 CPU - I/O Model

Procedure:

1. The mean nonoverlapped service demand, NO, and overlapped service demand, OL, are the same as in Example 3-1 where NO=.4 and OL=.1.

2. The augmented secondary subsystem model is identical to the one in Example 3-1, the throughputs, $T_a(n)$, for n=1,...,4 being listed in Table 3-2. In addition to the throughputs, the mean channel queue length is needed at each augmented secondary subsystem population. These are given in Table 4-3, and were also obtained by evaluating the augmented secondary subsystem in isolation.

Table 4-3

Augmented Secondary Subsystem Channel Queue Lengths

Busy Servers	Mean Channel Queue Length
1	.20
2	.46
3	.80
4	1.24

3. The multi-entrance queue model for this example is illustrated in Figure 4-11. The model is evaluated for populations from one to ten.

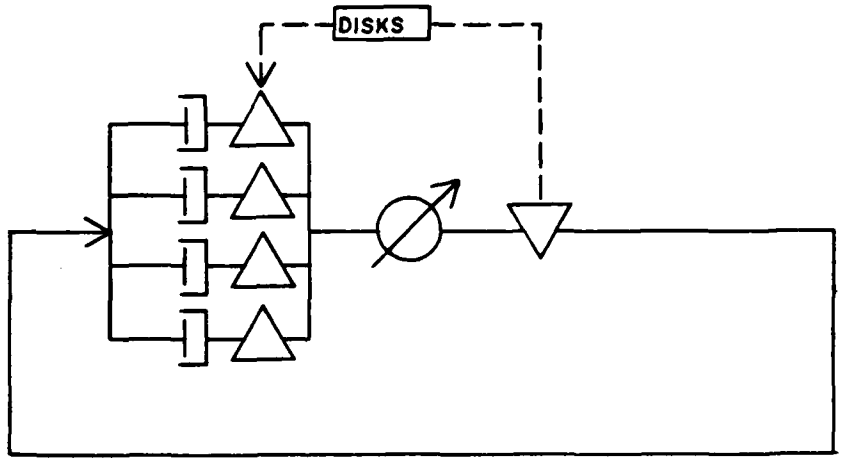


Figure 4-11. Multi-entrance Queue Evaluated in Isolation

a. The relative throughputs are given by equation 4-19:

$$\theta_i = \sum_{j=1}^{k} \theta_j P_i \qquad \text{(4-19) repeated}$$

where it was chosen to set θ_1=1.0.

Solving for the relative throughputs yields

θ_i=1.0 for all i=1,...,4. Also note that

$$P_i = \frac{\theta_i}{\theta_1+\ldots+\theta_k} = \frac{1}{4} = .25$$

b. The queue dependent service rates are given by equation 4-20:

$$U_m(n) = T_a(n), \text{ for } n=1,\ldots,4 \qquad (4\text{-}20)$$

repeated

and are listed in Table 4-4

Table 4-4

Example 4-1. Queue Dependent Service Rates

Population (n)	Unnormalized Service Rate [$U_m(n)$]
1	2.000
2	3.846
3	5.493
4	6.893

c. The multi-entrance queue model is solved and the population dependent throughputs, $T_m(n)$, are obtained for n=1,...,10. They are given in Table 4-5. Algorithm 4-2 was used so that the required subnetwork performance parameters could be calculated. The mean allocation queue length, L_i, and the set of busy queue probabilities, $PB^n(q)$, were required to obtain the necessary subnetwork performance parameters, and are listed for each subnetwork population in Table 4-6.

Table 4-5

Example 4-3. Population Dependent Throughputs

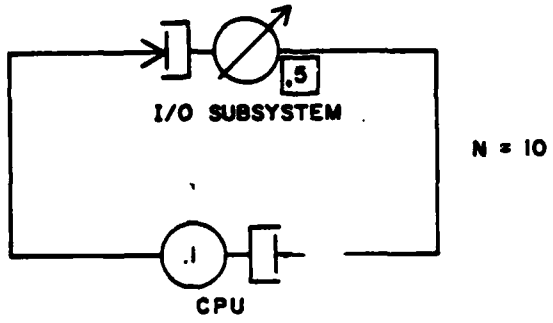
Population	Throughput $[T_m(n)]$	Unnormalized Service Rate $[U_f(n)]$
1	2.000	1.000
2	3.125	1.562
3	3.837	1.919
4	4.326	2.163
5	4.681	2.340
6	4.950	2.475
7	5.162	2.581
8	5.332	2.666
9	5.472	2.736
10	5.588	2.794

Table 4-6

Example 4-3. Multi-entrance Queue Performance Parameters

Population	Mean Entry Queue Length	Busy Server Probabilities PB^1	PB^2	PB^3	PB^4
1	0.00	1.00	0.00	0.00	0.00
2	.10	.39	.61	0.00	0.00
3	.25	.19	.60	.21	0.00
4	.42	.11	.51	.35	.03
5	.62	.06	.42	.44	.08
6	.83	.04	.35	.48	.13
7	1.04	.03	.29	.50	.18
8	1.27	.02	.24	.51	.23
9	1.49	.01	.21	.51	.27
10	1.72	.01	.18	.50	.31

4. The final model to be solved is illustrated in Figure 4-12. The mean service demand SD_f = .5, and the normalized service rates are given by equation 4-24 and are listed in Table 4-5.



Example 4-3 Final Network
Figure 4-12

The network in Figure 4-12 is product form and is solved using conventional methods. The required performance parameters obtained from analyzing this model are given in Table 4-7. Also needed from this model is the marginal distribution of customers at the variable rate queue. These values are listed in Table 4-8, rounded to three significant figures.

Table 4-7

Example 4-3. Performance Parameters

Performance Parameter	Mean Value
Throughput	5.425
Cycle Time	1.843
CPU Utilization	.543
CPU Queue Length	1.106
I/O Node Queue Length	8.894

Table 4-8

Example 4-3. I/O Subsystem Marginal Distribution of Customers

Number	Probability [$P_f(n)$]
0	.000
1	.001
2	.001
3	.004
4	.009
5	.019
6	.038
7	.075
8	.140
9	.256
10	.457

5. Obtain the subnetwork performance parameters.

a. The mean allocation queue length can be approximated by using equation 4-25

$$L_i = \sum_{n=1}^{N} L_i(n) P_f(n) \qquad (4\text{-}25)$$

repeated

where the values of $L_i(n)$ are listed in Table 4-6, and the values of $P_f(n)$ are listed in Table 4-8. In this case

$$L_i = L_1 P_f(1) + \ldots + L_{10} P_f(10)$$

$$= (0.0)(.001) + \ldots + (1.72)(.457)$$

$$= 1.47$$

b. The mean channel queue length, X, can be approximated by using equation 4-26

$$X = \sum_{n=1}^{N} P_f(n) \sum_{j=1}^{k} PB^n(j) X_a(j) \qquad (4\text{-}26)$$

repeated

where the values of $P_f(n)$ are listed in Table 4-8, the values of PB^n in Table 4-6, and the values of $X_a(n)$ in Table 4-3. In this case

$$X = P_f(1)[PB^1(1)X_a(1) + \ldots + PB^1(4)X_a(4)]$$

$$+ \ldots + P_f(10)[PB^{10}X_a(1) + \ldots + PB^{10}(4)X_a(4)]$$

$$= (0.00)[(1.00)(.20) + \ldots + (0.00)(1.24)]$$

$$+ \ldots + (.457)[(.01)(.20) + \ldots + (.31)(1.24)]$$

$$= (0.00)(.20) + \ldots + (.457)(.87)$$

$$= .83$$

All of the analytic performance parameters obtained by the multi-entrance queue procedure are compared against the 'true' values obtained by simulation in Table 4-9. It should be noted that not all of the performance estimates are within the confidence intervals of the simulated results. The possible causes of error are the topic of discussion in the next chapter.

Table 4-9

Example 4-3. Performance Comparisons

Performance Parameter	Multi-entrance Queue Value	Simulation Value	97% Confidence Interval
Throughput	5.425	5.63	(5.51-5.75)
Cycle Time	1.843	1.77	(1.74-1.81)
CPU Utilization	.543	.56	(0.54-0.58)
Mean CPU Queue Length	1.106	1.07	(0.99-1.16)
Mean I/O Node Queue Length	8.894	8.92	(8.84-9.01)
Mean Entry Queue Length	1.47	1.44	(1.42-1.47)
Mean Channel Queue Length	.83	.89	(0.85-0.92)

It must again be emphasized that the procedure presented in this chapter, and those in Chapter 3, are approximate. They usually do not provide exact solutions. The focus of the following chapter will be to analyze the sources of error in the procedure just presented. Limited comparisons will also be made with the other methods.

Although the multi-entrance queue procedure is more limited in its applicability than those presented in Chapter 3, it is applicable to most type two simultaneous resource possession problems of interest. The main advantage that it has over the other methods is the ability to obtain performance estimates of parameters within the subnetwork with simultaneous resource possession.

V. Comparison Of Techniques

Introduction

In the previous chapter the new multi-entrance queue procedure was presented for solving type two simultaneous resource possession problems. Two examples were given to illustrate the procedure. In this chapter its usefulness will be demonstrated by comparing its performance with the other two available methods and to results obtained by simulation. Finally the potential sources of error will be identified.

Initial Tests

The initial tests performed involved using all three methods to analyze the I/O model and the multiprocessor models discussed in Chapter 3. The I/O model used is illustrated in Figure 5-1 and was analyzed for 20 different parameter sets. In all 20 tests, the mean CPU service time, total I/O mean service time, number of disks, disk access probabilities, and the network population were held constant. The values used are listed in Table 5-1. All service times in the model were exponential.

The varying parameters consisted of the seek time, SD_s, rotational latency plus transfer time, SD_{ch} (channel access time), and the number of available channels, c. For each test the model was solved using the three available analytic methods and simulation.

The tests were broken into four sets, each consisting of five tests, where the number of available channels was held constant at one, two, three, and four. The four channel set was used as a control. As the multi-entrance queue procedure is equivalent to a central server representation when the number of disks equals the number of channels, the

analytic networks were exactly solved for this test set. Within each set five different seek/channel time combinations were used as shown in Table 5-2.

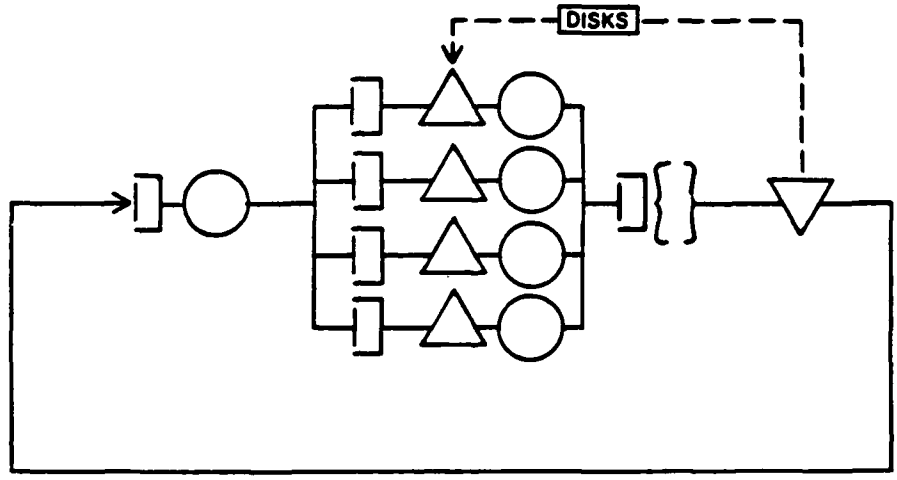


Figure 5-1. Base I/O Model Control Variables

TABLE 5-1

Base I/O Model Control Variables

Parameter	Value
Mean CPU service time	0.1
Total I/O Total I/O time	0.5
Number of available disks	4
Disk access probability (all disks)	.25
Network population	5

TABLE 5-2

Base I/O Model Specifications

Set Number	Model Number	Number Channels	Mean Seek Time	Mean Channel Time	Degree of Overlap
1	1-1	1	.10	.40	.80
	1-2	1	.20	.30	.60
	1-3	1	.30	.20	.40
	1-4	1	.40	.10	.20
	1-5	1	.45	.05	.10
2	2-1	2	.10	.40	.80
	2-2	2	.20	.30	.60
	2-3	2	.30	.20	.40
	2-4	2	.40	.10	.20
	2-5	2	.45	.05	.10
3	3-1	3	.10	.40	.80
	3-2	3	.20	.30	.60
	3-3	3	.30	.20	.40
	3-4	3	.40	.10	.20
	3-5	3	.45	.05	.10
4	4-1	4	.10	.40	.80
	4-2	4	.20	.30	.60
	4-3	4	.30	.20	.40
	4-4	4	.40	.10	.20
	4-5	4	.45	.05	.10

The multiprocessor model used is illustrated in Figure 5-2, and was analyzed for five different customer populations from 6 to 10. It was solved analytically using only the ECM procedure and the multi-entrance queue procedures.

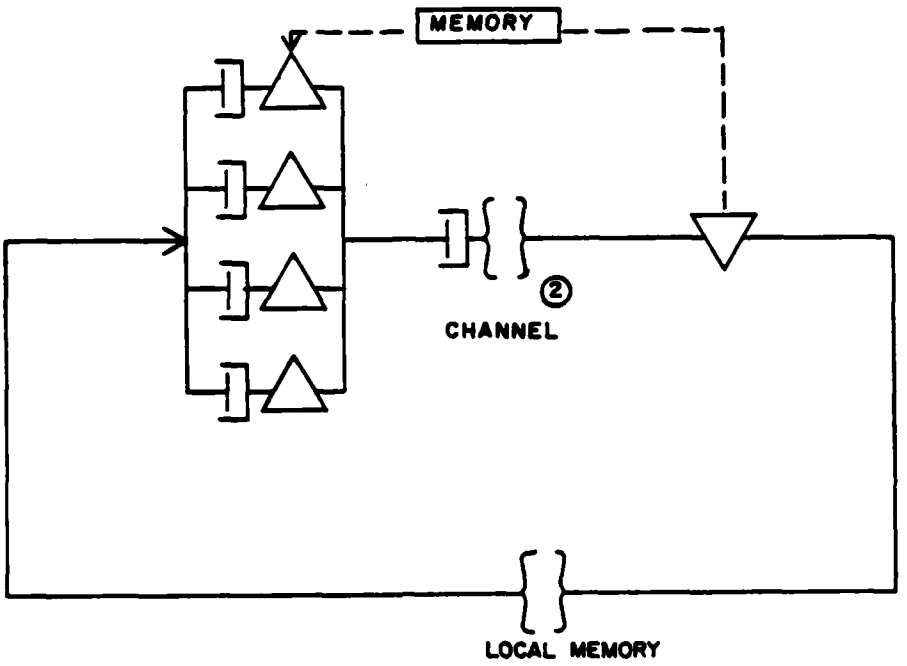


Figure 5-2. Test Set 5 - Loosely Coupled Multiprocessor

The model numbers are listed in Table 5-3, Base Multiprocessor Model Specifications, with the network population. All other parameters were held constant as identified in Figure 5-2.

TABLE 5-3

Test Set 5 - Base Multiprocessor Model Specifications

Model Number	Network Population
5-1	6
5-2	7
5-3	8
5-4	9
5-5	10

The values of the throughputs obtained for each model are given in Table 5-4. The table contains the throughput and percent error for each of the analytical methods, and the throughput and 97% confidence interval for the simulation results. The throughputs are also graphically displayed for each test set in Figures 5-3 through 5-7. The confidence intervals were obtained by making five separate simulation runs for each model and

Table 5-4
Initial Model Throughputs

Model Number	MEQ Method Throughput	MEQ Method Percent Error	ECM Method Throughput	ECM Method Percent Error	Surrogates Method Throughput	Surrogates Method Percent Error	Simultion Throughput	Simultion 97% CI
1-1	2.45	-.4	2.46	0.0	2.49	1.2	2.46	2.37-2.56
1-2	3.04	-2.6	3.05	-2.2	3.22	3.2	3.12	3.01-3.22
1-3	3.75	-4.1	3.75	-4.1	4.03	3.1	3.91	3.82-4.01
1-4	4.37	-4.0	4.37	-4.0	4.43-4.61*		4.55	4.49-4.62
1-5	4.55	-2.6	4.55	-2.6	4.58-4.61*		4.67	4.61-4.74
2-1	4.14	-2.8	4.21	-1.2	4.20	-1.4	4.26	4.12-4.41
2-2	4.40	-4.8	4.42	-4.3	4.35-4.61*		4.62	4.50-4.74
2-3	4.54	-5.4	4.45	-5.2	4.54-4.61*		4.80	4.70-4.90
2-4	4.60	-3.8	4.60	-3.8	4.61-4.61		4.78	4.71-4.86
2-5	4.61	-2.5	4.61	-2.5	4.62-4.61		4.73	4.65-4.80
3-1	4.58	-3.6	4.61	-2.9	4.54-4.61*		4.75	4.61-4.89
3-2	4.60	-5.3	4.61	-5.1	4.60-4.61*		4.86	4.75-4.98
3-3	4.61	-5.3	4.61	-5.3	4.62-4.61		4.87	4.78-4.98
3-4	4.61	-3.8	4.61	-3.8	4.62-4.61		4.79	4.72-4.87
3-5	4.61	-2.5	4.61	-2.5	4.62-4.61		4.73	4.66-4.80
4-1	4.61	-3.6	4.61	-3.6			4.78	4.65-4.93
4-2	4.61	-5.4	4.61	-5.4			4.87	4.76-4.99
4-3	4.61	-5.5	4.61	-5.5	Not Run		4.88	4.78-4.98
4-4	4.61	-3.8	4.61	-3.8			4.79	4.72-4.87
4-5	4.61	-2.5	4.61	-2.5			4.73	4.66-4.80
5-1	1.59	0.0	1.64	3.1			1.59	1.56-1.63
5-2	1.73	0.0	1.78	2.9			1.73	1.68-1.77
5-3	1.83	.5	1.88	3.3	Not Run		1.82	1.77-1.88
5-4	1.89	0.0	1.94	2.6			1.89	1.82-1.95
5-5	1.93	0.0	1.97	2.1			1.93	1.86-2.00

assuming the throughputs were independent samples from a normal distribution. In all intermediate steps of the analytical methods, parameters were recorded using the full accuracy of the computer. Only the final results were rounded to three significant figures.

After analysis of the initial test results several observations were made:

1. The method of surrogates did not converge for most of the models tested.

2. Most of the throughputs obtained analytically fell outside the confidence intervals of the simulated results, including the four channel control group.

These points will be discussed in turn.

It was discovered during these runs that the method of surrogates did not always converge. These runs are illustrated by the presence of two throughputs under the method of surrogates throughput column in Table 5-4. The first value listed is the most recent value of the secondary contention model throughput before divergence was detected, and the second value is the most recent value of the primary contention model throughput before divergence. If divergence occurred on the first iteration then the first iterate values are listed. This method was not used in test sets four and five; but, of the first three sets (15 runs) only four runs converged, which required an average of five iterations. Of the twelve that did not converge, six had throughput values from the initial secondary contention

model iterate that were higher than the throughput values of the initial primary contention model iterate. In the remaining six runs, the sequences of secondary contention model throughputs were not bounded. Table 5-5 lists the iterate sequences for model 2-4, which had the type of divergence first mentioned. Table 5-6 lists the iterate sequences for model 1-4, which had the type of divergence mentioned second. The method of surrogate throughput sets marked with an asterisk in Table 5-4, had the second type of divergence, the others had the first type.

Most of the throughputs obtained analytically by the multi-entrance queue procedure fell outside the confidence intervals of the simulated results. In all cases where this occurred the analytic results predicted lower throughputs than the actual results. What was surprising was that all of the analytic throughputs in the four channel control group were significantly less than the actual throughputs. This was not expected as the multi-entrance queue procedure is equivalent to a central server representation of the CPU - I/O system with this test group, and hence has product form and can be solved exactly. After these results were obtained a separate simulation was run using a different procedure and programming language which yielded similar results.

The throughput responses for the models in the four channel control group are graphically displayed in Figure 5-3. Although none of the throughputs are significantly different in this group, more simulations were run on model 4-5 and model 4-3 to obtain additional accuracy. These simulations yielded throughputs of the two models that were significantly different. Model 4-5 had a mean throughput of 4.74 with a 97 percent confidence interval of (4.68-4.79), and model 4-3 had a mean throughput of 4.89 with a 97 percent confidence interval of (4.81-4.96). This was

TABLE 5-5

Method of Surrogate Iterations (Model 2-4)

Primary Contention Model				Secondary Contention Model		
Ident	Input Queueing Delay	Throughput	Primary Resource Residency	Input Queueing Delay	Throughput	Resource Residency
1	0	4.609	.9236	.4236	4.614	5.092
2	.0092	4.590	.9195	.4195		

TABLE 5-6

Method of Surrogate Iterations (Model 1-4)

Primary Contention Model				Secondary Contention Model		
Ident	Input Queueing Delay	Throughput	Primary Resource Residency	Input Queueing Delay	Throughput	Resource Residency
1.	0	4.609	.9236	.4236	4.430	.5558
2.	.0558	4.491	.8995	.3995	4.513	.5578
3.	.0578					

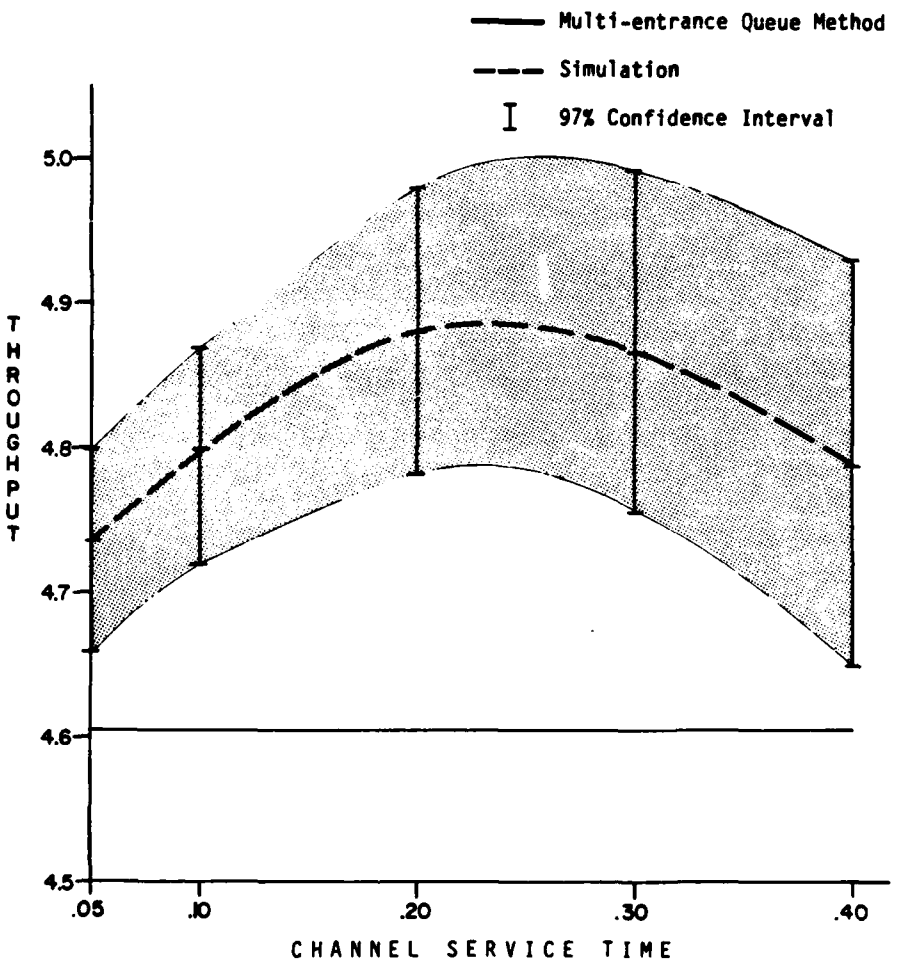


Figure 5-3. Test Set Four Final Network Throughputs

initially very counter intuitive, as it was felt that with four channels the ratio of the seek time to the channel time (degree of overlap) should not matter.

A very similar throughput response, illustrated in Figure 5-4, was obtained for the three channel test group. In fact the absence of one channel had almost no impact on the throughput means and the shape of the response curve. All of the simulated throughputs were significantly different from the multi-entrance queue throughputs. The throughputs of the two channel test group also produced a similar response shape, which is illustrated in Figure 5-5. The lack of two channels finally showed up in the throughputs of model 2-4 and model 2-5, and were lower than the throughputs of the previous groups. The throughput of model 2-5 was significantly lower than the throughputs of model 3-5 and model 4-5. The throughputs for model 2-1, model 2-2, and model 2-3 however, were significantly higher than the analytic four disk central server model throughputs (the analytic result with four channels)! The throughputs for the multi-entrance queue models were significantly different than the simulated throughputs except for model 2-1.

There were no surprises in the results of the one channel test group where the throughput response curves were similar (Figure 5-6). The simulated throughputs of model 1-1, model 1-2, and model 1-3 were significantly different form the multi-entrance queue throughputs, however these differences were always less than five percent.

The simulated throughput responses of the multiprocessor test group (test set five) agreed almost exactly with the multi-entrance queue throughputs. This is illustrated in Figure 5-7.

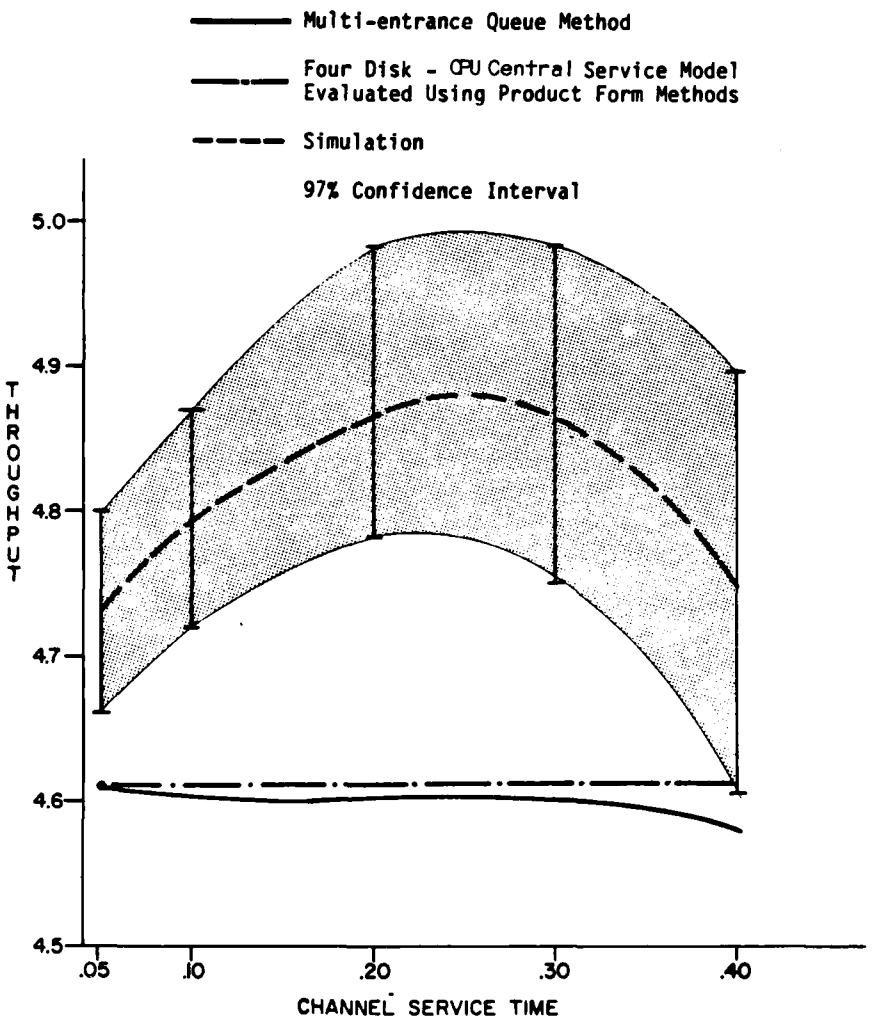


Figure 5-4. Test Set Three Final Network Throughputs

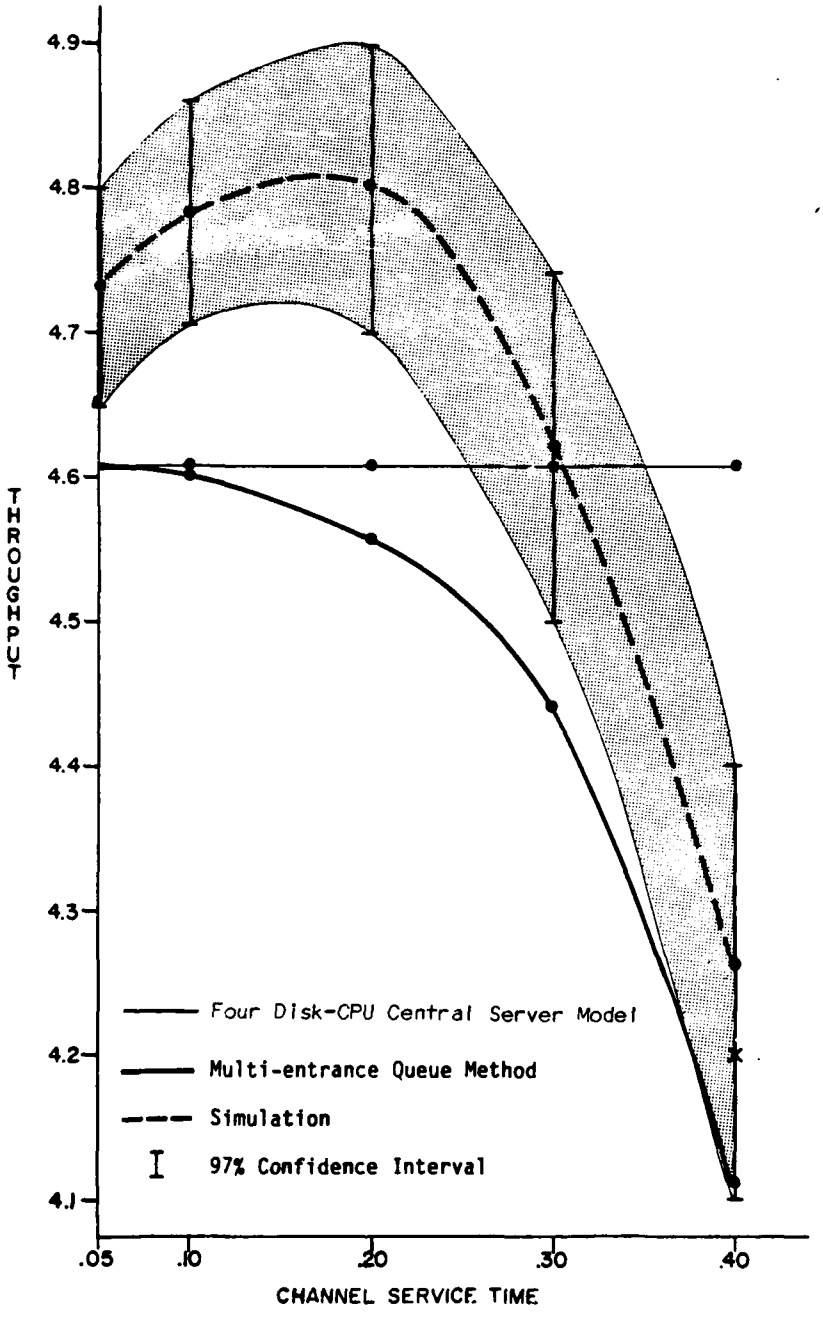


Figure 5-5. Test Set Two Final Network Throughputs

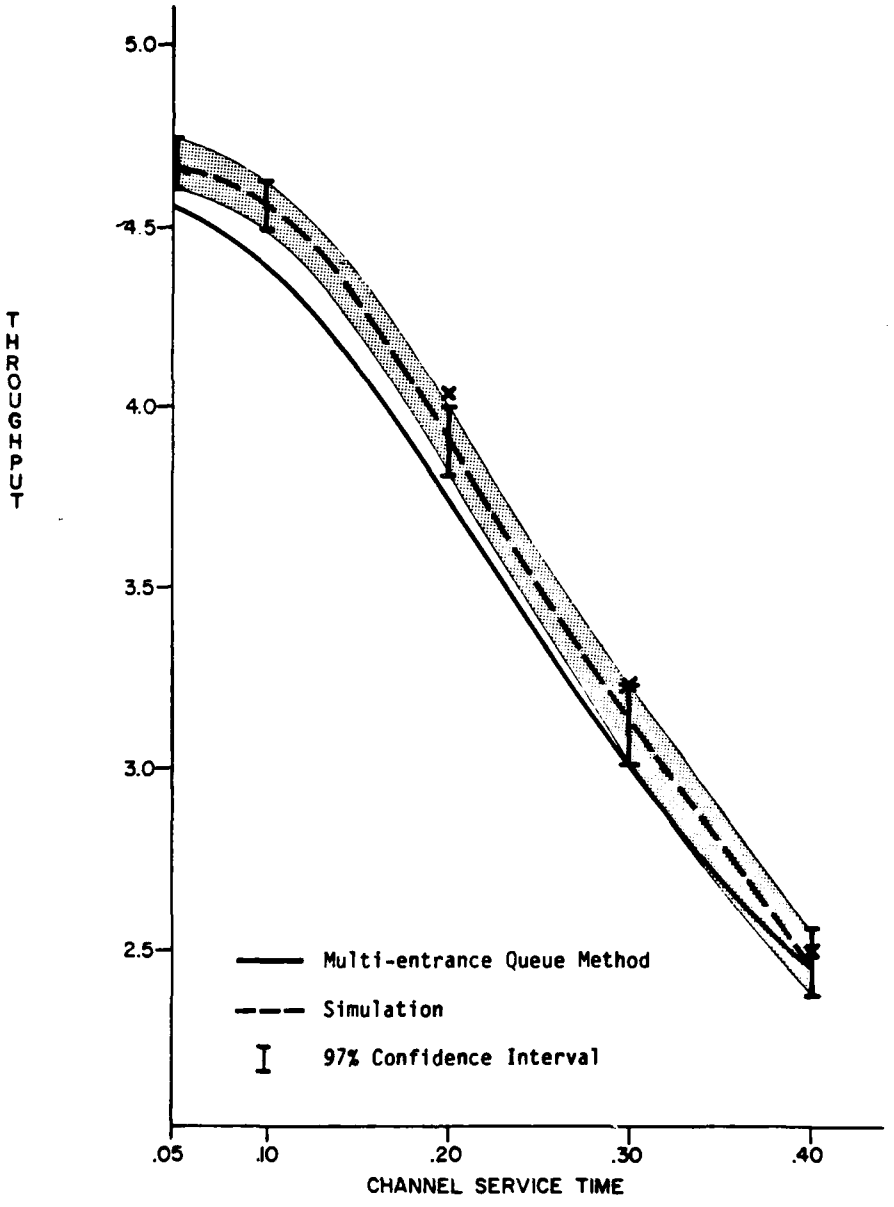


Figure 5-6. Test Set One Final Network Throughput

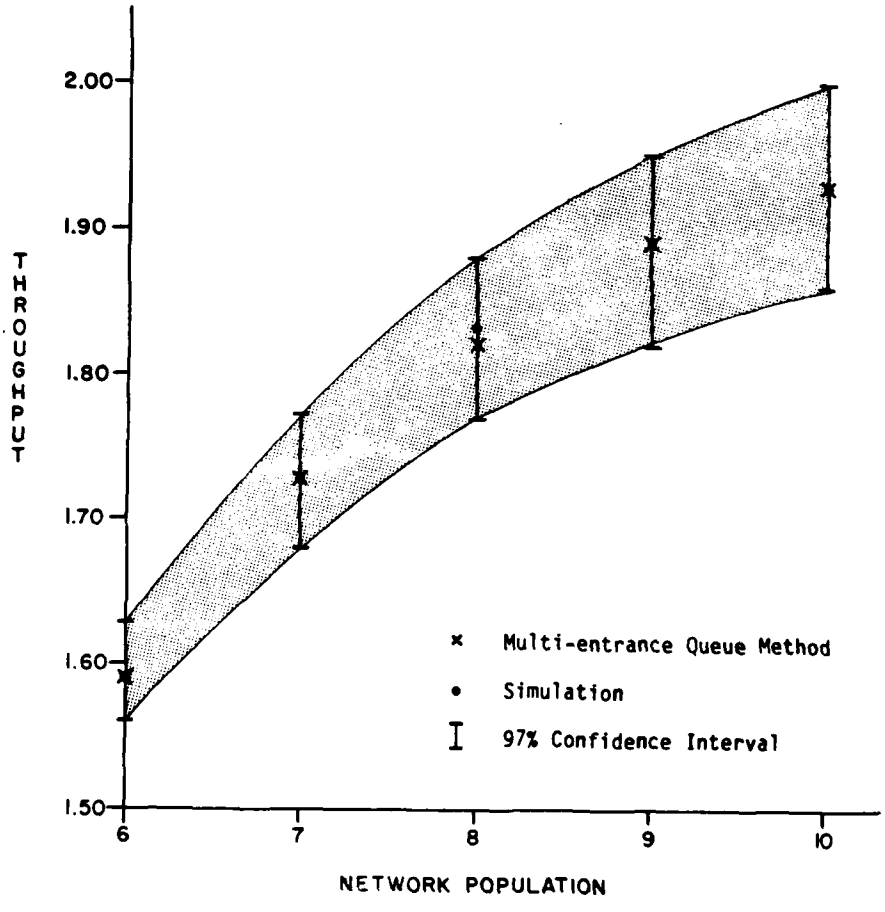


Figure 5-7. Test Set Five Final Network Throughputs

Overall, the error present in the multi-entrance queue derived throughputs was always less than six percent, and always lower than the actual result (when the error was significant). However, the shape of the response curves of test sets two, three, and four were not predicted by the multi-entrance queue procedure. For example, in the two channel test group the mean throughput for model 2-3 was significantly higher than the analytic result of the central server model (which had four channels). These differences led to the conclusion that whatever the source of the error, the cause was not represented in the network that was solved analytically. Note that in the four channel control group, this network can be solved exactly, and hence the error must be due to differences in the analytically solved network and the true representation.

It was believed that any possible error due to inaccuracies of the multi-entrance queue procedure in solving the analytic network representation was greatly overshadowed by the network representation error.

Additional tests were performed in order to isolate the sources of error and also to compare in a limited way the multi-entrance queue procedure to the other available methods. The error analysis is treated in three areas.

1. Departures of the MEQ method from the simulated results,

2. Limited comparison of the multi-entrance queue procedure and the ECM procedure, and

3. Limited comparison of the multi-entrance procedure and the method of surrogates.

A detailed analysis of the departures of the new method from the simulated results will be treated in the next sections. However, only a limited discussion will be provided to identify those sources of error in the other two methods or identify when they may be prone to error. Nevertheless, because the ECM procedure is similar to the MEQ procedure some of the analysis in the following section will apply to the ECM procedure as well. The comparisons of the MEQ procedure with the other two procedures consists of two examples evaluated at several different populations. The examples are intended to illustrate that the other methods also have errors, but are not meant to characterize or identify the sources of error.

It is hoped that by analyzing and characterizing the possible sources of error in the new procedure, confidence in the technique will be gained. Also an indication will be obtained as to when the procedure will perform well and when it may not.

Error Analysis of the New Procedure

It was believed that the error present in the analytic results of the initial test runs could have been caused by either of the following:

1. Inadequate or oversimplified representation of the subnetwork with simultaneous resource possession. This error is caused by transforming the original network into another nonequivalent network and analytically solving the nonequivalent network exactly. This error will be called the representation error.

2. That error due to decomposing nonproduct form networks. This error will be called decomposition error. Decomposition error can occur when a nonproduct form network is decomposed using product form networks. Consider the nonproduct form network illustrated in Figure 5-8. Assume that the network is nonproduct form as a result of the behavior in the subnetwork. Suppose that the subnetwork is analyzed exactly in isolation and population dependent throughputs are obtained and used in a variable rate queue in place of the subnetwork, as illustrated in Figure 5-9. Further suppose that the new network is solved using product form methods. The error in the performance parameters caused by replacing the subnetwork with a variable rate queue and using product form methods to solve the network (Figure 5-9) is called decomposition error.

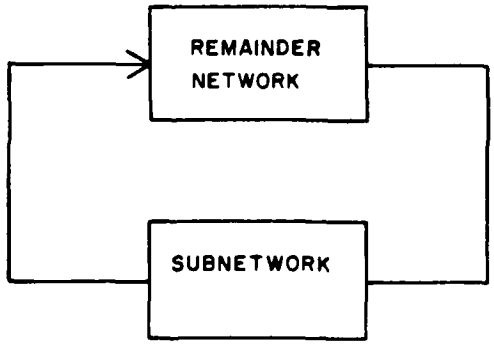


Figure 5-8. Original Network

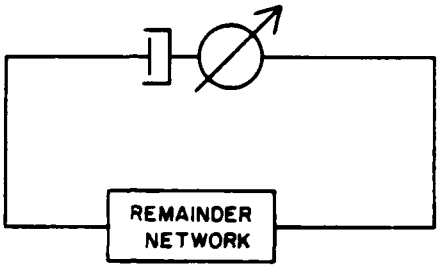


Figure 5-9. Original Network With the Subnetwork Replaced With a Variable Rate Queue

The analysis that follows partitions the sources of error present in the multi-entrance queue procedure into the above two categories. However, the representation error present in the multi-entrance queue method will be analyzed by evaluating the subnetwork in isolation exactly using simulation, and observing the difference in the multi-entrance queue obtained throughputs and the throughputs obtained using simulation. This will allow easy separation of the representation and decomposition errors, but requires the assumption that nonproduct form networks can be decomposed exactly as long as all the service time distribution aspects are carried along with the population dependent service rates in the aggregrated variable rate queues.

Representation Error

As stated earlier and illustrated by Figures 5-3 through 5-7, the throughput response of the initial test models were not totally expected. In fact, the shape of the analytically obtained response curves was not even

similar to the actual response curves in most cases. The initial conclusion was made that most of the inaccuracies were probably due to representation error. As a result a closer analysis of the I/O model was made.

Consider the four disk I/O Model illustrated in Figure 5-10.

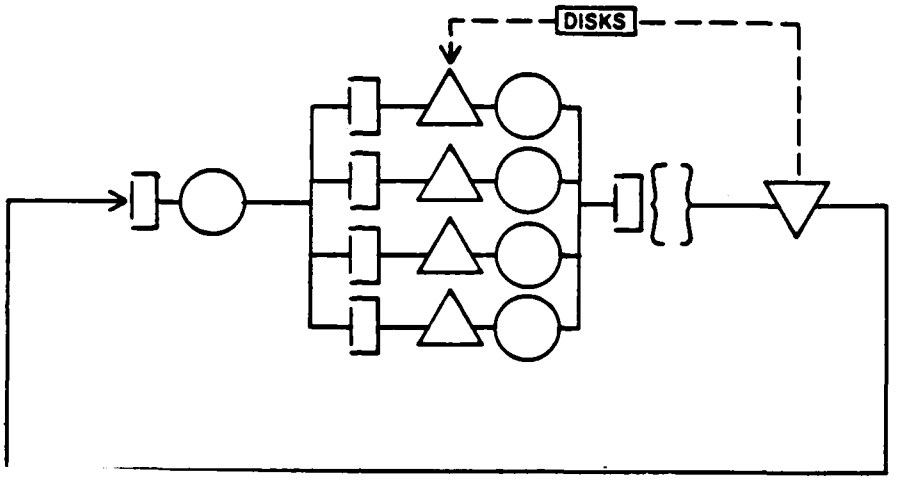


Figure 5-10. CPU - I/O Model

The multi-entrance queue procedure involves analyzing the I/O subsystem in isolation and obtaining a set of population dependent throughputs to be used in a variable rate queue, replacing the I/O subsystem. These networks are illustrated in Figures 5-11 and 5-12.

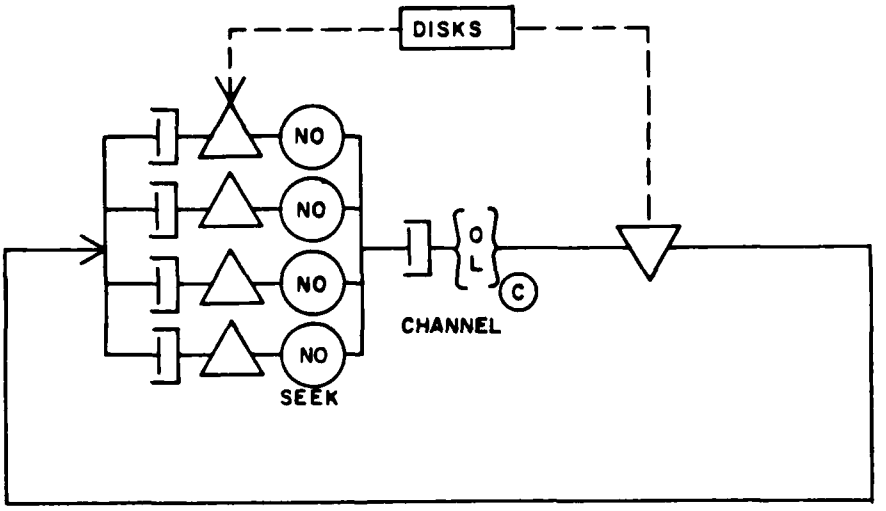


Figure 5-11. I/O Subsystem evaluated in Isolation

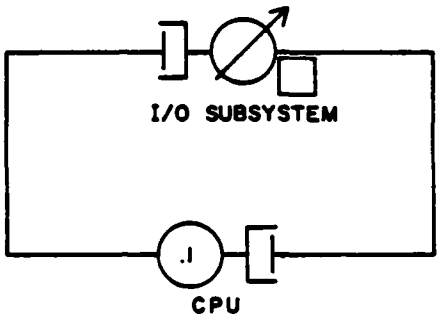


Figure 5-12. Final Network

The I/O subsystem as shown in Figure 5-11 is analyzed by using the augmented secondary subsystem model to determine a set of queue dependent service rates which are transformed into population dependent service rates by the multi-entrance queue model.

Recall that in the development of the augmented secondary subsystem model there was a departure from the actual representation (Figure 5-11) to the one illustrated in Figure 5-13. It was argued that these two networks were identical as long as all mean nonoverlapped service times were equal. An analysis was then made of the seek and channel queues together in isolation, the augmented secondary subsystem model, allowing their replacement with a variable rate queue yielding the network in Figure 5-14. It was this network that was solved exactly using the multi-entrance queue. Note that in the four physical channel models, the network in Figure 5-14 can be equivalently represented by the network in Figure 5-15, and therefore has a conventional product form solution. This is true because the throughput through each

primary resource allocation queue is not impacted by the states of the other queues (This is not true where there are less than four channels.) and the throughput capability through each queue is constant for all queue populations. In this case, the service rates that depend on the number of nonempty queues are identical for all feasible states to the service rates that are population dependent. Note that the network in Figure 5-15 can be described by a service rate function that is dependent on the number of nonempty queues. For example, if there are customers at two of the four queues in Figure 5-15, the aggregated service rate will be 2/(NO+OL).

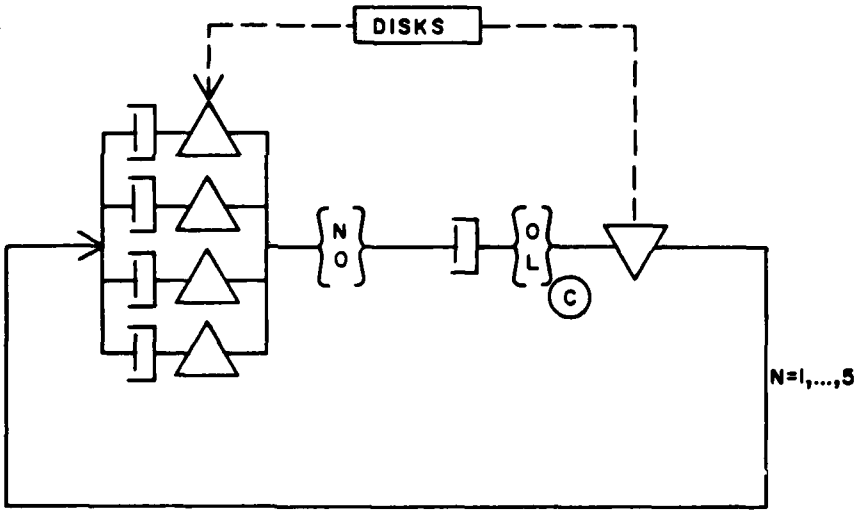


Figure 5-13. I/O Subnetwork Evaluated in Isolation With the Nonoverlapped Service Requirements Replaced With an Infinite Sever Queue

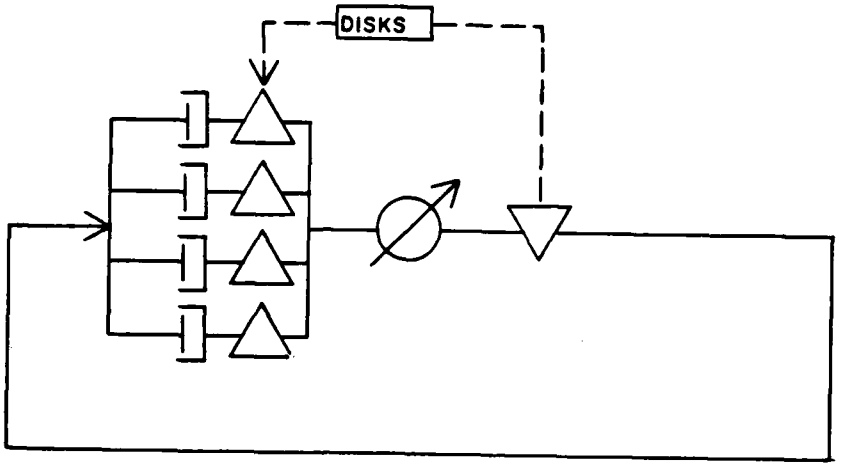


Figure 14. Multi-entrance Queue Model

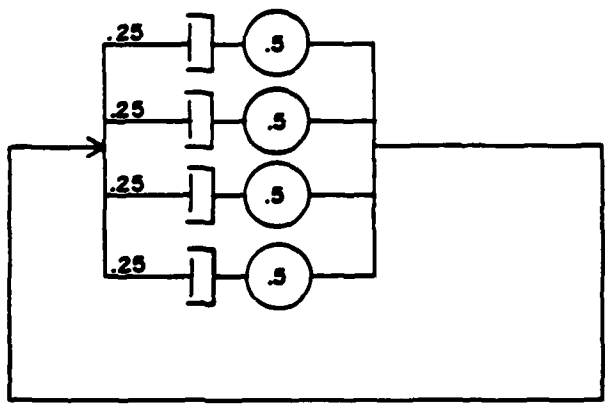


Figure 5-15. Equivalent Four Channel Model Subnetwork Representation

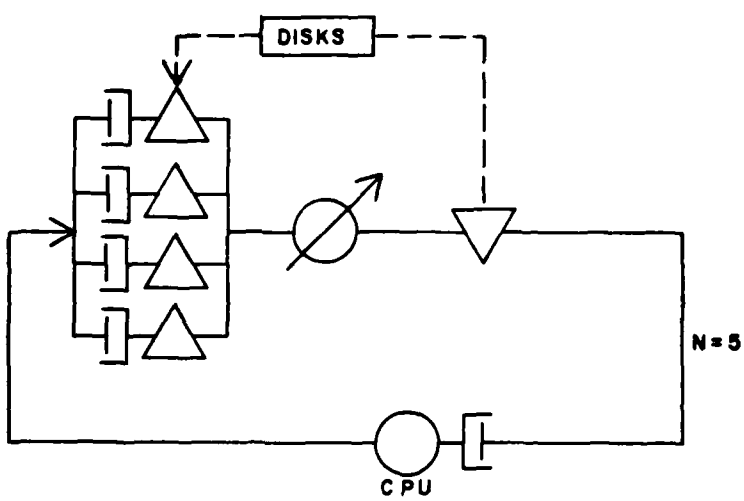


Figure 5-16. Multi-entrance Queue CPU - I/O Subsystem Representation

As a result of this special case the network in Figure 5-16 can be analytically solved exactly when there are the same number of secondary resources as primary resources, and therefore performance differences between the network in Figure 5-10 and the network in Figure 5-16 must be due to representation error. Hence, the only error that can be present in the four channel control group is representation error. By making a closer analysis of the network in Figure 5-13, the causes of this error were found.

Consider the network in Figure 5-13 and suppose there are four channels available and one customer is in the network. There would never be any waiting in the I/O subsystem and the mean throughput would be 1/(NO+OL). In this case, the network can be modeled exactly using product form techniques as long as the service time distributions have rational Laplace transforms. Now suppose there are two customers in the network and customer one is in service at disk i and customer two has just completed service and is arriving at the I/O subsystem. If customer two arrives at disk j, as long as $j \neq i$ he will proceed immediately into service. However, if he arrives at disk i, he must wait for customer one to complete service. Let us consider this situation more closely.

The above situation is depicted in Figure 5-17. Customer one is currently in service at disk i and is located somewhere in the dashed box. Customer two is arriving at the i-th primary resource allocation queue (entry queue). Notice that customer one proceeds through two stages of

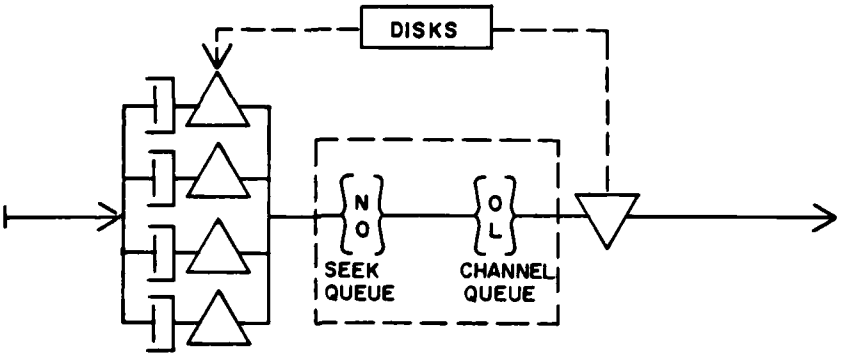


Figure 5-17. I/O Subnetwork

exponential service, the first stage consists of the seek queue, and the second the channel queue. Customer one never waits at these queues so arriving customer two at disk i sees customer one being serviced with a two stage hypoexponential service time distribution. His mean waiting time is determined by the residual life of the service time of customer one which is a function of the first two moments of the service time distribution. It can also be expressed in terms of the mean service time and service time coefficient of variation. Hence, it is expected that the mean waiting time function of an arriving customer involves the coefficient of variation of the service time distribution along with other parameters. Note that in this example the total waiting time for customer two consists of residual life service time of customer one.

Similar situations occur when there are more than two customers in the network. However, there can never be more than the number of primary resources, in this example four, in service simultaneously. If the seek

and the channel service times are considered together, then an arriving customer will always see a two stage hypoexponential service time distribution as illustrated in Figure 5-18.

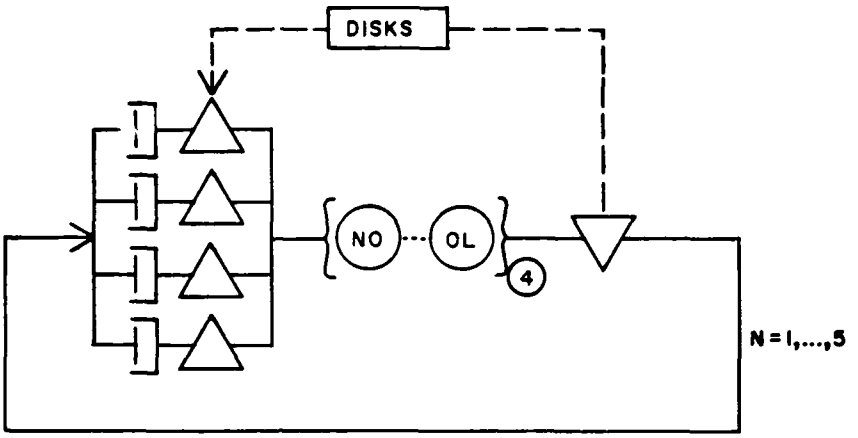


Figure 5-18. Subnetwork With a Two Stage Hypoexponential Service time Distribution

Now consider the network in Figure 5-13 but modified so that the number of physical channels, c, is less than the number of disks. As long as the network population is less than or equal to c, the above arguments apply and an arriving customer always sees a two stage hypoexpontential service time distribution. Now suppose that the network population is strictly greater than c. Such a network is illustrated in Figure 5-19. The channel and seek queues can no longer be treated as a two stage hypoexponential queue because the possibility for queueing exists at the channel. Furthermore, the amount of queueing at the channel can vary depending on the number of customers in service. This waiting time at the channel effectively changes the mean service time and the service time distribution that an arriving customer sees at the allocation queue. Thus when there are more than c customers in the network, an arriving customer has the possibility of seeing different mean service demands and service time distributions. The differences will depend on the current number of customers in service and consequently on the number of nonempty queues at the arrival instant.

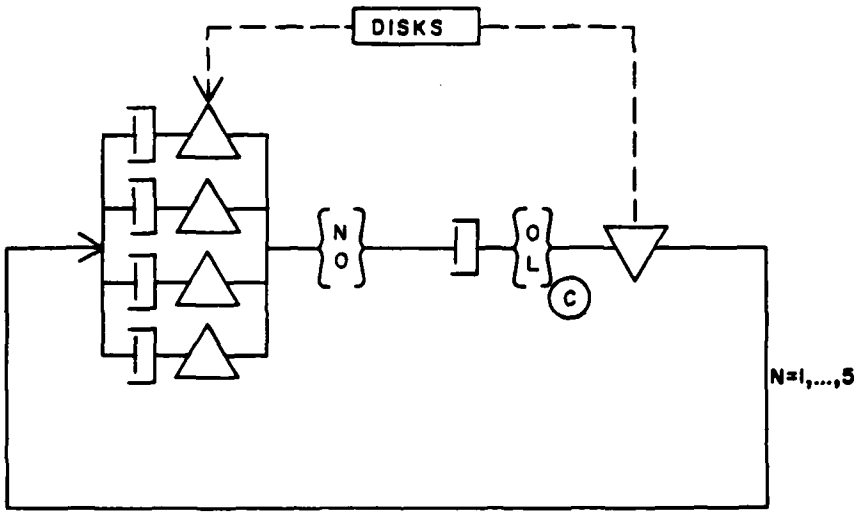


Figure 5-19. Subnetwork With a Variable Service Time Distribution

As an example consider the network in Figure 5-19 with two channels and four customers. Suppose that three customers are currently in service, one at disk one, one at disk two, and one at disk three. Since there are three customers in service and only two channels, queueing can (and will on the average) occur at the channel. This increases the mean time the customers in service will hold the primary resource, and consequently the service time the arriving fourth customer will see at the entry queue. Now suppose that one customer is in service at disk one, and two are queued at disk one. The arriving fourth customer will see the two stage hypoexponential service time distribution with no queueing because queueing can not occur at the channel. Note that in both cases there were four customers in the I/O subsystem (including the arriving customer). Hence, it is possible that with the subnetwork population held constant, an arriving customer can still have the possibility of seeing different mean service demands and service time distributions. This example confuses the definition of the base service demand of the subnetwork. In order to keep the definitions consistent new terminology will be defined.

Suppose that the stages of service and queueing in the subnetwork can be exactly represented by a single variable rate queue with the appropriate service time distribution. This network is illustrated in Figure 5-21 and is identical in performance (as viewed externally) to the subnetwork. The variable rate queue has a base service demand and rate function which describes the number of servers available. The service demand is constant for all states of the subnetwork and is given by equation 3-10. The total time spent holding the primary resource will be called the primary resource residency time. This is the "service time" an arriving customer sees.

Now consider the differences between the actual representation of the subnetwork, Figure 5-19, and the representation considered by the multi-entrance queue model, Figure 5-14. The actual representation of the subnetwork can be exactly represented by the network in Figure 5-20. The differences between this network and the one in Figure 5-14 are more easily shown. In the exact representation, Figure 5-20, the server may not be exponential and queueing can occur, where as in the multi-entrance queue model, Figure 5-14, the entire primary resource residency time is treated as the service time, which is exponential. Additionally, the server in the exact representation divides the service capacity to customers in FCFS order in most cases of interest, whereas the multi-entrance queue divides the service capacity to the customers by processor sharing. However, the exact representation service queue, Figure 5-20, may be considered to have a processor sharing service discipline as long as the residency time distribution is treated as the service time distribution for the processor share queue (This can be done because the customer service requirements are homogeneous, regardless of the entry queue. Hence, the FCFS discipline divides the service capacity the same as the processor share discipline in the long run). As a result, the differences between the two service disciplines can be ignored when making comparisons as long as the residency time distribution is used. The two models are identical only when the residency time distribution is exponential. Note however that even when the server in the exact representation is exponential, the residency time may no be.

To conclude, it is believed that the departures of the residency time distribution from exponential is responsible for error caused by representation error. However, this difference will not totally describe the degree of error present in the population dependent throughputs obtained from the multi-entrance queue model.

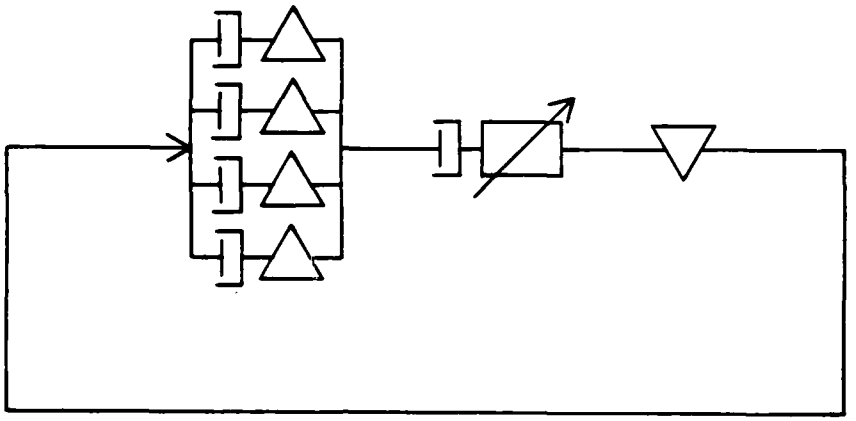


Figure 5-20. Subnetwork With the Service Requirement Represented by a Black Box Queue

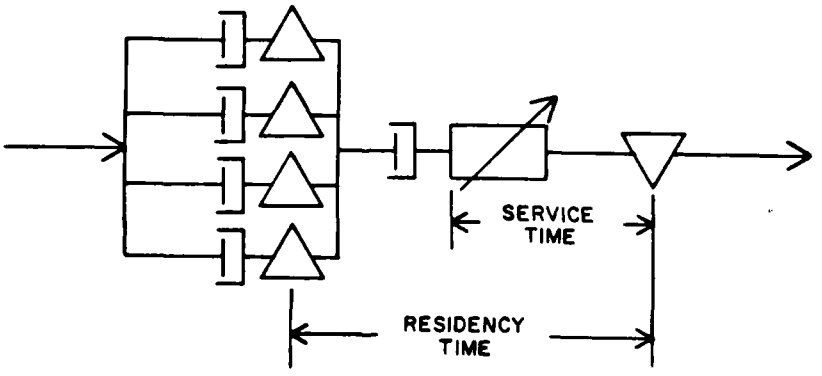


Figure 21. Subnetwork Service and Residency Times

Several parameters were thought to impact the degree of representation error present. These include:

1. The type of residency time distribution of the subnetwork. Recall that the mean service time (demand) consists only of the time a customer is in service, and does not include time queued within the augmented secondary subsystem. The residency time distribution will be characterized in terms of its mean and its coefficient of variation (really its first two moments). If is believed that as the coefficient of variation departs from 1.0, the degree of representation error will increase, and hence the potential for differences in throughput. If the coefficient of variation is 1.0, or close to 1.0, the first two moments will equal, or approximate those of an exponential distribution. The expected representation error will be very small. If the residency time distribution is exponential, no representation error can occur because the multi-entrance queue model is exactly solved. The degree of impact of the additional parameters listed hinges on the presence of the residency time distribution differences from exponential.

2. The degree of queueing, including but not limited to external contention that is present at the primary resource allocation centers. This is also a function of many parameters. Consider the subnetworks in Figure 5-22 and Figure 5-23. Suppose that in each case there are k

resources and N customers in the network. In Figure 5-22 queueing can occur with as

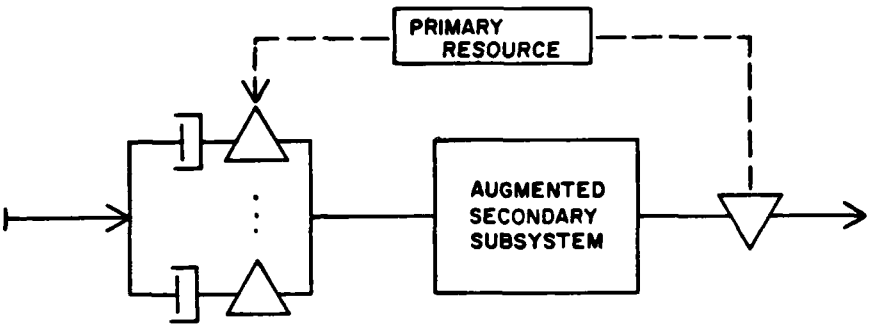


Figure 5-22. Subnetwork With External Contention Possible

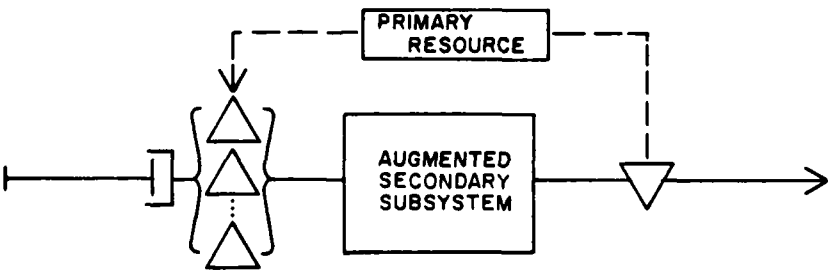


Figure 5-23. Subnetwork With External Contention Not Possible

little as two customers in the subnetwork, whereas in the network in Figure 5-23 no queueing can occur until at least k are in the subnetwork. During those times when there is no queueing, the allocation center can be removed and the network is product form. Hence, it is believed that the amount of time blocking occurs, and to which the allocation center is necessary, is directly proportional to the degree of difference possible (This example is an illustration of why type one problems are better behaved than type two problems). This will be measured as the ratio of time spent in the entry queue to the total residence time or cycle time of the subnetwork.

3. The degree of utilization of the augmented secondary subsystem, called the secondary utilization, which is defined as follows. Consider the augmented secondary subsystem of a subnetwork with simultaneous resource possession, as shown in Figure 5-24. The augmented secondary

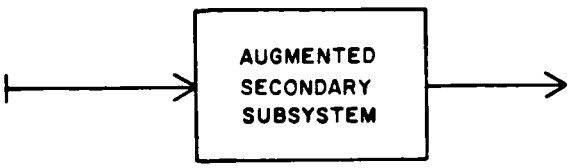


Figure 5-24. Augmented Secondary Subsystem

subsystem has a maximum mean throughput capability, called T_{max}, at the maximum possible population allowed in the augmented secondary subsystem. Recall that this is the thoughput before the effects of external contention are considered. Suppose that T_m represents the mean throughput through the subnetwork. Then the secondary utilization, X, is given by

$$X = \frac{T_m}{T_{max}} \qquad (5\text{-}5)$$

and is the ratio of the maximum mean throughput permitted to the maximum possible mean throughput capable through the augmented secondary subsystem. The secondary utilization is important because, as the secondary utilization approaches 1.0, the throughput and waiting time differences an arriving customer 'sees' due to the service time distribution differences will decrease. Also, it has meaning when the subnetwork is analyzed in isolation and when analyzed in the presence of the remainder network.

It is believed that the above three parameters can explain most of the differences in throughput caused by representation error. The intent is not to quantify these relationships but to show that their existence seems reasonable. To even prove their existence is beyond the scope of this paper, but willbe the subject of future investigation.

All of the aspects discussed seem to be considered by the multi-entrance queue model except the true residency distribution which can be different from exponential. The degree of distribution difference can be estimated by measuring the residency time mean and the coefficient of variation by the use of simulation. The distribution differences will impact the waiting at the entry queues but not the throughputs obtained from the augmented secondary subsystem model. This is because the flow of customers within the augmented secondary subsystem is not blocked or inhibited in any unusual way.

Decomposition Error

In the last section the error due to misrepresentation of the subnetwork with simultaneous resource possession was considered. The main focus involved examining the I/O subsystem, however, the concepts are the same for other examples. In this section the error due to decomposing nonproduct form networks will be considered. In this analysis it will be assumed that an exact procedure is available for obtaining the population dependent service rates. The source of error must then be caused by using the population dependent service rates in a variable rate queue when the underlying process cannot be totally characterized using exponential service time distributions.

The existence of decomposition error can be illustrated by a simple example. Consider the network in Figure 5-25 where customers alternate between a single CPU and a single I/O device. Let the CPU be represented by a single sever queue that has an exponential service time distribution with mean .1, and the I/O device be represented by a two server queue that has a ten stage Erlangian service time distribution with mean .5. Suppose there are five customers in the network. It is desired to decompose the I/O queue, and replace it with a variable rate queue that has an exponential service time distribution. If the I/O queue is analyzed in isolation for all possible customer populations, the service rates given in Table 5-7, I/O Node Throughputs, are obtained.

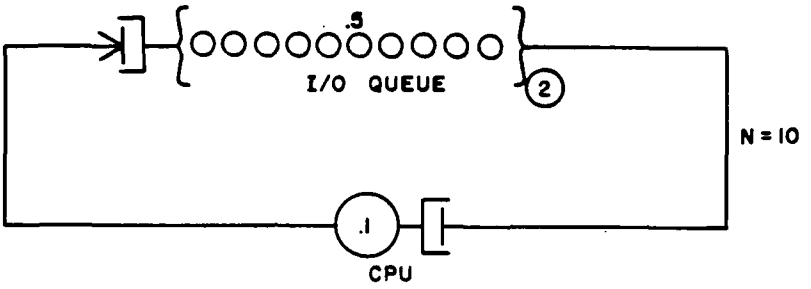


Figure 5-25. CPU and Single I/O Device

TABLE 5-6

I/O Node Throughputs

Customer population	Throughput	Service demand	Normalized service rate
1	2.0	.5	1.0
2	4.0	.5	2.0
>2	4.0	.5	2.0

It is not necessary to calculate the throughputs of the I/O queue evaluated in isolation because the number in the I/O node is always equal to the customer population of the network. Hence, the throughput is equal to the service rate at that population, regardless of the service time distribution. If the throughputs in Table 5-7 are used in a variable rate queue as illustrated in Figure 5-26, and the network is solved using

product form methods, an incorrect result is obtained. This result and the correct result, obtained by simulation, are listed in Table 5-8, I/O Node Throughputs.

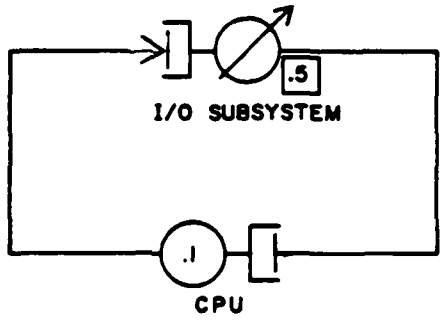


Figure 5-26. CPU and I/O Device Replaced With a Variable Rate Queue

TABLE 5-7

I/O Model Throughputs

Analytical Result	Simulation Result	97% Confidence Interval
3.957	3.997	(3.994-4.000)

Even though this was a trivial example it is apparent that the normalized service rates obtained from analyzing the network in isolation are identical to the original queue normalized service rates (because it was a two server queue). Hence, the only difference in the analytic and simulation evaluations was that the true distribution of the I/O queue was

considered in the simulation. This led to the belief that departures of the service time distribution of the variable rate node from exponential was responsible for the error. The next question was how would these departures impact the queue performance in the presence of multiple, and possibly fractional or variable rate servers.

In the previous section the service time residual life was used to explain the waiting time and resulting throughput differences caused by the service time distributions departures from exponential. It is desired to find a way to measure the differences in this case also. However, an arriving customer at a multiple server queue that sees all servers busy, and no customers queued, will wait the minimum of the residual life of all customers in service. This generally will not be equal to the residual life of the service time distribution, and therefore it was felt that the service time residual life would not useful in determining the differences in waiting times due to service time departures from exponential for multiple, and possible fractional server queues. However, a measure of this value was still desired so that it could be observed in simulation analysis. In order to determine a relevant measure of the minimum residual service wait, a closer analysis of multiple server and variable rate queues was made.

Consider a three server queue with unspecified arrival and service time distributions. The residual service wait will only occur when an arriving customer sees three or more customers currently at the queue. Figure 5-27 illustrates an arbitrary set of service demands with all three servers busy. Note that the minimum residual life of the busy servers is equal to the interdeparture residual life. This is also true when all servers are not busy, but these situations do not impact the queue performance because there is no queueing.

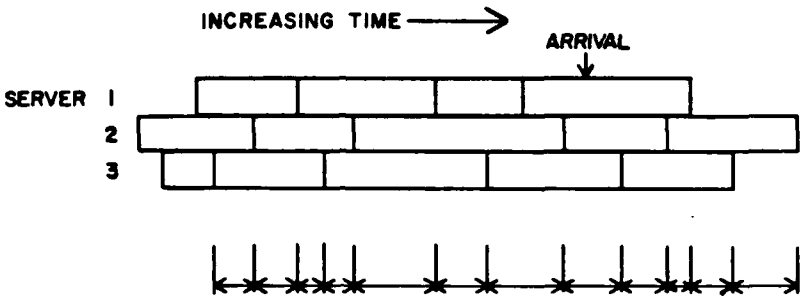


Figure 5-27. Interdeparture Time Residual Lives

Hence an estimate of the minimum residual life of all busy servers can be obtained by determining the interdeparture time residual life when all servers are busy. This can be done by analyzing the queue in isolation with the appropriate population using simulation.

The relationship is not as straightforward when more complicated variable rate queues are considered, such as the exact representation of the I/O subsystem illustrated in Figure 5-28 with a variable rate queue. This is caused by three factors. First, the queue may have fractional servers. Second, the possibility for queueing exists at the entry queues when the I/O subsystem is below its maximum throughput. Third, an arriving

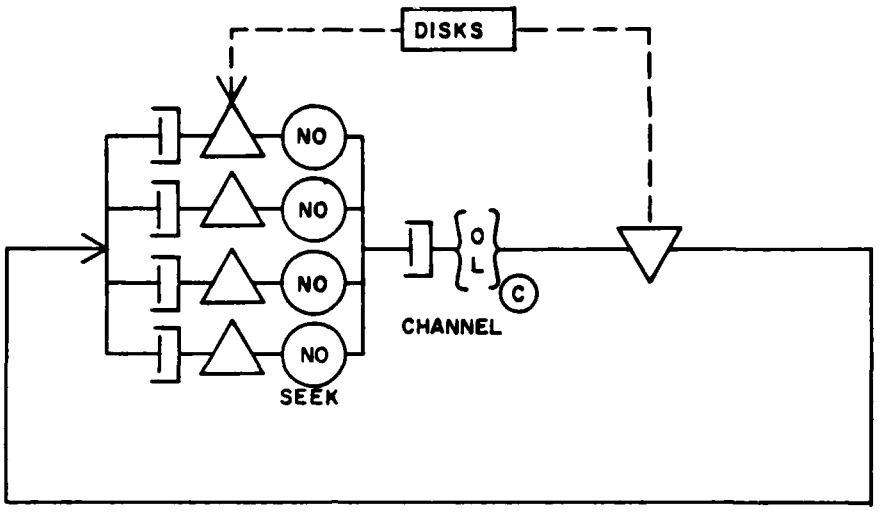


Figure 5-28. I/O Subsystem Evaluated in Isolation

customer can see different service time means and service time distributions at different queue populations. If the I/O subsystem is considered as a variable rate queue which has a FCFS service discipline and a fractional number of servers in parallel, a similar argument can be made as in the multiple server case. However, the possibility for queueing exists at the entry queues at service rates below the maximum. The mean residual wait will be different at each queue population because there is a different number of servers available, and possibly a different service time mean and distribution. The I/O subsystem can be evaluated in isolation for each possible population and an estimate obtained of the interdeparture time distribution and its associated mean residual life It is not known if the interdeparture mean residual life accurately reflects the mean residual wait an arriving customer sees at a variable rate queue when queueing occurs. However, it seems to be a natural extension of the multiple server case if it is measured for each population. The interdeparture time distribution mean can also be used to

determine the throughput. Since the residual life is a function of the first two moments of the distribution, the interdeparture mean and coefficient of variation might be good measures in determining the the true behavior of the variable rate queue.

The point behind the above analysis is that if a variable rate queue has an exponential server, such a server can be inserted into a product form network and the resulting network will be product form. If however the variable rate queue does not have an exponential server, then decomposition error may occur as the network may not be product form (recall the FCFS discipline is used). Hence, the service time differences from exponential in the variable rate queue are responsible for decomposition error, yet this difference does not seem to totally describe the degree of error possible in the final results.

Several parameters were thought to impact the degree of decomposition error present in the variable rate queue replacing the subnetwork. These parameters are very similar to the three described in characterizing representation error. They include:

1. The type of service time distribution of the variable rate queue. Since the interdeparture time distribution variation from exponential was identified as determining when differences in queue behavior were possible, it is believed that as the interdeparture time coefficient of variation departs from 1.0, the degree of decomposition error possible will increase. If the interdeparture time distribution is exponential then no decomposition error will be possible. Hence, the interdeparture time distribution difference from exponential totally characterize when decomposition error can exist, but not necessarily when it will occur.

In the above argument the implication was made that a subnetwork can be treated as 'black box', and by monitoring its output behavior for every possible input condition, the subnetwork can be accurately summarized. It is believed, and assumed, that the subnetwork can be exactly aggregrated into a variable rate queue with a set of population dependent throughputs as long as the service time distribution is accurately represented in the queue for each population. It is not believed that this queue can be inserted into a product form network and the resulting network still be product form.

The degree of impact of the additional parameters listed hinges on the presence of interdeparture time distribution differences from exponential.

2. The degree of queueing present at the variable rate queue. This is a function of many parameters within the subnetwork. Note that it might be possible to have queueing present at any subnetwork population greater than two. It is believed that the amount of time queueing occurs (equivalent to blocking at the subnetwork level) is proportional to the amount of decomposition error possible. When no queueing occurs the queue is effectively an infinite server queue and can be inserted into a product form network, preserving the product form, if the server has a rational Laplace transform. This quantity also influences the degree of representation error possible and as a result the two error are not independent

3. The utilization of the primary resources. This utilization is equivalent to the utilization of a variable rate node exactly representing the subnetwork, and is important because as it approaches 1.0, the possible throughput and waiting time differences an arriving customer 'sees' due to the interdeparture time distribution variation from exponential will decrease. Hence, the decomposition error possible will decrease as the

utilization approaches 1.0. Note that the primary resource utilization is not equivalent to the secondary utilization of the subnetwork.

It is believed that the above three parameters can explain most or all of the differences in performance parameters from the true values caused by decomposition error. The intent, as with representation error, is to show that the heuristic arguments presented seem reasonable.

In the previous two sections, two types of possible error were discussed: representation error and decomposition error. It was hypothesized that representation error is caused by the primary resource residency time distribution departures from exponential, and that decomposition error is caused by the service time distribution departures from exponential of the variable rate queue not considered by product form solution techniques. In the next section four network models are analyzed in order to illustrate the above heuristics and identify when the error might be more or less significant.

Additional Tests

Four of the initial test networks were selected for further analysis, models 1-4, 2-3, 3-4, and 5-1. The purpose of the analysis was to devise tests that would partition the potential sources of error in the analytic results into either representation error or decomposition error. The representation error was identified by analyzing the subnetworks with simultaneous resource possession in isolation using simulation for each feasible network population. The population dependent throughputs were then compared with those obtained analytically using the multi-entrance queue procedure. The impact on the final solution caused by representation error was estimated by using the 'exact' population dependent throughputs in a variable rate queue, replacing the subnetwork with simultaneous resource possession in the original network, and solving the resulting network (Figure 5-9) using product form methods. The parameters obtained by this procedure are called the multi-entrance queue parameters corrected for representation error.

Estimates of the difference in performance parameters caused by decomposition error were determined by comparing the parameters corrected for representation error with the exact parameters obtained by solving the original network using simulation. The partitioning of error is illustrated in Figure 5-29. The total error in the multi-entrance queue procedure is made up of that part due to representation and that part due to decomposition.

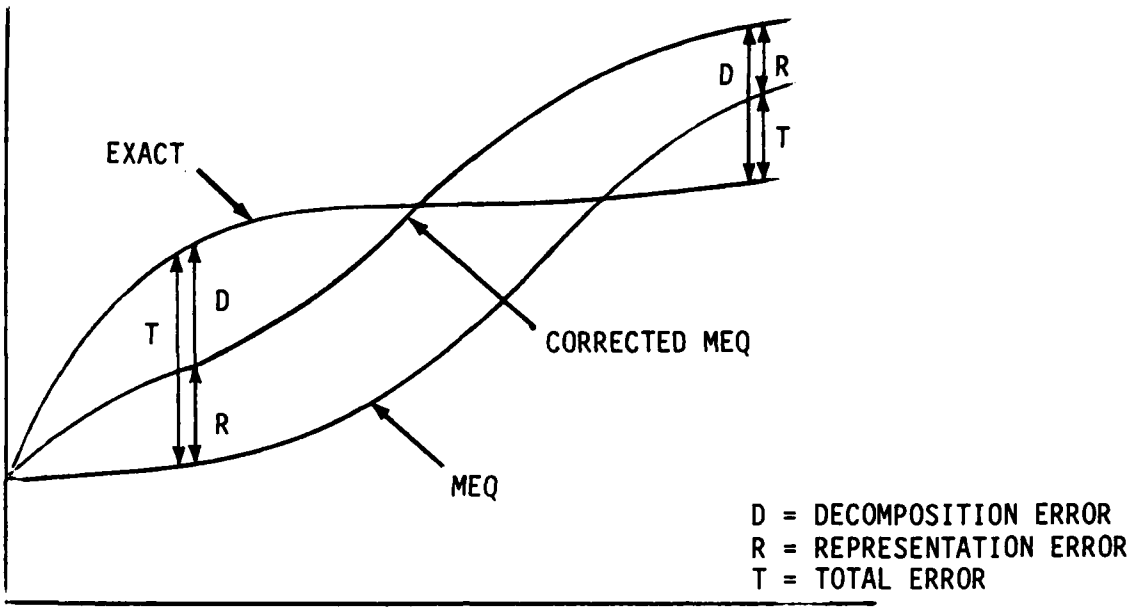


Figure 5-29. Error Partitioning

In the discussion that follows, an analysis of the representation error present will be given followed by an analysis of the decomposition error present in the additional test networks. Final conclusions will be made.

Representation Error. Because of the similarity of the results for model 1-4, model 2-3, and model 3-4, they will be discussed together. Model 5-1 will be discussed separately.

The throughput responses for the subnetworks of model 1-4, model 2-3, and model 3-2 were evaluated in isolation for both the multi-entrance queue method and simulation, and are given in Table 5-9 through Table 5-11. Along with the throughputs, the mean primary resource residency time and residency time coefficient of variation obtained by simulation are given. All of the analytically obtained throughputs of the subnetworks evaluated in isolation were significantly different from those obtained using simulation in model 1-4 and model 2-3, and all but one were significantly

TABLE 5-9

Model 1-4. Isolated Subnetwork Performance Parameters

		SIMULATED			
Network Population	Analytic Throughput	Secondary Util	Throughput	Mean Residency Time	Residency Time Coefficient of Variation
1	2.00	0.29	2.00	.500	.82
2	3.13	.47	3.21 (3.15-3.27)	.516 (.512-.522)	.81 (.78-.83)
3	3.84	.58	3.99 (3.91-4.07)	.527 (.522-.532)	.79 (.77-.82)
4	4.33	.65	4.51 (4.44-4.58)	.537 (.531-.542)	.78 (.76-.81)
5	4.68	.71	4.87 (4.79-4.95)	.543 (.540-.545)	.78 (.75-.81)

TABLE 5-10

Modél 2-3. Isolated Subnetwork Performance Parameters

		SIMULATED			
Network Population	Analytic Throughput	Secondary Util	Throughput	Mean Residency Time	Residency Time Coefficient of Variation
1	2.00	0.27	2.00	.500	.72
2	3.20	.45	3.32 (3.28-3.36)	.501 (.497-.506)	.73 (.70-.75)
3	3.97	.57	4.19 (4.13-4.25)	.506 (.500-.511)	.72 (.69-.75)
4	4.51	.65	4.78 (4.69-4.87)	.511 (.505-.517)	.72 (.69-.75)
5	4.90	.71	5.19 (5.05-5.33)	.516 (.509-.522)	.71 (.68-.75)

TABLE 5-11

Model 3-4. Isolated Subnetwork Performance Parameters

		SIMULATED			
Network Population	Analytic Throughput	Secondary Util	Throughput	Mean Residency Time	Residency Time Coefficient of Variation
1	2.00	0.25	2.00	.500	.82
2	3.20	.41	3.27 (3.20-3.35)	.502 (.488-.515)	.83 (.80-.87)
3	4.00	.52	4.13 (4.02-4.24)	.502 (.488-.515)	.83 (.80-.87)
4	4.55	.59	4.71 (4.58-4.85)	.505 (.491-.519)	.83 (.80-.86)
5	4.95	.64	5.13 (4.98-5.30)	.510 (.495-.524)	.82 (.79-.86)

different in model 3-4. Hence, representation error was present in these models. The throughput comparisons are illustrated graphically in Figure 5-30 for Model 1-4, Figure 5-31 for Model 2-3, and Figure 5-32 for Model 3-4. In each case when the throughputs were significantly different, the analytic throughputs yielded lower values than the simulated throughputs.

Recall that is was said the the primary resource residency time distribution must be different from exponential for representation error to exist. In each model, the residency time coefficient of variation was significantly different from 1.0 for every population. For example, in model 1-4, the residency time coefficient of variation varied from .82, for a subnetwork population of one, to .78, for a network population of five. The differences in these values were not significant, however, and were not significant in model 2-3 or model 3-4 either. Hence even though it was possible for different residency time distributions to exist, the differences if any could not be detected with the accuracy of the simulations used. The residency time coefficients of variation are illustrated graphically in Figure 5-33 for model 1-4, Figure 5-34 for model 2-3, and Figure 5-35 for model 3-4. Because the residency time coefficient of variation was less than 1.0, it was expected that the throughputs obtained by simulation would be higher than the throughputs obtained analytically. Hence, it was predicted that an arriving customer would have a slightly lower residual wait for the customer in service to finish. This would result in a slightly lower cycle time and consequently a slightly higher throughput.

Another stated prerequisite for representation error was that queueing must exist at the primary resource allocation centers, which it did. The degree of queueing present was measured as the proportion of the subnetwork

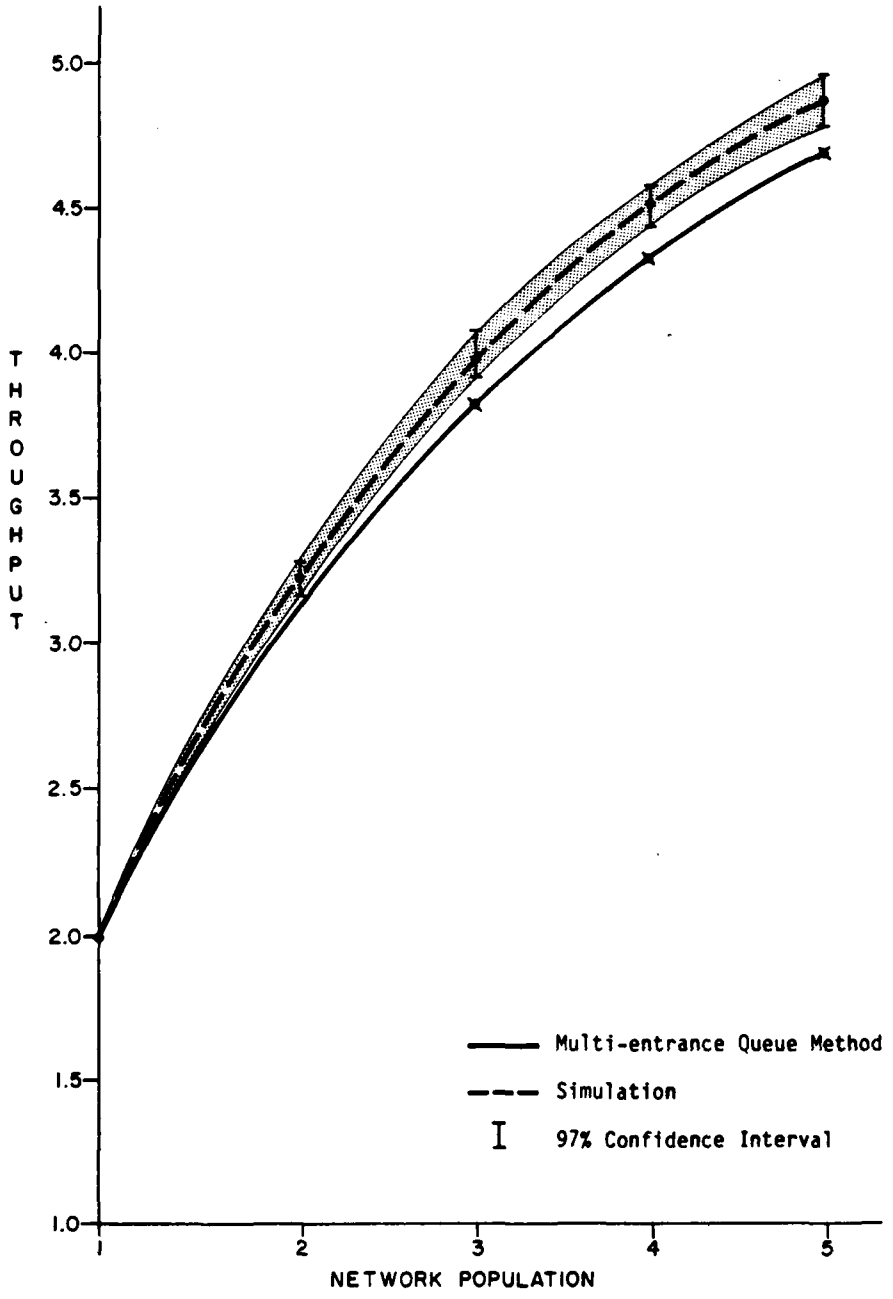


Figure 5-30. Model 1-4 Throughputs of Subnetwork Evaluated in Isolation

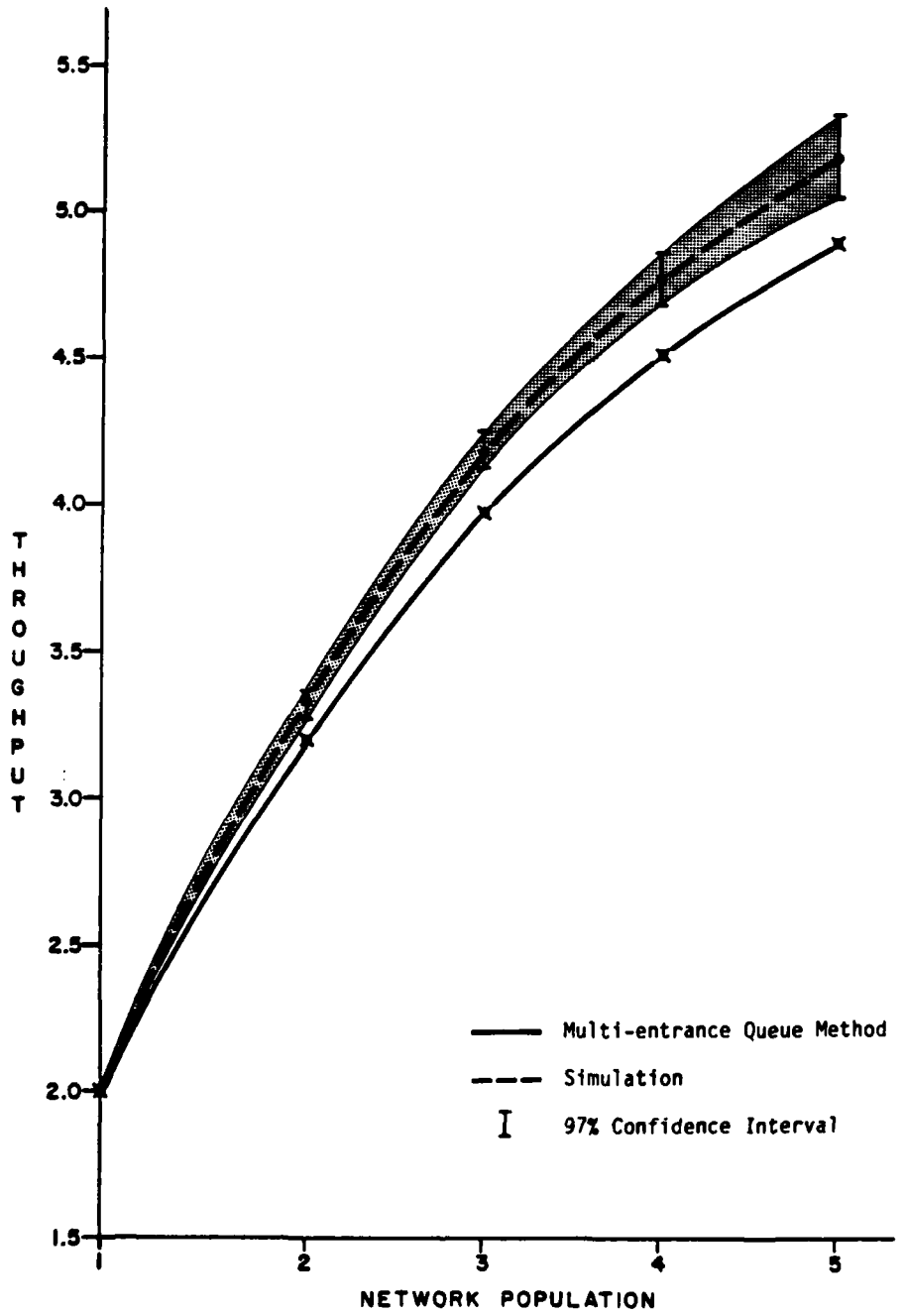


Figure 5-31. Model 2-3 Throughputs of Subnetwork Evaluated in Isolation

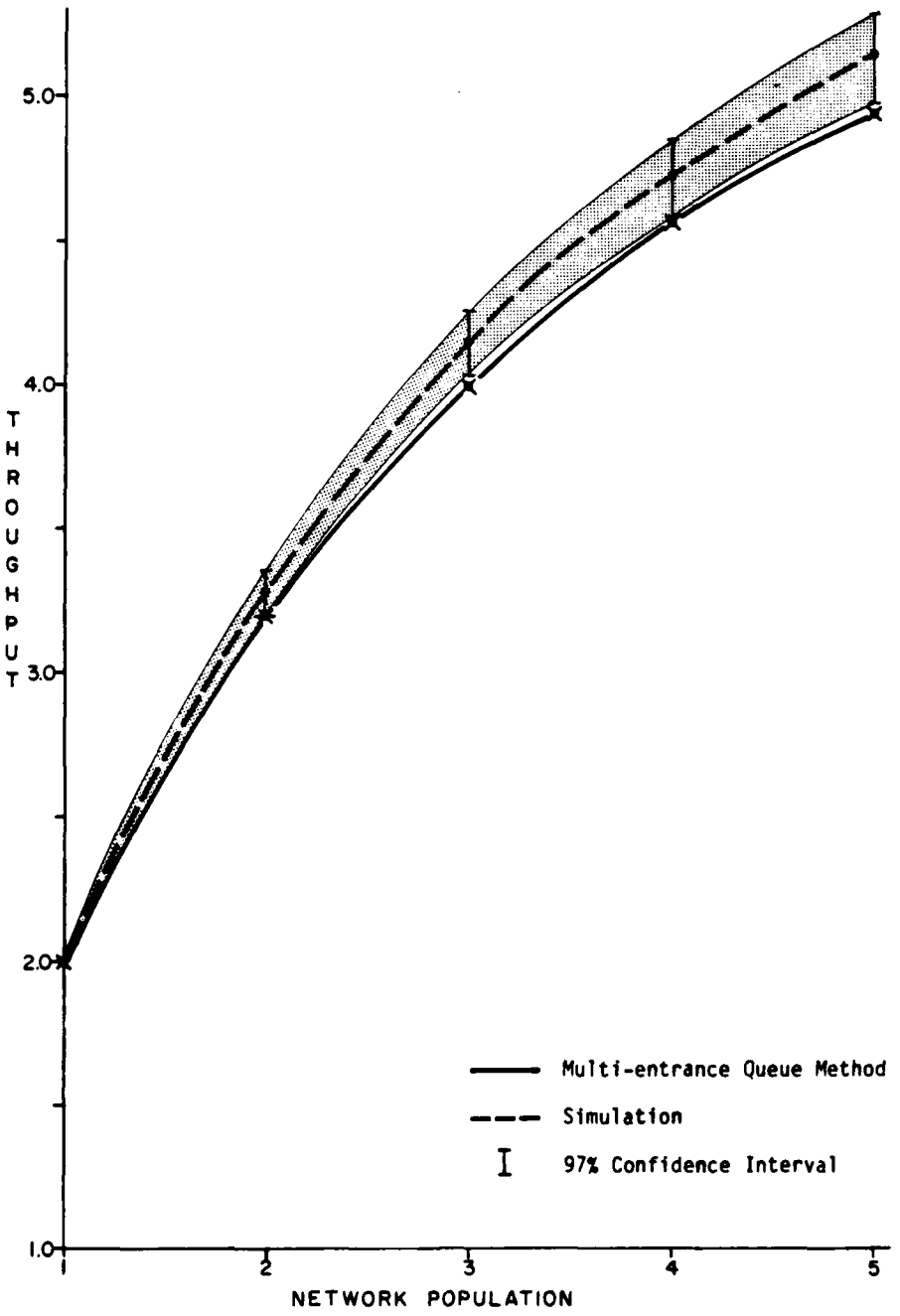


Figure 5-32. Model 3-4 Throughputs of Subnetwork Evaluated in Isolation

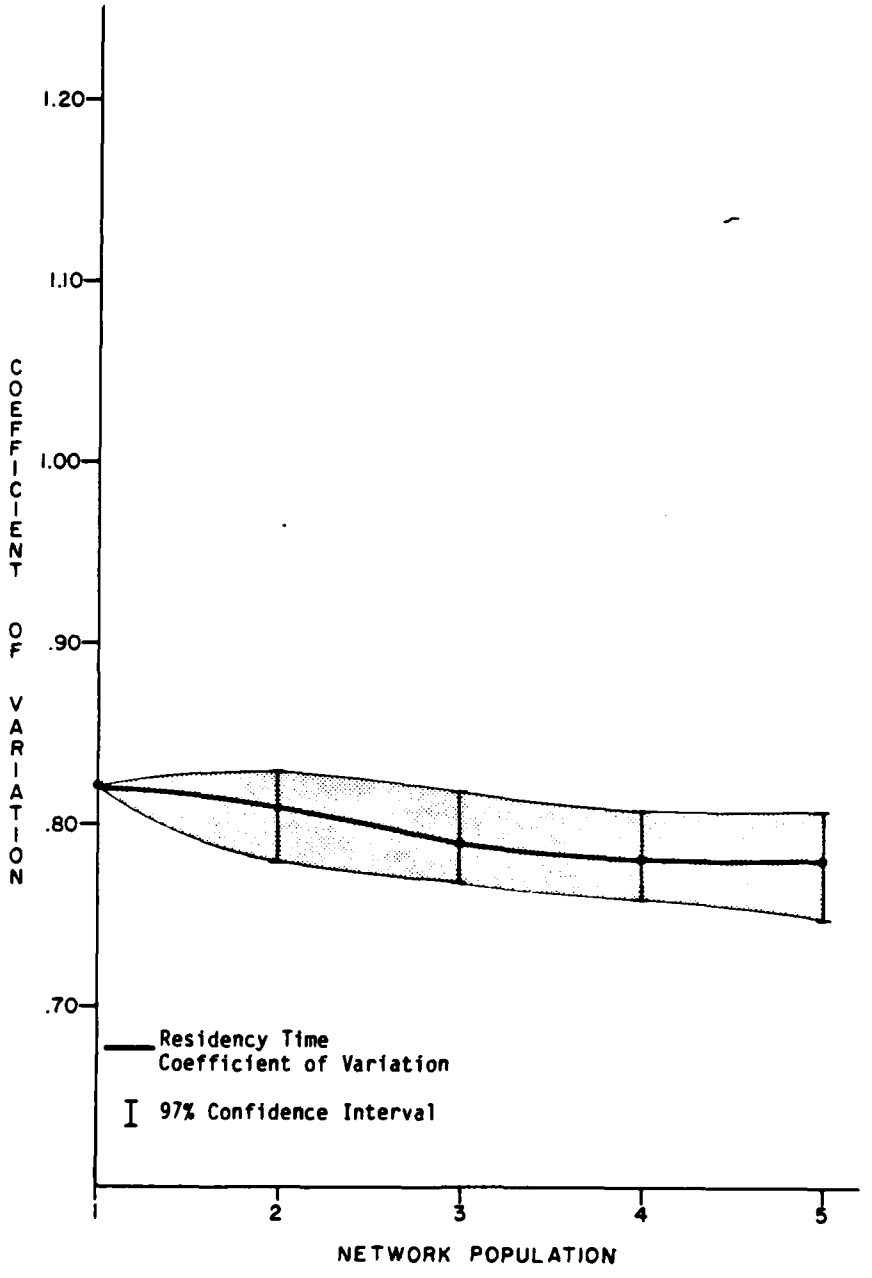


Figure 5-33. Model 1-4 Residency Time Coefficients of Variation With the Subnetwork Evaluated in Isolation

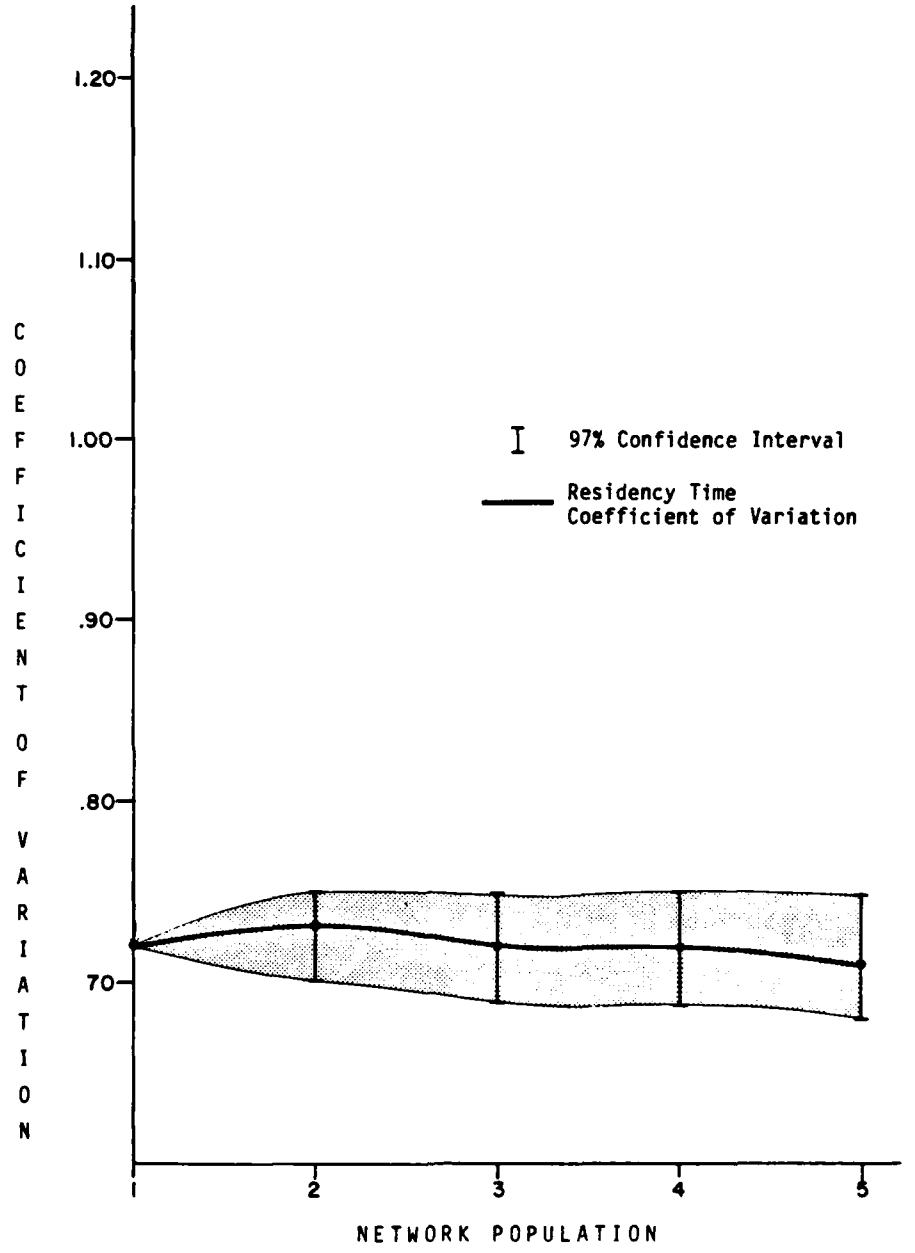


Figure 5-34. Model 2-3 Residency Time Coefficients of Variation With the Subnetwork Evaluated in Isolation

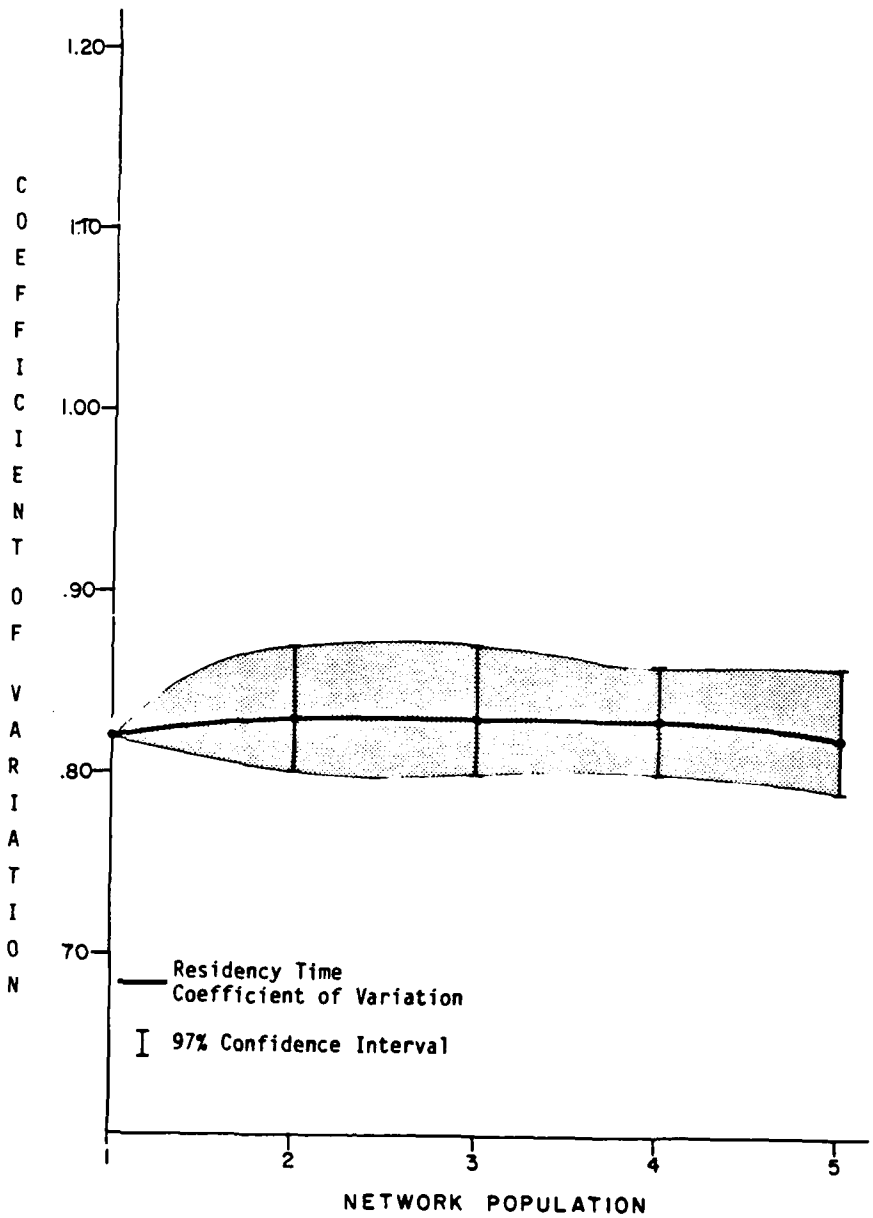


Figure 5-35. Model 3-4 Residency Time Coefficients of Variation With the Subnetwork Evaluated in Isolation

cycle time spent queued. These values, obtained by simulation, are given in Table 5-16 for model 1-4, Table 5-17 for model 2-3, and Table 5-18 for model 3-2, and are nonzero for all populations greater than one.

The secondary utilizations of the three models were all less than 1.0 and are given in Table 5-9, for model 1-4, Table 5-10 for model 2-3, and Table 5-11 for model 3-4. The values ranged between .29 and .70 for model 1-4, between .27 and .71 for model 2-3, and between .25 and .64 for model 3-4. Since all population dependent secondary utilizations were less than 1.0, it was expected that the residency time distribution departures from exponential could cause differences between the population dependent multi-entrance queue throughputs and the population dependent throughputs obtained by simulation when the subnetworks were evaluated in isolation.

The impact of representation error on the final model throughputs is illustrated in Table 5-12 by comparing the multi-entrance queue (MEQ) throughputs to the multi-entrance queue throughputs corrected for representation error. The multi-entrance queue throughputs corrected for representation error were obtained by using the simulation derived population dependent throughputs of the subnetwork evaluated in isolation in the final variable rate queue, as opposed to the population dependent throughputs obtained by the multi-entrance queue model. Notice that the difference accounts for almost all of the error present in the multi-entrance queue procedure. For example the final throughput obtained by the multi-entrance queue procedure for model 1-4 was 4.37, whereas the throughput corrected for representation error was 4.53. In this case the corrected throughput was within the confidence interval for the 'exact' throughput. This was true for model 2-3 and model 3-4, also.

Table 5-12

Throughput Comparisons

Analytic Results			Simulation		
Model	MEQ Throughput	Corrected MEQ Throughput	Primary Utilization	Throughput	97% Confidence Interval
5-1	1.59	1.59	0.81	1.59	(1.56-1.63)
1-4	4.37	4.53	.94	4.55	(4.49-4.62)
2-3	4.54	4.77	.93	4.80	(4.70-4.90)
3-4	4.61	4.72	.94	4.79	(4.72-4.98)

The subnetwork of model 5-1 was evaluated in isolation for both the multi-entrance queue method and simulation, and the throughputs are given in Table 5-13, along with the mean primary resource residency time and residency time coefficient of variation obtained using simulation. For each population, the throughput obtained by the multi-entrance queue procedure was within the confidence intervals for the throughput obtained by simulation. In fact, in most cases the multi-entrance queue obtained throughputs were exactly, or almost exactly equal to the throughput means obtained by simulation. This is illustrated graphically in Figure 5-36. Hence, if representation error was present in this model it was minimal.

Even though the representation error present was small or nonexistent, the residency time mean and distribution varied considerably between different subnetwork populations. For example, the residency time mean ranged from 1.00 for a population of one and two, to 1.49 for a population of six. The residency time coefficient of variation also varied from 1.00 for a population of one, to .79 for a population of six. The differences in the residency time coefficient of variation were significantly different between most populations and are illustrated graphically in Figure 5-37. As a result, it was possible that representation error could occur because the residency time distribution varied from exponential for most populations.

Table 5-13

Model 5-1. Isolated Subnetwork Performance Parameters

		SIMULATED			
Network Population	Analytic Throughput	Secondary Util	Throughput	Mean Residency Time	Residency time Coefficient of Variation
1	1.00	0.50	1.00	1.00	1.00
2	1.60	.80	1.59 (1.55-1.64)	1.00 (0.96-1.04)	1.01 (0.96-1.05)
3	1.82	.91	1.81 (1.75-1.89)	1.15 (1.11-1.20)	.93 (0.90-0.96)
4	1.90	.95	1.89 (1.86-1.92)	1.30 (1.28-1.32)	.86 (0.84-0.89)
5	1.94	.97	1.94 (1.86-2.03)	1.40 (1.35-1.46)	.82 (0.80-0.83)
6	1.96	.98	1.96 (1.88-2.05)	1.49 (1.45-1.53)	.79 (0.76-0.81)

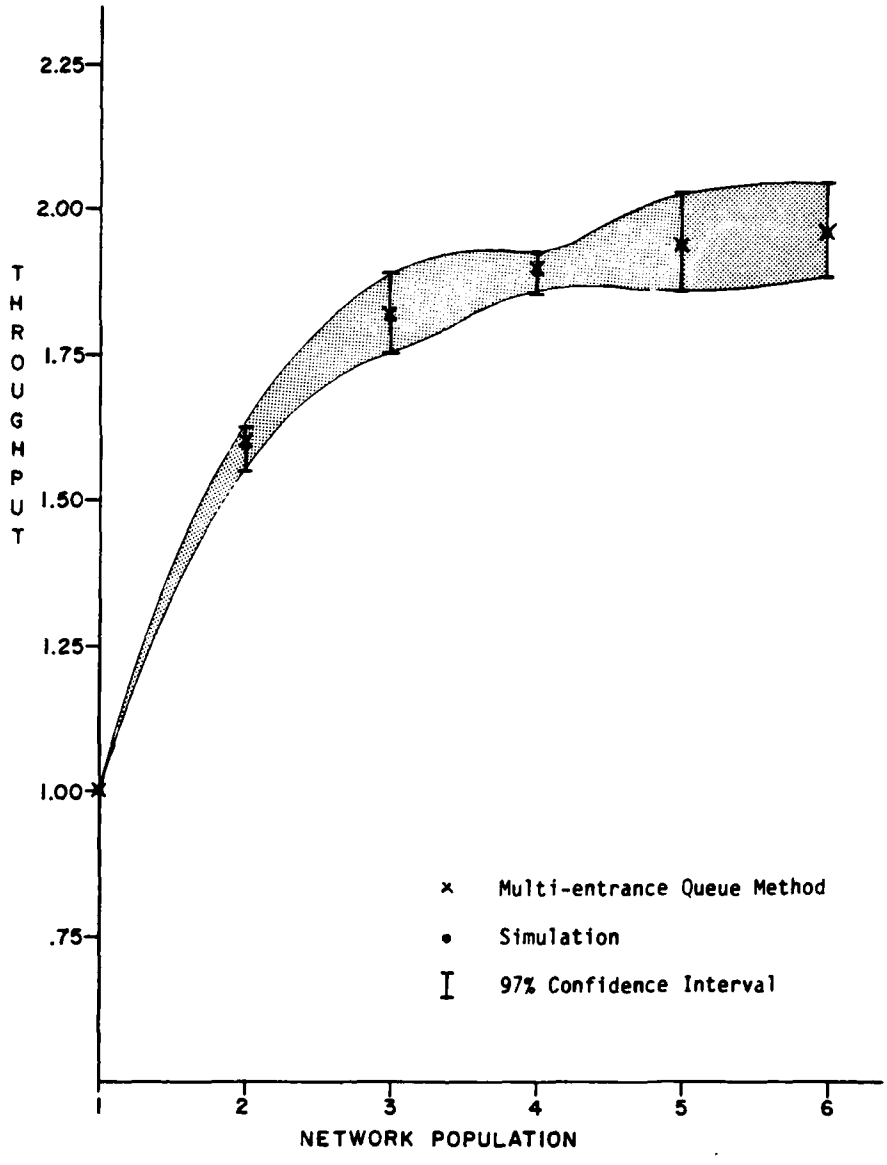


Figure 5-36. Model 5-1 Throughputs of Subnetwork Evaluated in Isolation

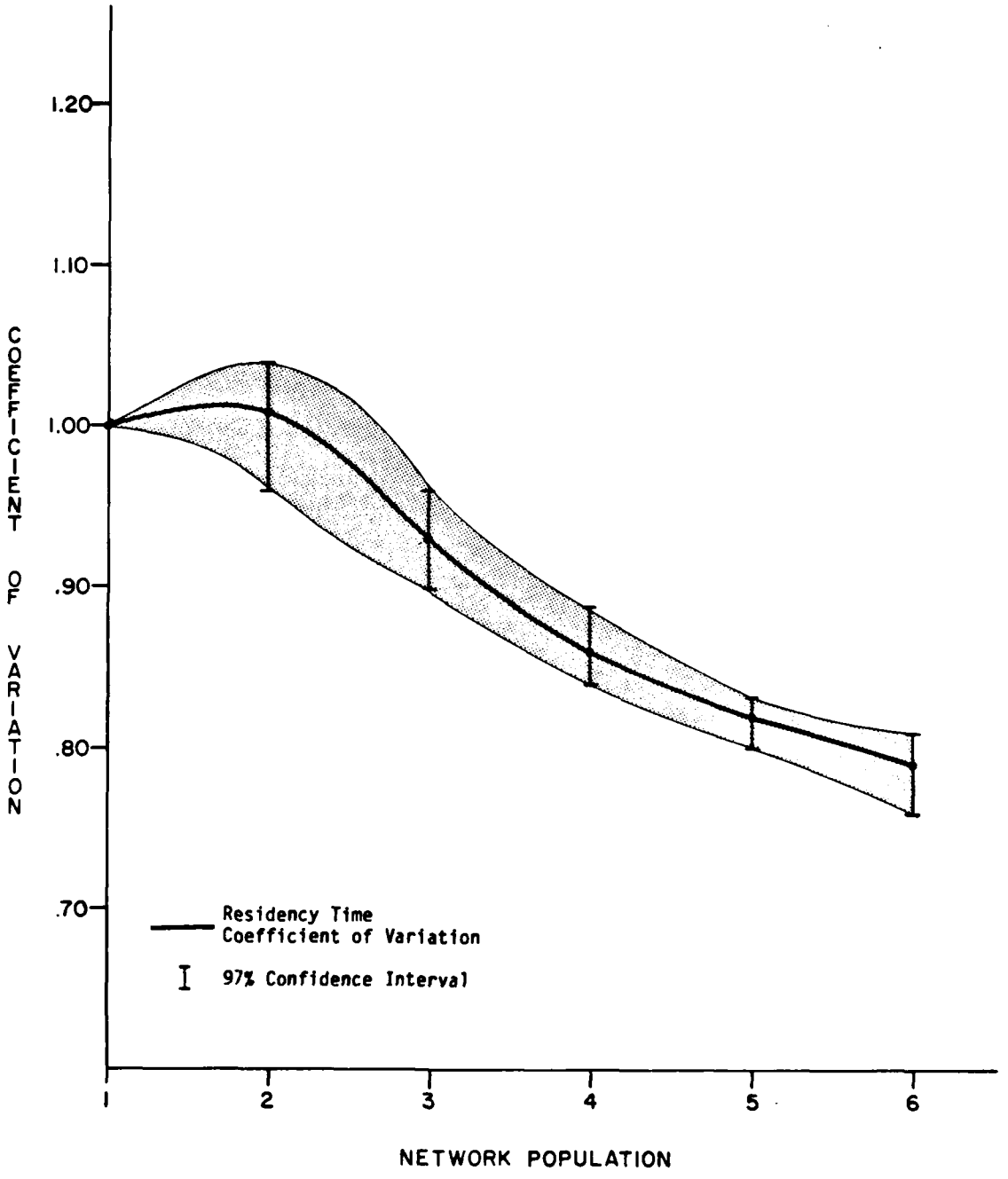


Figure 5-37. Model 5-1 Residency Time Coefficients of Variation With the Subnetwork Evaluated in Isolation

Queueing was also present at the primary resource allocation queues for all populations greater than one. The queueing ratios are listed in Table 5-12. Since queueing was present, it was possible for representation error to occur.

It was believed that the representation error present in model 5-1 was minimal because the secondary utilization at most subnetwork populations was high. The secondary utilizations ranged from .50 at a subnetwork population of one, to .98 at a subnetwork population of six. Recall that it was stated that as the secondary utilization approached 1.0, the amount of throughput difference due to residency time distribution departures from exponential would approach zero. Since the attained secondary utilization was very close to 1.0 for the higher subnetwork populations, it was expected that the throughput departures would be minimal even though the residency time coefficient of variation was significantly different from 1.0 at these populations. At the lower subnetwork populations, the secondary utilization was not high, but the residency time coefficient of variation was close to 1.0.

The multi-entrance queue procedure did not obtain estimates within the simulation confidence intervals for all performance parameters of interest. As an example the mean number of busy primary resources for each network population obtained by simulation and the multi-entrance queue procedure were compared. These values are listed in Table 5-14. The multi-entrance queue obtained values were significantly lower than those obtained by simulation for populations where the residency time coefficient of variation was greatly different from 1.0. For example, at a customer population of six, the residency time coefficient of variation was .79, and the mean number of primary resources obtained by the multi-entrance queue procedure was 2.84, whereas the simulated value was significantly different at 2.92.

However, at a customer population of two, the residency time coefficient of variation was not significantly different from 1.0, and the mean number of busy primary resources obtained by both methods was 2.09.

Table 5-14

Mean Number of Busy Primary Resources

Network Population	Mean Number of Busy Primary Resources		
	MEQ	Simulation	97% Confidence Interval
1	1.00	--	--
2	1.60	1.60	(1.59-1.61)
3	2.09	2.09	(2.07-2.11)
4	2.43	2.46	(2.41-2.51)
5	2.67	2.72	(2.67-2.78)
6	2.84	2.92	(2.87-2.97)

As there was minimal representation error present in the multi-entrance queue throughputs, it was expected that the throughput of model 5-1 corrected for representation error would be identical, or almost identical to the uncorrected throughput. In fact, in this example the two throughputs were the same at 1.59. However, there was error in some of the other subnetwork related performance parameters, such as the mean number of busy primary resources (recall that this is the value estimated by the primary contention model of the ECM procedure).

The above examples illustrate some of the heuristic concepts identified as impacting the degree of representation error present in a network solved by the multi-entrance queue procedure. No examples were given to show that when queueing is not present at the primary resource allocation centers, no representation error is possible. However, in this case the network reduces to a sequence of queues without population constraints on the subnetwork. Hence blocking does not occur, and therefore the network

will be product form if it otherwise would have been. The examples are not intended to provide significant empirical evidence proving or quantifying the relationships. In the following paragraphs the degree of decomposition error present in the four test models will be discussed.

Decomposition Error. The amount of decomposition error present in the throughputs obtained using the multi-entrance queue procedure for the additional models was determined by comparing the throughputs corrected for representation error to the actual throughputs obtained by simulation. These values are given in Table 5-12. The throughputs corrected for representation error for model 5-1 agreed exactly with the mean of the actual simulated throughputs. The throughputs corrected for representation error for model 1-4, model 2-3, and model 3-4, were all slightly lower than the actual throughputs, however the differences were not significant. It was not known if decomposition was present, or just not measurable with the accuracy of the simulations used.

In the previous section, the differences in the interdeparture time distribution from exponential were said to determine when decomposition error was possible in the variable rate queue replacing the subnetwork with simultaneous resource possession. The interdeparture time mean and coefficient of variation was measured using simulation for each feasible population of the subnetwork for the four additional models. These values are given in Table 5-15 for model 5-1, Table 5-16 for model 1-4, Table 5-17 for model 2-3, and Table 5-18 for model 3-4, and are graphically displayed in Figures 5-38 through 5-41.

The interdeparture coefficient of variation for model 5-1 varied from 1.12, which was significantly different from 1.0 at a subnetwork population of two, to 1.02, which was not significantly different from 1.0 at a network population of six. Only the interdeparture coefficients of variation at subnetwork populations of one through three were significantly

different from 1.0. Queueing occurred at all populations over one and the primary utilization was only .81 so it was felt that decomposition error was possible, although it was believed that the degree of decomposition error possible was small for model 5-1, and would decrease as the probability of low populations in the subnetwork decreased. The throughput corrected for representation error and the actual throughput were both equal in the final results at 1.59. It was not believed that the network with the variable rate queue in place of the subnetwork was an exact representation of the original network, however.

The interdeparture coefficients of variation for model 1-4 and model 2-3 were significantly different from 1.0 for every subnetwork population. Model 1-4 had values ranging from .82 at a subnetwork population of one, and .93 at a network population of two, to .95 at a subnetwork population of five. None of the values were significantly different from each other where queueing occurred (populations two to five). Model 2-3 had interdeparture coefficients of variation that ranged from .72 at a network population of one, and .86 at a network population of two, to .89 at a network population of five. Similar to model 1-4, none of the values were significantly different from each other where queueing occurred. However, it is believed that with more accurate simulations, significant differences would occur in the interdeparture coefficients of variation between different populations for both models. In both models queueing occurred at the entry queues at all populations over one and the primary utilizations were .94 for model 1-4, and .93 for model 2-3. Hence, because the interdeparture coefficients of variation were significantly different from 1.0, it was believed that decomposition error could occur. The differences between the throughputs corrected for representation error and the actual throughputs of the original network were not significantly different for either model 1-4 or model 2-3. However, in both cases the actual throughputs were slightly higher than the corrected throughputs. It is not known

Table 5-15

Model 5-1. Isolated Subnetwork Performance Parameters (Cont)

Network Population	Interdeparture Coefficient of Variation	Proportion Of Service Time Spent Queued	Throughput Percent Error
1	1.00	0.00	---
2	1.12 (1.08-1.16)	.20	0.6
3	1.08 (1.05-1.11)	.31	0.6
4	1.04 (1.00-1.09)	.39	0.6
5	1.02 (.99-1.05)	.46	0.0
6	1.02 (.99-1.05)	.51	0.0

Table 5-16

Model 1-4. Isolated Subnetwork Performance Parameter (Cont)

Network Population	Interdeparture Coefficient of Variation	Proportion Of Service Time Spent Queued	Throughput Percent Error
1	.82	0.00	---
2	.93 (.91-.96)	.17	-2.5
3	.94 (.91-.97)	.30	-3.8
4	.94 (.93-.96)	.39	-4.0
5	.95 (.92-.98)	.47	-3.9

Table 5-17

Model 2-3. Isolated Subnetwork Performance Parameters (Cont)

Network Population	Interdeparture Coefficient of Variation	Proportion Of Service Time Spent Queued	Throughput Percent Error
1	.72	0.00	---
2	.86 (.84-.87)	.17	-3.6
3	.88 (.85-.90)	.29	-5.3
4	.89 (.86-.91)	.39	-5.6
5	.89 (.87-.91)	.46	-5.6

Table 5-18

Model 3-4. Isolated Subnetwork Performance Paramters (Cont)

Network Population	Interdeparture Coefficient of Variation	Proportion Of Service Time Spent Queued	Throughput Percent Error
1	.82	0.00	---
2	.95 (.92- .98)	.18	-2.1
3	.97 (.92-1.01)	.31	-3.1
4	.96 (.90-1.02)	.41	-3.4
5	.95 (.92- .99)	.48	-3.5

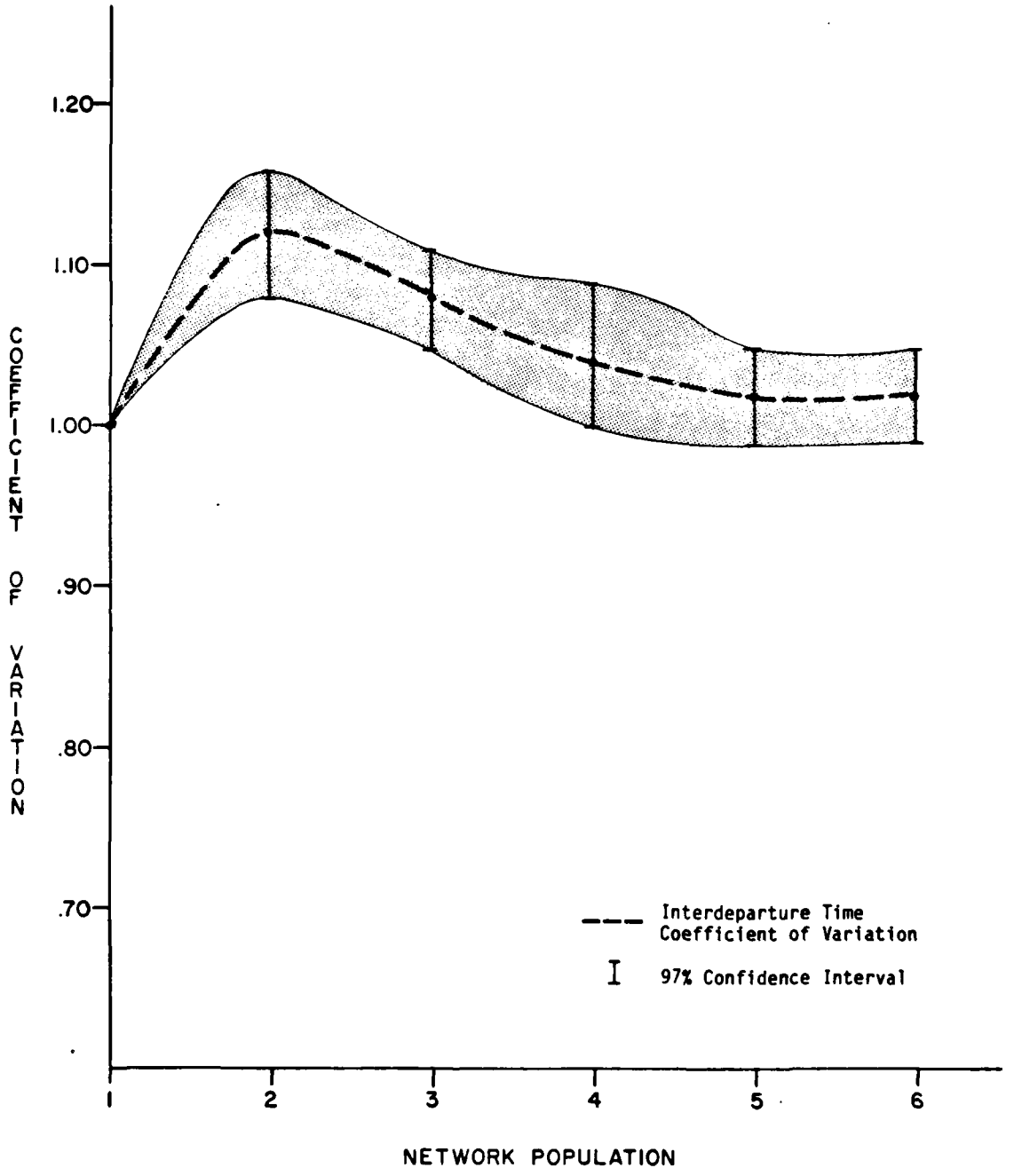


Figure 5-38. Model 5-1 Interdeparture Time Coefficients of Variation With the Subnetwork Evaluated in Isolation

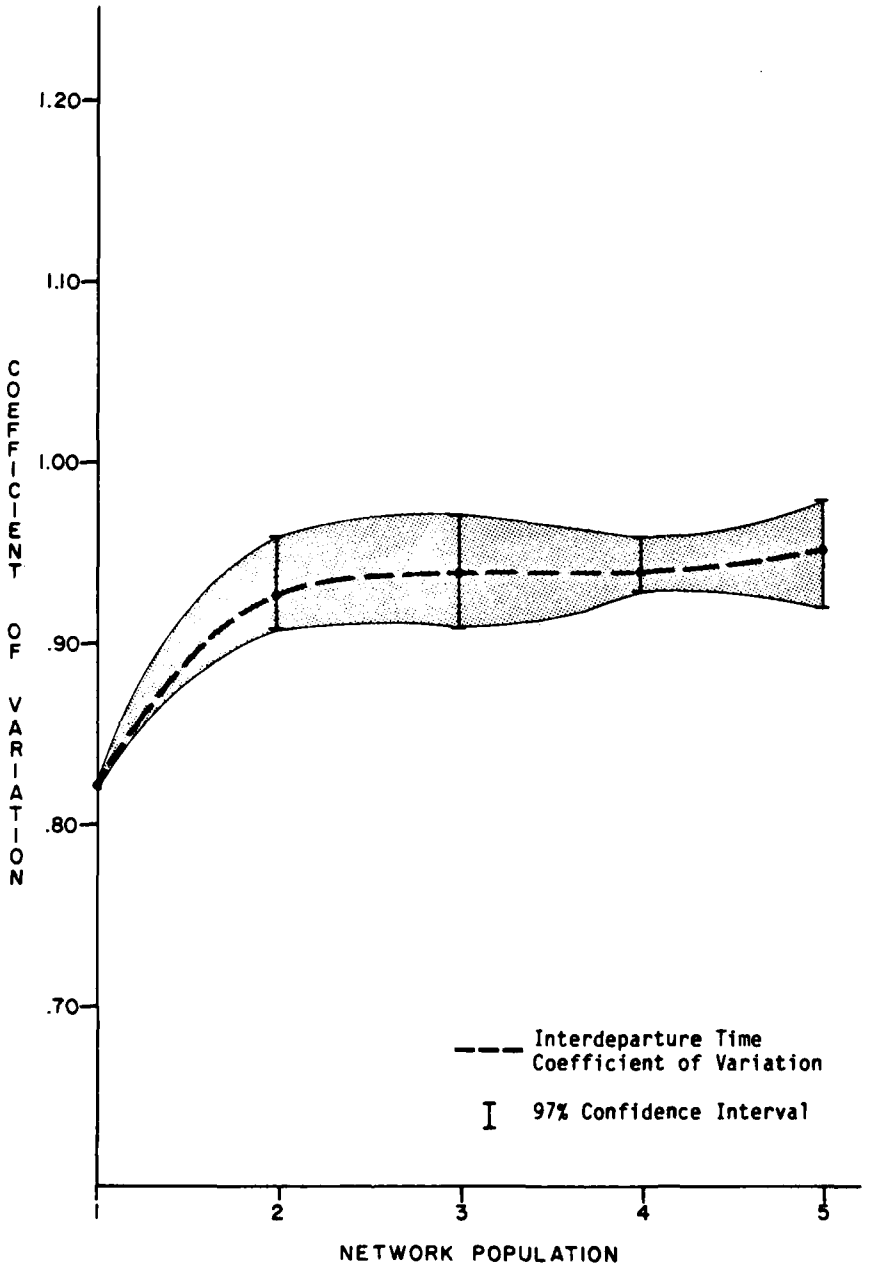


Figure 5-39. Model 1-4 Interdeparture Time Coefficients of Variation With the Subnetwork Evaluated in Isolation

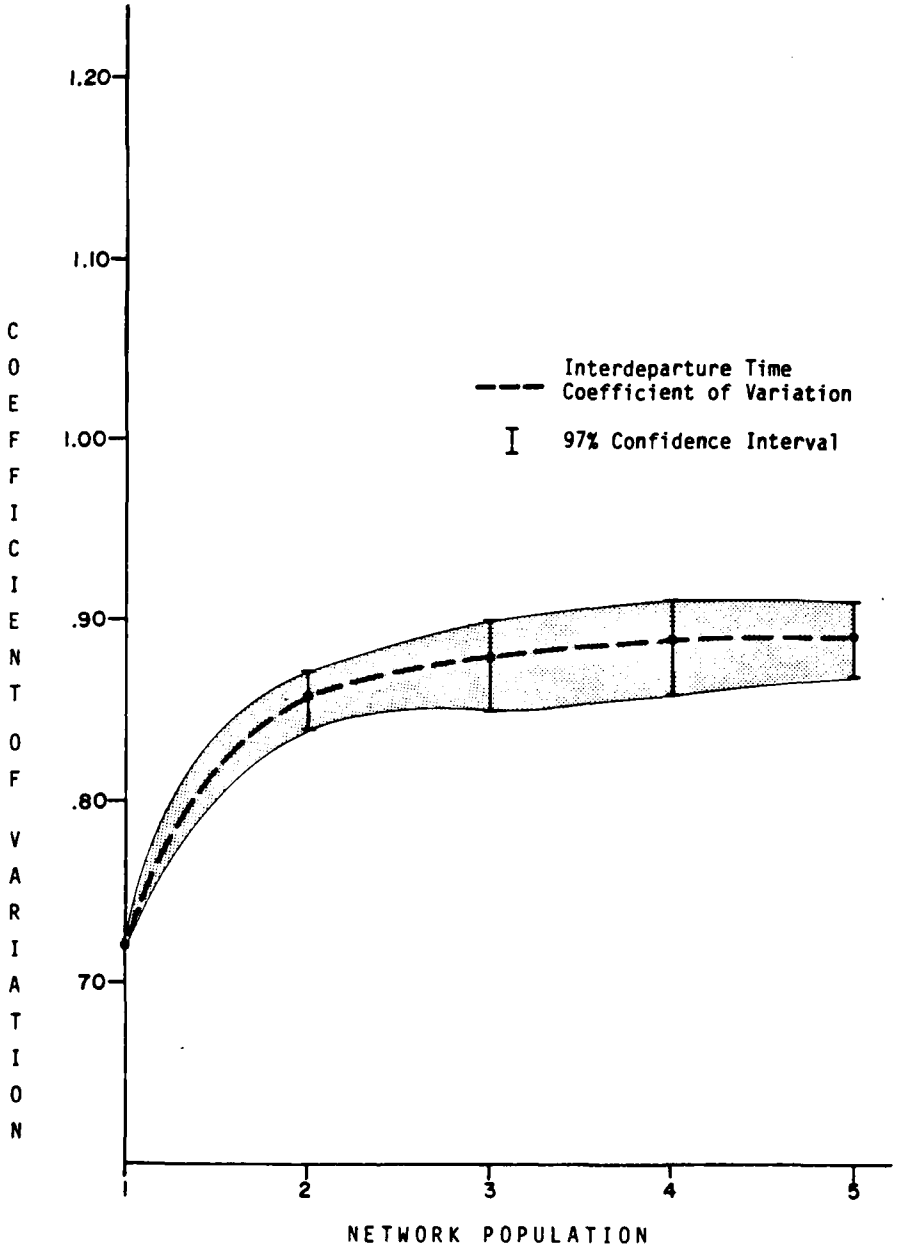


Figure 5-40. Model 2-3 Interdeparture Time Coefficient of Variation With the Subnetwork Evaluated in Isolation

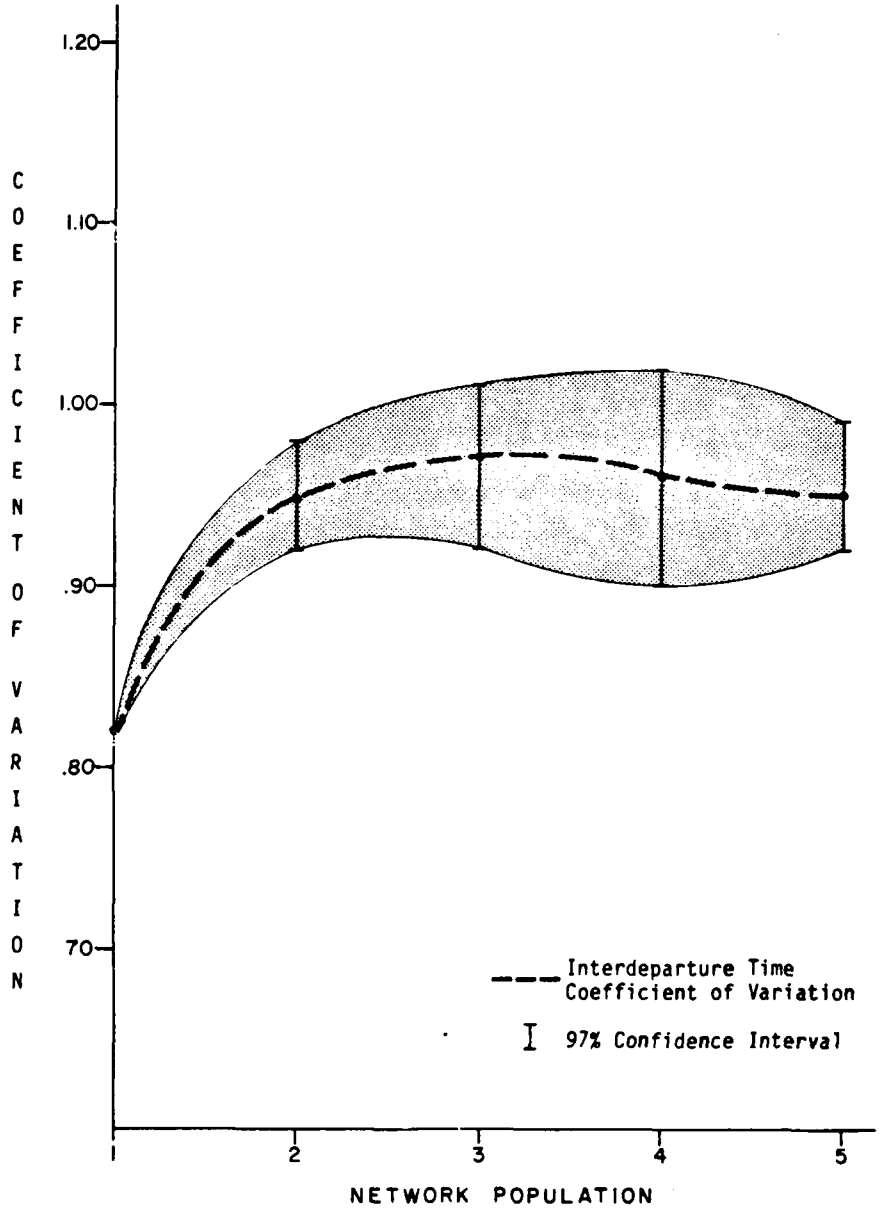


Figure 5-41. Model 3-4 Interdeparture Time Coefficients of Variation With the Subnetwork Evaluated in Isolation

if decomposition error was present and not detectable with the accuracy of the simulations or if it was not present, although it was felt that more accurate simulations would find significant differences between the throughputs.

Model 3-4 had interdeparture coefficients of variation that ranged from .82 at a subnetwork population of one, and .95 at a subnetwork population of two, to .97 at a network population of three. The confidence intervals for this model were looser than the other three models, and two of the interdepartue coefficients of variation, at populations three and four, were not significantly different from 1.0. However, it was believed that this was due to the higher inaccuracy in the simulation. Queueing occurred at all subnetwork populations over one, and the primary utilization was .94, hence it was felt that decomposition error was possible. The throughput corrected for representation error was 4.72 as compared to the actual throughput of 4.79, although the difference was not significant.

The above analysis attempts to illustrate some of the heuristic concepts identified as impacting the degree of decomposition error possible in the networks solved by the multi-entrance queue procedure. As none of the comparisons were significant, it can not be said that decomposition error was present in the above models. However, it is believed that there was decomposition error present in the results, which was on the order of 1.0 percent. It was too costly to run the simulations long enough to obtain the required accuracy to determine if the differences were significant.

Conclusions. The analysis in the previous two sections is not sufficient to empiricially support the heuristics presented or even to show their existence. Such a task is beyond the scope of this document. What

it is intended to show is where further investigation should be focused. It is believed that with sufficient empirical studies and analysis that the some of the sources of error can be quantified.

The multi-entrance queue was analyzed for error by partitioning the error into two possibilities, representation error and decomposition error. This partitioning required the assumption that the behavior of a non-product form subnetwork could be summarized exactly using a set of population dependent service times and service time distributions. Representation error was caused by the assumptions placed on the residency time distribution of the multi-entrance queue model. The residency time distribution differences from exponential were identified as a necessary condition for representation error to occur. Additionally, the degree of queueing at the primary resource entry queues and the secondary utilization were identified as influencing the magnitude of the representation error. Decomposition error was caused by the restrictions placed on the interdeparure time distribution of the variable rate queue replacing the subnetwork with simultaneous resource possession. Interdeparture time differences from exponential were identified as a necessary condition for decomposition error. The degree of queueing present at the primary resource center and the primary utilization were identified as influencing the magnitude of the decomposition error.

In the following sections a the multi-entrance queue procedure will be compared to the method of ECM and the method of surrogates.

Comparison of the ECM method and Multi-entrance Queue Procedures

There are two factors that separate the ECM procedure form the multi-entrance queue procedure. These are listed as follows:

1. The ECM procedure can be applied in situations where the multi-entrance queue model is not applicable. These include cases where all nonoverlapped service requirement are not equal and cases where type two simultaneous resource possession is present but the entry queue may allocate more than one primary resource.

2. In those cases where the multi-entrance queue is applicable, the ECM procedure provides an approximate solution to the model, Figure 4-4, solved exactly by the multi-entrance queue procedure.

As no comparison can be made where the multi-entrance queue procedure is not applicable this section will cover two topics.

1. The departures of the ECM solution from the multi-entrance queue model solution, and

2. The relative merits of both procedures.

ECM Departures from the Multi-entrance Queue. The basic assumptions that motivate the ECM procedure are:

1. That external contention present in a subnetwork is only a function of the relative throughputs at each entry queue and the ratio of differences between the mean service demand at each entry queue (In cases

of interest here the mean service demand at each entry queue must be the same). The size of the service demand is not involved just the ratio of differences.

2. The external contention can be represented as an effective population allowed passage through the entry queues of a subnetwork for a particular subnetwork population.

Since external contention can be represented as a mean population allowed passage, the service rate of the subnetwork can be approximated for each possible subnetwork population by using the augmented secondary subsystem throughput with the mean population allowed passage. This is illustrated in Figure 5-37. However, the mean populations allowed entry into the augmented secondary subsystem may be fractional. Hence interpolation is required to estimate the throughputs, and error is introduced as a result.

If linear interpolation is used the error will be due to the difference in the throughput response from linear between the two points of interpolation. This error is usually small. However, if the throughput response of the augmented secondary subsystem has abrupt changes in slope at a particular population then additional error can be introduced if the estimates of external contention for any population exactly equal that number (no interpolation is required). This is best illustrated by an example.

Consider Model 5-5, which is the loosely coupled multiprocessor system evaluated with ten processors (customers). A set of ten population dependent throughputs, $U_f(n)$, for $n=1,\ldots,10$, are needed for the variable rate queue replacing the subnetwork. These values are obtained by equation 3-8. The throughputs of the augmented secondary subsystem are required and

are listed in Table 3-6. They are also plotted in Figure 5-42. Notice that there is an abrupt change in the slope of the throughput response when an augmented secondary subsystem population of two is reached. This is because the secondary utilization reaches 1.0 at this population (Recall that the augmented secondary subsystem for this model is a simple two server queue evaluated in isolation). The estimates of external contention are measured by the primary contention model for each subnetwork population, and are interpreted as the mean population allowed entry into the augmented secondary subsystem ($U_p(n)$, for n=1,...,10). These values are listed in Table 5-19. Notice that at a subnetwork

Table 5-19

Model 5-5. Estimates of External Contention

Subnetwork Population	Mean Passage Population [$U_p(n)$]
1	1.000
2	1.600
3	2.000
4	2.286
5	2.500
6	2.667
7	2.800
8	2.909
9	3.000
10	3.077

population of three, the mean passage population is 2.0, which exactly equals the population at the point of the augmented secondary subsystem throughput slope discontinuity. The impact of this occurrence can be seen by comparing the isolated subnetwork throughputs obtained by the ECM

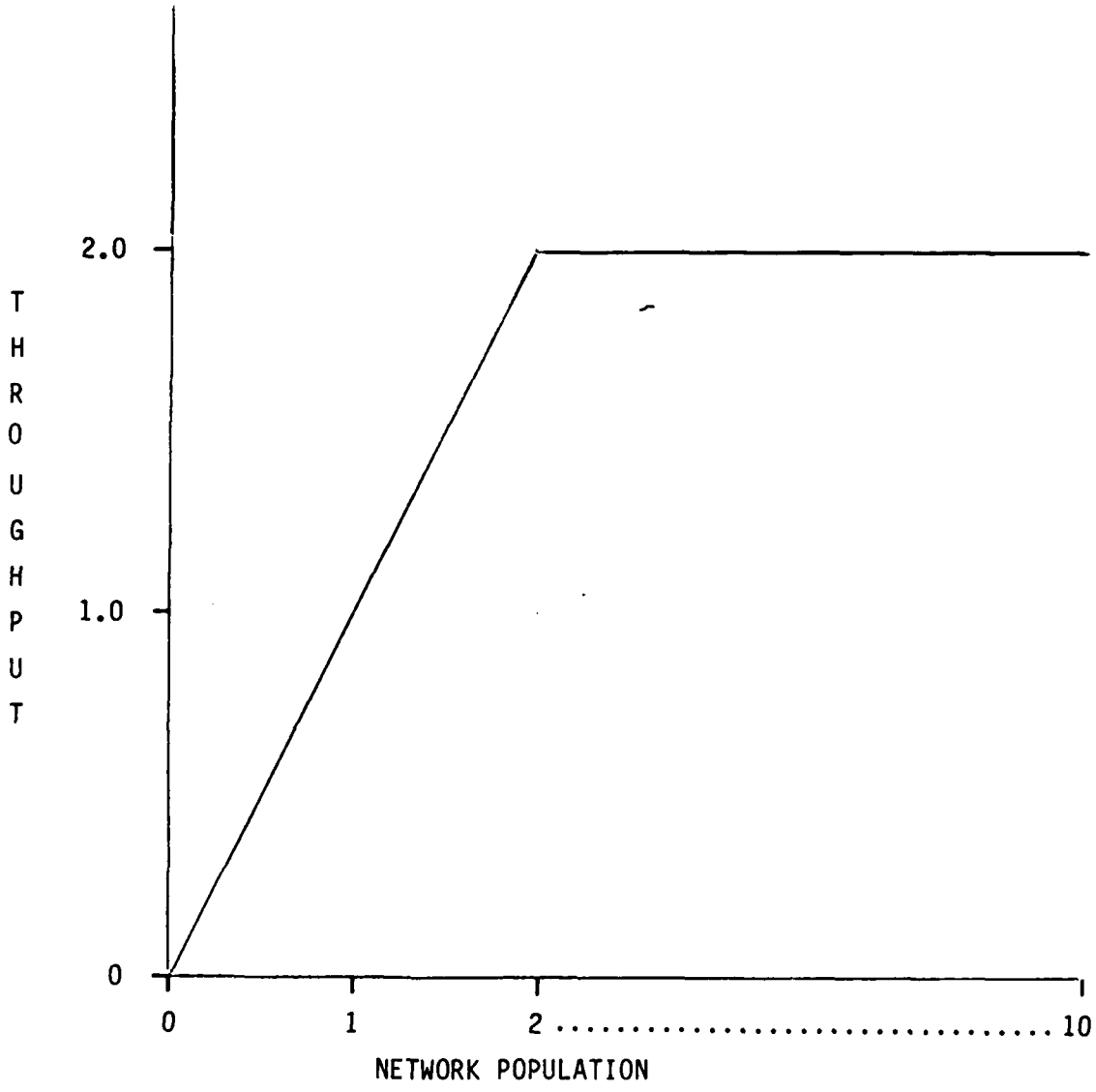


Figure 5-42. Augmented Secondary Subsystem Throughputs

procedure with the exact throughputs obtained by the multi-entrance queue procedure. This comparison is presented in Table 5-20, and graphically in Figure 5-43. Notice that for a subnetwork population of 3, the ECM

Table 5-20

Model 5-5. Isolated Subnetwork Throughputs

Subnetwork Population	Exact Throughput	ECM Throughput
1	1.000	1.00
2	1.600	1.60
3	1.818	2.00
4	1.905	2.00
5	1.944	2.00
6	1.965	2.00
7	1.976	2.00
8	1.983	2.00
9	1.988	2.00
10	1.991	2.00

procedure yields a throughput of 2.00 while the correct throughput is 1.90. There is error in all the remaining throughputs also. The cause of this larger than normal interpolation error is due to the interpretation of the augmented secondary subsystem throughputs. The ECM procedure treats these values as mean throughputs averaged over all possible population. For example, in determining the final throughput for subnetwork population 2, the primary contention model predicts that 1.6 will be allowed entry. Hence, interpolation is required between the augmented secondary subsystem throughput at populations of 1 and 2. As a result, the interpolation considers the behavior of the augmented secondary subsystem at a population of two and a correct result is obtained. In fact an exact answer is obtained by linear interpolation because the augmented secondary subsystem throughput response is linear in this region. Consider now what happens at a subnet-

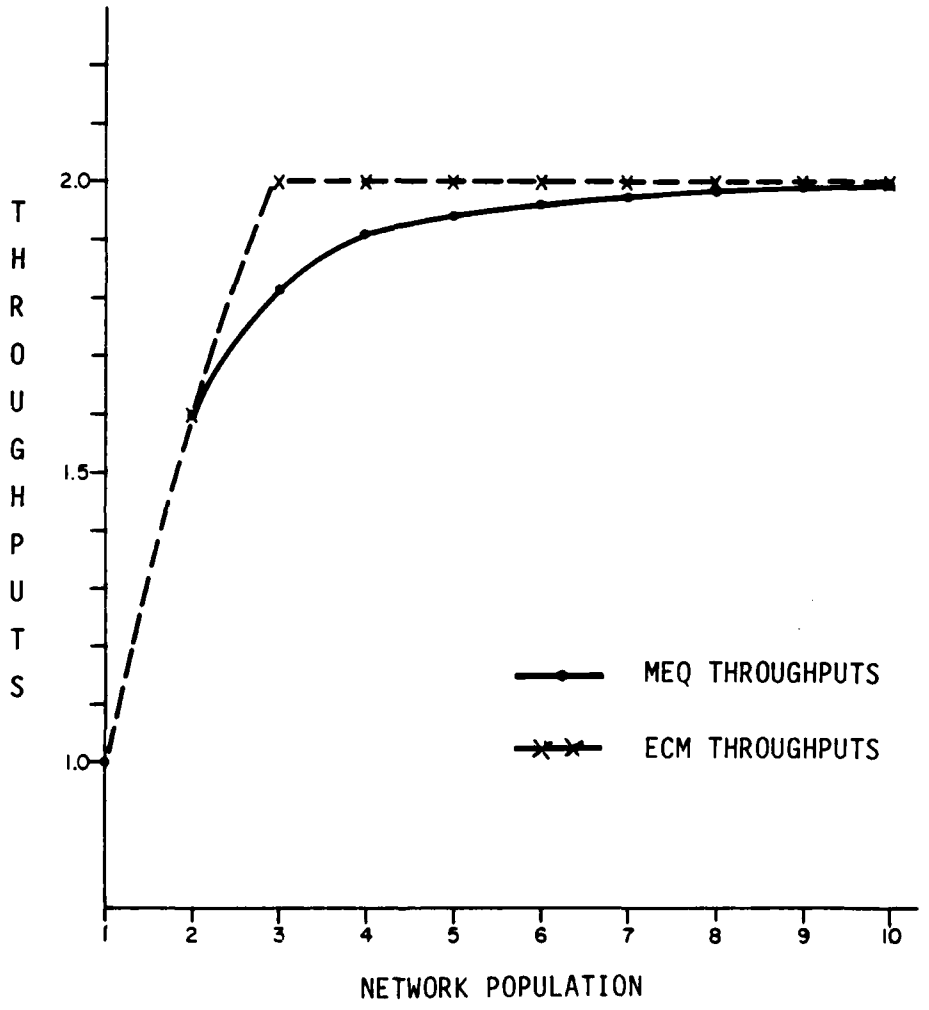


Figure 5-43. Model 5-5 Isolated Subnetwork Throughput Comparisons

work population of three. The primary contention model predicts that 2.0 customers will be allowed entry. Hence, no interpolation is required. However, the primary contention model considers this a mean value. There are nonzero probabilities that 1,2 or 3 customers can be in the augmented secondary subsystem.

Since the probability of one customer being in the augmented secondary subsystem is nonzero, where the throughput is 1.0, and the probability of two or three customers being in the augmented secondary subsystem is less than 1.0, where the throughput is 2.0, the mean throughput over all populations must abe less than 2.0. The error occurs because the augmented secondary subsystem throughput with a population of two does not consider the possibility for more customers to be present. The impacts in the final performance parameters can be noticeable. For example, the difference in throughputs of Models 5-1 through 5-5 in the ECM and multi-entrance queue procedures listed in Table 5-4 are due to this error.

In conclusion, the greatest potential for error exists when there are discontinuties in the slope of the throughput response of the augmented secondary subsystem. This situation usually occurs when congestion or a secondary utilization of 1.0 occurs at an augmented secondary subsystem population less than the maximum.

Relative Merits. The relative merits of the multi-entrance queue procedure over that of the ECM procedure includes:

1. The multi-entrance queue procedure solves the simplified representation of the subnetwork exactly where as the ECM procedure does not, and

2. The multi-entrance queue procedure allows determination of individual subnetwork performance parameters where as the ECM procedure does not. (They can be roughly estimated, however).

The main advantage that the ECM procedure has over the multi-entrance queue procedure is that:

1. It is simple to use and no special program is required and
2. It is applicable to a wider range of problems than the multi-entrance queue procedure.

It is recommended, however, that the multi-entrance queue procedure be used where applicable because it reduces the potential sources of error and is very easy to program. Algorithms are provided that give all the necessary information to obtain subnetwork performance estimates. In cases where the multi-entrance queue procedure does not apply, the ECM procedure appears most attractive.

Comparison of the Method of Surrogates and the Multi-entrance Queue Procedure

A minimal amount of error analysis was performed on the method of surrogates and five additional runs made using Model 1-3, except evaluated at populations from six to ten. The error analysis of the method of surrogates will be presented first followed by a comparison of the multi-entrance queue procedure versus the method of surrogates.

Error Analysis of the Method of Surrogates. The only attempt to identify sources of error in the method of surrogates was to see how accurately the representation of the subnetwork with simultaneous resource possession yielded the correct population dependent throughputs when analyzed in isolation. The 'correct' population dependent throughputs were determined by analyzing the subnetwork with simultaneous resource possession in isolation for each feasible subnetwork population using simulation. Model 1-3 was chosen, partly because most of the 'correct' population dependent throughputs were already available and partly because the method of surrogates converged in this example.

The primary contention model for model 1-3 is illustrated in Figure 3-18 and the secondary contention model in Figure 3-19. The representation of the subnetwork in the primary contention model is illustrated in Figure 5-44, and the representation in the secondary contention model in Figure 5-45. Chosen for

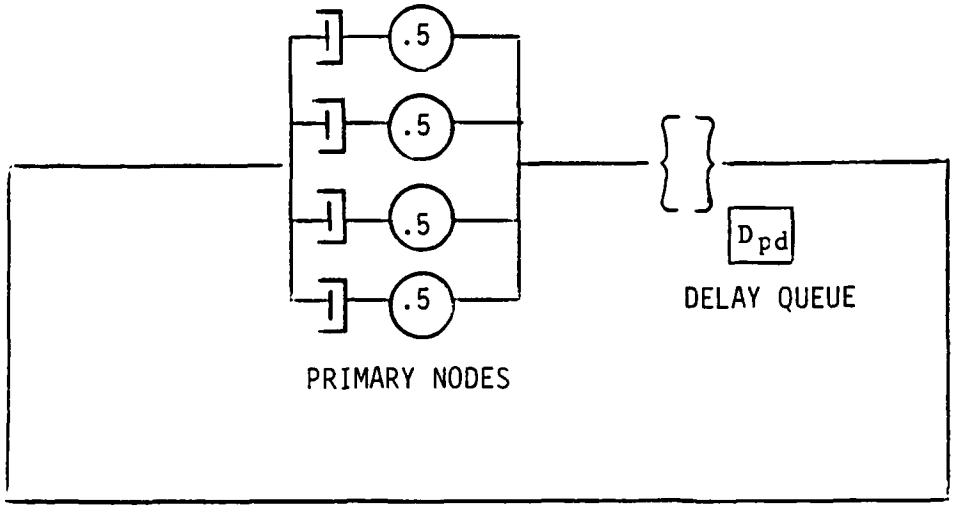


Figure 5-44. Method of Surrogate Primary Contention Model Representation of the I/O Subnetwork

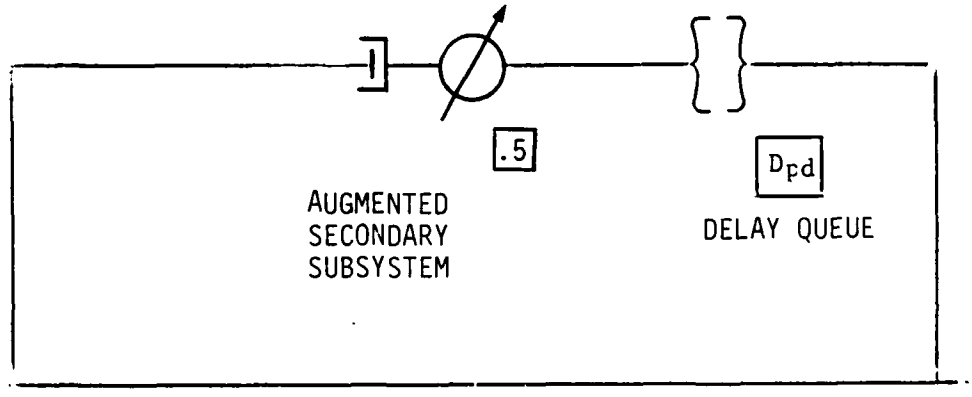


Figure 5-45. Method ot Surrogates Secondary Contention Model Representation of the I/O Subnetwork

further analysis was Model 1-3 evaluated at populations of six, eight, and ten. The simulated final throughputs for these models for all populations between five and ten are given in Table 5-21, along with the throughputs obtained by the method of surrogates and the multi-entrance queue procedure.

Table 5-21

Model 1-3. Performance Parameters at Different Populations

		ANALYTIC		SIMULATED			
Model Number	Network Population	Multi-entrance Throughput	% Error	Surrogate Delay Throughputs SCM	PCM	% Error	Throughput/ 97% Confidence Intervals
6-0	5	3.75	-3.8	4.03	4.04	3.1	3.91 (3.82-4.01)
6-1	6	3.92	-3.9	4.27	4.28	4.7	4.08 (3.97-4.20)
6-2	7	4.05	-3.6	4.45	4.46	6.0	4.20 (4.09-4.32)
6-3	8	4.14	-3.5	4.56	4.58	6.3	4.29 (4.14-4.44)
6-4	9	4.22	-3.2	4.64	4.66	6.4	4.36 (4.23-4.51)
6-5	10	4.28	-2.9	4.70	4.72	6.6	4.41 (4.28-4.55)

Three populations were chosen for further analysis, models 6-1, 6-3 and 6-5, so some idea of the error trend could be deduced. The mean service demand of the primary and secondary delay queues are given in Table 5-22

AD-A124 877 A NEW APPROACH FOR SOLVING SIMULTANEOUS RESOURCE POSSESSION PROBLEMS IN C.. (U) AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOOL OF ENGI.. D J FREUND

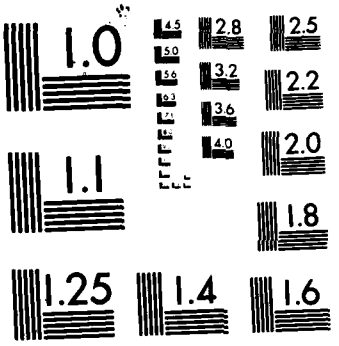
UNCLASSIFIED DEC 82 AFIT/GCS/MA/82D-2 F/G 12/1 NL

3/3

END

FILMED

DTIC



for populations five through ten, and were those obtained after convergence.

Table 5-22

Primary and Secondary Contention Model Queueing Delays

Model Number	Network Population	PCM Mean Queueing Delay D_{pd}	SCM Mean Queueing Delay D_{sd}
6-0 (1-3)	5	.269	.322
6-1	6	.367	.380
6-2	7	.480	.430
6-3	8	.611	.469
6-4	9	.759	.501
6-5	10	.923	.525

Table 5-23 contains the isolated subnetwork throughputs obtained by simulation along with the throughputs obtained by evaluating the surrogate delay representations of both models in isolation. The population dependent throughputs obtained from each model were different and hence each has a separate column in the table. For each model the first column

Table 5-23

Subnetwork Representations Evaluated in Isolation

Subnetwork Population	Simulated Throughput	Method Of Surrogates: Model 6-1 Throughput PCM	Model 6-1 Throughput SCM	Model 6-3 Throughput PCM	Model 6-3 Throughput SCM	Model 6-5 Throughput PCM	Model 6-5 Throughput SCM
1	2.00	1.15	1.14	0.90	1.03	0.70	0.98
2	3.03 (3.00-3.05)	2.13	2.16	1.71	1.98	1.36	1.88
3	3.58 (3.52-3.64)	2.95	3.04	2.44	2.82	1.98	2.69
4	3.90 (3.83-3.97)	3.62	3.75	3.08	3.52	2.55	3.39
5	4.10 (4.01-4.19)	4.16	4.25	3.63	4.06	3.07	3.94
6	4.22 (4.12-4.32)	4.61	4.54	4.11	4.41	3.54	4.32
7	4.32 (4.20-4.44)			4.53	4.61	3.97	4.55
8	4.38 (4.28-4.49)			4.88	4.70	4.35	4.68
9	4.44 (4.32-4.57)					4.68	4.73
10	4.48 (4.37-3.58)					4.98	4.75

is the set of throughputs obtained by evaluating the representation in the primary contention model in isolation, and the second column the set obtained by evaluating the representation in the secondary contention model in isolation. Notice that in most cases the population dependent throughputs of the primary contention model and secondary contention model representations are not identical. Figure 5-46 through Figure 5-48 graphically illustrate the isolated throughputs obtained from the method of surrogates along with the simulated throughputs for each model (6-1, 6-3, and 6-5). Notice that both surrogate delay representations of the subnetwork under estimate the throughput at low populations and over estimate the throughputs at high populations. Also observe that the surrogate delay representations increasingly under estimate the throughputs at the low

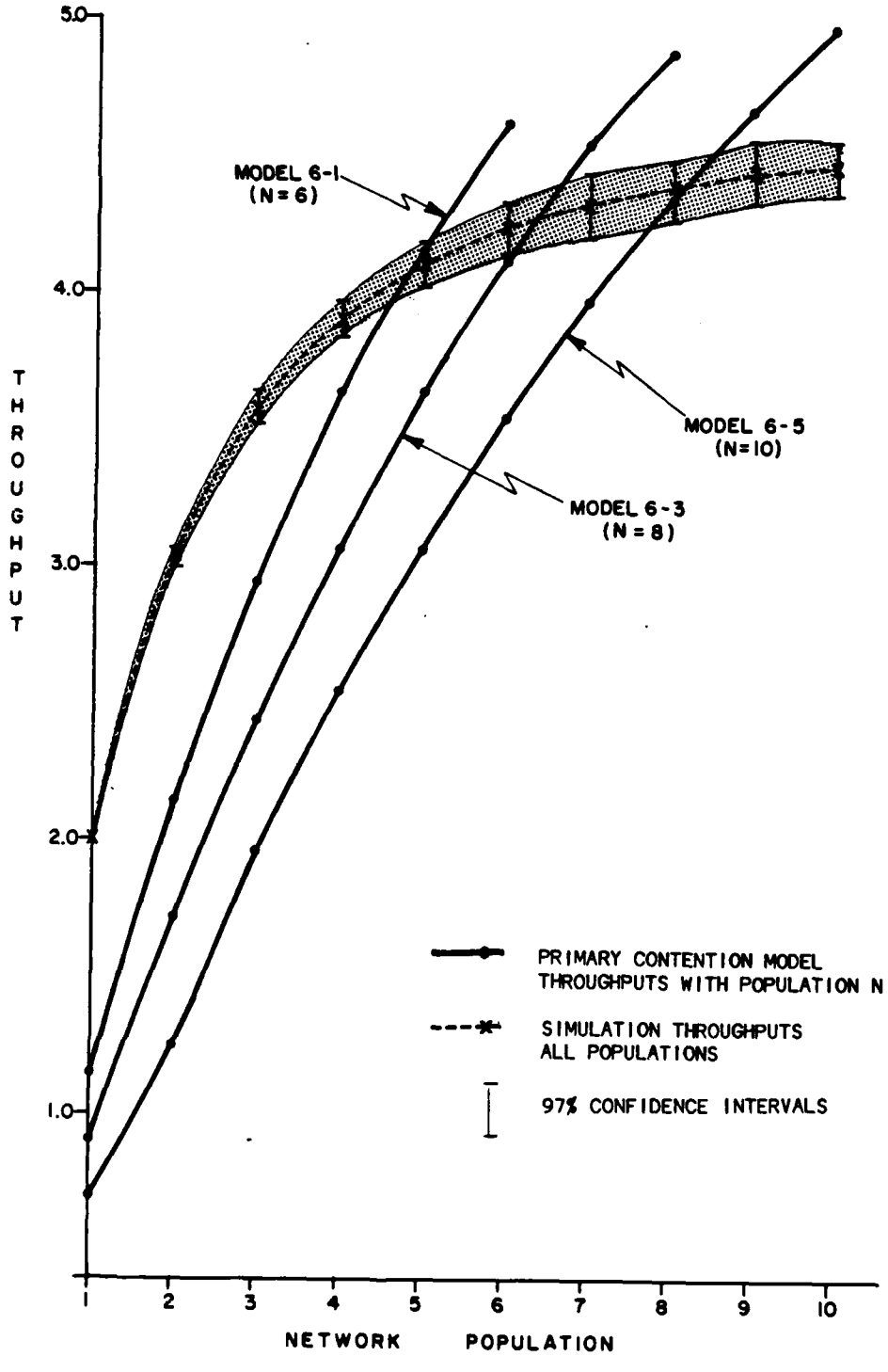


Figure 5-46. Throughput Comparisons of the Primary Contention Model Representation of the I/O Subsystem Evaluated in Isolation

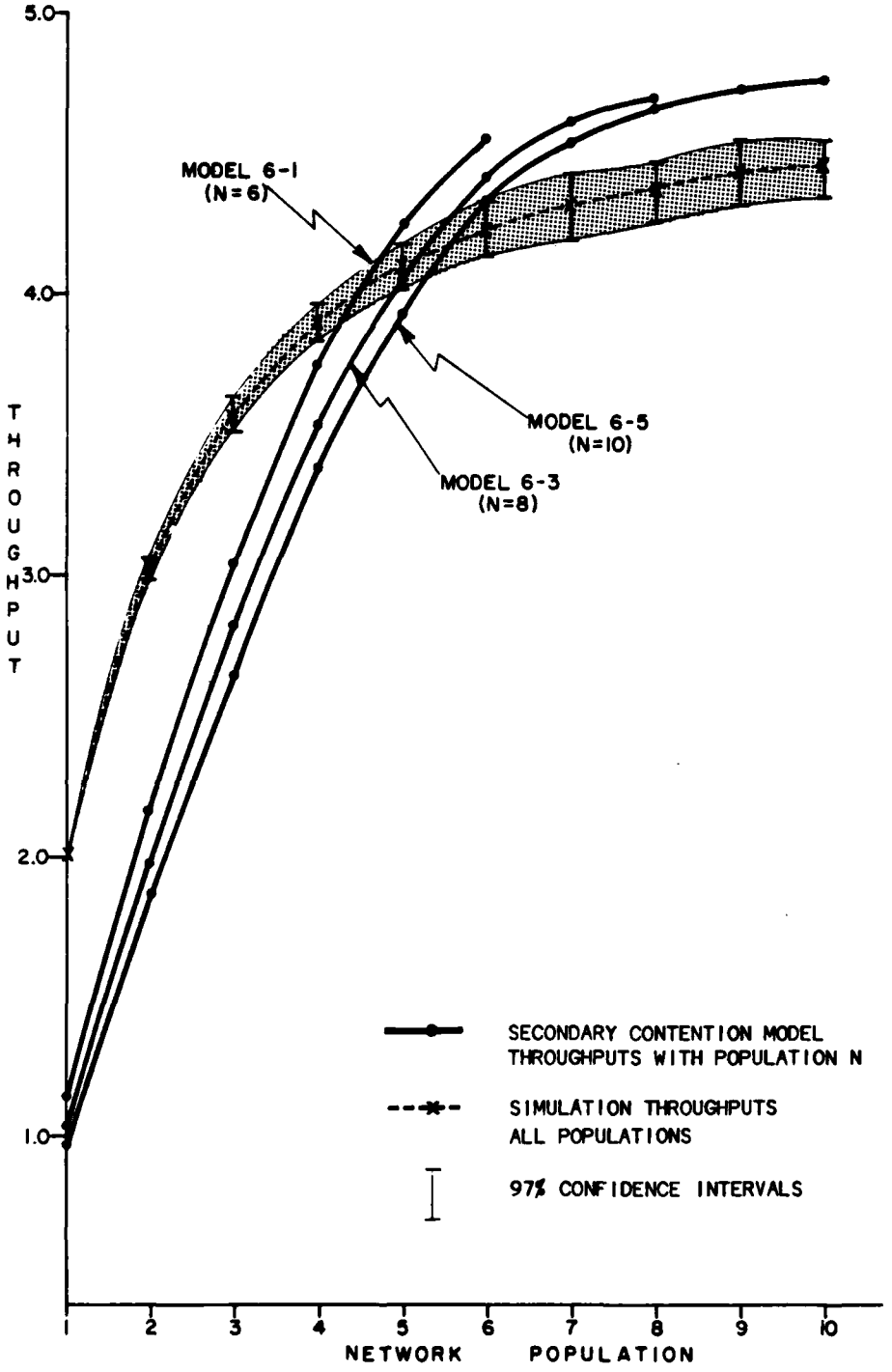


Figure 5-47. Throughput Comparisons of the Secondary Contention Model Representation of the I/O Subsystem Evaluated in Isolation

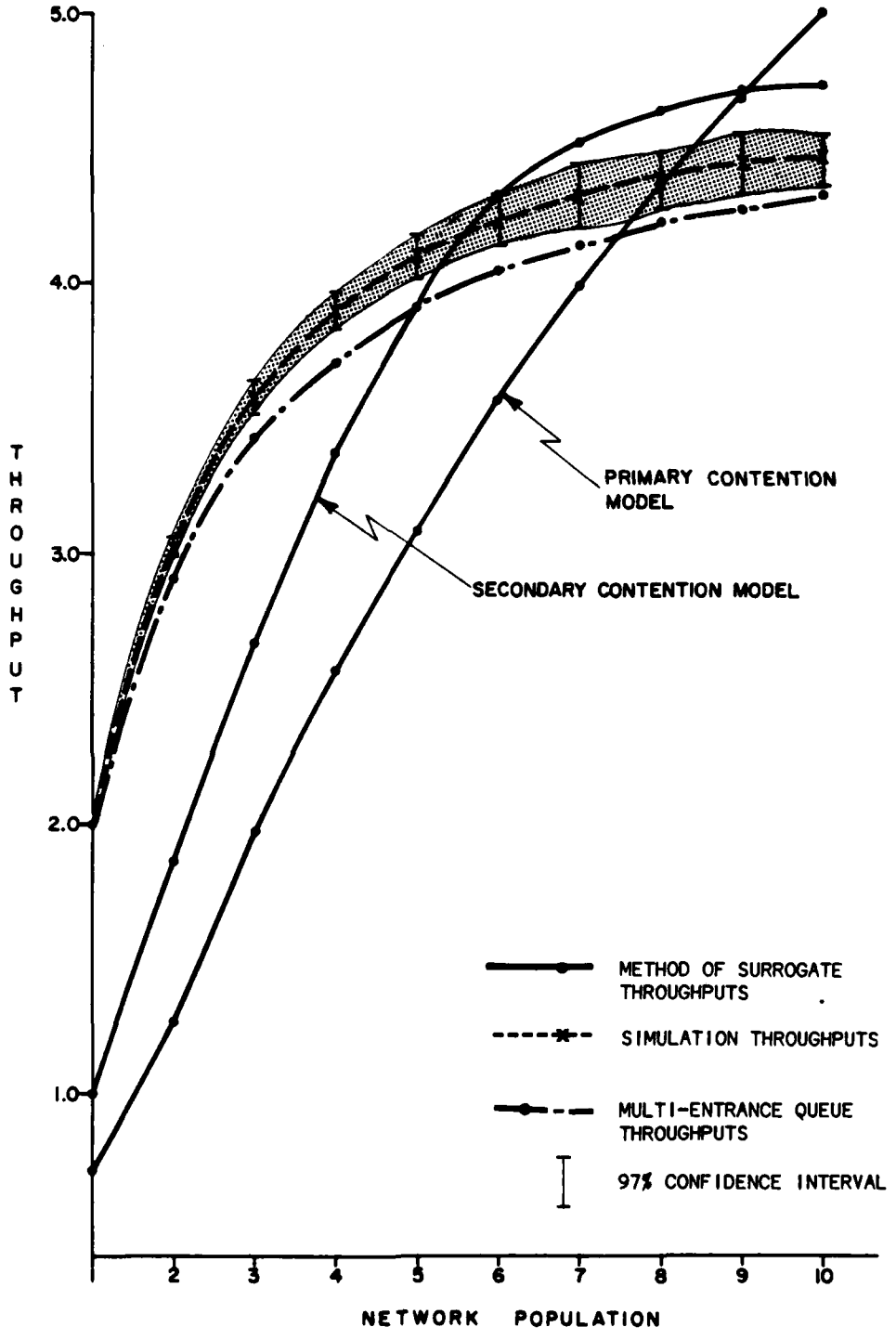


Figure 5-48. Model 6-5 Throughputs of I/O Subsystem Evaluated in Isolation

populations and increasingly over estimate the throughputs at the high populations as the original network population increases. For example, the exact throughput through the subnetwork evaluated in isolation with one customer present is 2.0. The throughput estimated by the primary contention model representation for model 6-1 is 1.15, for model 6-2 is 0.90, and for model 6-5 is 0.70.

The above results prompted the belief that representation of the subnetwork used by the method of surrogates really does not characterize its behavior over more than one population. It is believed that a better representation could be obtained if the subnetwork was analyzed in isolation using the method of surrogates for each feasible population and using the throughputs in a variable rate queue replacing the subnetwork. The amount of work required might be prohibitive, however.

Relative Merits of the Method of Surrogates. The method of surrogates requires further testing to determine its sources of error. In all the models solved using this procedure, the error was always less than 10 percent when the iterates converged, however the main disadvantages of the procedure are viewed as:

1. The number of iterates required for convergence
2. The necessity to reaccomplish the entire procedure for each network population performance parameters are desired, and
3. The observation that the representations of the subnetwork do not characterize the subnetwork when evaluated in isolation.

In spite of the above disadvantages there were cases when the method of surrogates yielded answers closer to the correct ones than the other two procedures. However, if this procedure is to be of use, additional analysis is required to determine its sources of error.

VI. Conclusions and Recommendations

In the previous five chapters a volume of material was presented, some a review of current state-of-the-art, and some new material. In this chapter a quick review will be made of the important points discussed, and some recommendations for further research.

The work in this thesis required the classification of simultaneous resource possession problems. A broad class of networks were said to fit into this group. For example, queues that have non-exponential severs can be represented as a series of exponential stages and modeled into a type one simultaneous resource possession problem. Networks that have blocking or population constants can also be fit into a simultaneous resource possession problem. However, a major class of networks that generally cannot be classified as simultaneous resource possession problems are networks that have complicated state dependent routing.

Networks that exhibit simultaneous resource possession were broken into two broad classes, type one and type two. Type one simultaneous resource possession problems were those where the subnetwork with simultaneous resource possession allocated its resources from a single queue or where it was not possible for a customer to wait at the primary resource allocation center whenever a resource was available. Type two simultaneous resource possession problems were those where the subnetwork with simultaneous resource possession allocated its resources in such a way that it was possible for a customer to queue for a specific resource even though others were available.

Within these two classes the primary resources could be active or passive. A passive resource was one where the primary resource acted solely to block entry into part of the network and had no service time associated exclusively with it. An active resource was one that had a service time associated with the resource. In many networks, it could be a matter of interpretation as to whether the resource is passive or active. The service time associated with the active resource was called the nonoverlapped service requirement, and the service time associated with the secondary subsystem was called the overlapped service requirement. The degree of overlap was said to be the ratio of the overlapped service demand to the total time the primary resource was held.

Little was said about the degree of overlap or the impact of active versus passive resources. This was because in most cases the problem could equivalently be considered as active or passive. This was an outcome of analysis of the augmented secondary subsystem model. For example, all of the I/O models analyzed could have been considered as having passive or active primary resources.

The major concept formalized in chapter three was the idea of external contention. External contention was said to be that queueing that occurred when customers were queued at primary resource allocation centers while resources were available that could not be used by those customers queued.

Two procedures were reviewed that would solve both type one and type two simultaneous resource possession problem, the method of external contention modification (ECM) and the method of surrogates. Both were developed primarily to solve type two problems. Each method deals with the issue of external contention differently. The ECM procedure compensated for external contention by adjusting the throughputs of a variable rate

queue, and the method of surrogates compensated for external contention by partitioning the queueing delay. Both methods were approximate and prone to error.

The multi-entrance queue procedure was developed specifically to solve a subset of type two simultaneous resource possession problems where each allocation queue only allocates one primary resource, and all nonoverlapped service requirements, if any, are equal for each resource. The procedure was based on representing the subnetwork with a multi-entrance queue and a set of exponential service times that depended on the number of nonempty queues. The multi-entrance queue model then determined population dependent rates that were used in a variable rate queue that replaced the subnetwork with simultaneous resource possession. The main advantage that the multi-entrance queue model had over the other methods was its ability to estimate performance parameters within the subnetwork with simultaneous resource possession. None of the other currently available methods could do this easily.

The multi-entrance queue procedure was analyzed for error in Chapter 5. This was done by partitioning the error into two possible causes representation error and decomposition error. Representation error was defined as error due to simplifying the representation of the subnetwork with simultaneous resource possession, and solving that network exactly in isolation. The cause of this error was found to be due to the assumption that the primary resource residency time was exponentially distributed. Additionally, the degree of queueing at the resource allocation queues and a newly defined measure, the secondary utilization, were said to impact the magnitude of the representation error possible.

Decomposition error was defined as the error due to the difference in performance measures of the actual results and the results of the analytic values connected for representation error. The cause of this error was said to be due to the interdeparture time distribution difference from exponential. The degree of queueing at the resource allocation queues and the primary resource utilization were identified as impacting the magnitude of decomposition error possible.

The partitioning of the error in the multi-entrance queue procedure required the assumption that the subnetwork with simultaneous resource possession could be exactly represented (as viewed externally) as a variable rate queue with a set of population dependent throughputs as long as the population dependent set of service time distributions was also considered as part of the queue.

The error present in the multi-entrance queue procedure solutions of all the models tested never exceeded six percent, and in the models chosen for additional analysis all of the significant error was due to representation error. However, it was hypothesized that some of error was due to decomposition error but was not detectable with the accuracy of the simulations used.

Finally, the multi-entrance queue was compared to the ECM procedure and the method of surrogates. It was found that the ECM procedure provided an approximate solution to the throughputs of the multi-entrance queue model where it was applicable, and that the procedures yielded similar results in most cases.

The method of surrogates was found to give more accurate results than the multi-entrance queue procedure in some cases and in other cases less accurate, or no solution at all. An investigation showed that the presen-

tations of the subnetwork used by the method of surrogates did not closely predict the behavior of the subnetwork evaluated in isolation but seemed to provide an approximate mean value analysis.

Finally it is believed that the multi-entrance queue procedure should be used wherever it is applicable. Its sources of error have been identified and measures developed to indicate the extent of error possible. When the multi-entrance queue procedure is not applicable, the ECM method is easy and straightforward to use, whereas the method of surrogates involves using an iterative process. However, the method of surrogates also has many cases where it yielded the best results, but it is not known under what circumstances the procedure will be more or less accurate. There are many implications and recommendations as a result of this work. Some of this will be discussed here.

One concept that needs further investigation is the issue of external contention. It appears that it can be characterized by a set of population dependent throughputs that are adjusted to reflect the mean population allowed entry. But what factors impact the mean population allowed entry? For example, it is known that the number and loading of the entry queues, congestion level of the sever, and at the type of service time distribution impact these values, and that the values are independent of the actual service time means as long as they are the same for each primary resource. But do any other parameters impact the values, and how? What happens when the mean service requirements through different entry queues are not all equal? This can be determined in the absence of server congestion (using product form methods), but how about in the presence of server congestion, and nonexponential servers? These questions remain unanswered.

Another area that has not been considered previously is classifying and identifying the sources of error in estimating the performance parameters of nonproduct form networks with various available techniques and product form networks. Little has been done in this area. This paper has identified the sources and suggested measures that influence the magnitude of error possible. Two types of errors were considered but they are basically the same; the difference was in the interpretation of what was error. Three measures were identified as impacting the error in replacing a subnetwork with a variable rate queue. Assuming the population dependent throughputs are known these were:

1. The interdeparture time coefficient of variation,
2. The degree of queueing at the queue, and
3. The utilization of the queue.

Presently the interdeparture coefficient of variation must be measured using simulation. However, an effort should be made to estimate these values analytically. This has already been started. For example, a paper by Chow (Ref 14) gives the solution to the cycle time distribution of a two queue cyclic network.

Another step should be to quantify the sources of error and devise correction or adjustment algorithms.

Emphasis should be placed in finding exact solutions to isolated subnetworks, such as the multi-entrance queue model, that has broad uses, and using that model as the multi-entrance queue model was used to develop an approximation procedure. The error could be investigated as it was in this thesis. It is believed that almost all network representations, when

Bibliography

1. Sauer, C. H. and Chandy, K. M. Computer Systems Performance Modeling. New Jersey: Prentice-Hall, Inc. 1981.

2. Jacobson, P. A. and Lazowska, E. D. 'Analyzing Queueing Networks with Simultaneous Resource Possession,' Communications of the ACM, 25: 142-151 (February 1982).

3. Baskett, F., Chandy, K. M., et al. 'Open, Closed, and Mixewd Networks of Queues with Different Classes of Customers,' Journal of the Association for Computing Machinery, 22: 248-260 (April 1975).

4. Chandy, K. M., Herzog, U., Woo, L. 'Parametric Analysis of Queueing Networks,' IBM Journal of Research and Development, 19: 43-49 (January 1975).

5. Sauer, C. H. 'Approximate Solution of Queueing Networks with Simultaneous Resource Possession,' IBM Journal of Research and Development, 25: 894-903 (November 1981).

6. Sauer, C. H. and Chandy, K. M. 'Approximate Solution of Queueing Models,' Computer, 13: 25-32 (April 1980).

7. Bexfield, J. N. Unpublished manuscript on approximation of closed queueing networks with simultaneous resource possession.

8. Bard, Y. 'An Analytic Model of the VM/370 System,' IBM Journal of Research and Development, 22: 498-508 (September 1978).

9. Chen, P. P. 'Queueing Network Model of Interactive Computer Systems,' Proceedings of the IEEEE, 63: 954-957 (June 1975).

10. Wilhelm, N. C. 'A General Model for the Performance of Disk Systems,' Journal of the Association for Computing Machinery, 24: 14-31 (January 1977)

11. Brown, R. M., Browne, J. C., and Chandy, K. M. 'Memory Management and Response Time,' Communications of the ACM, 20: 153-165 (March 1977).

12. Chandy, K. M. and Sauer, C. H. 'Approximate Methods for Analysis of Queueing Network Models of Computer Systems,' Computing Surveys, 10: 263-280 (September 1978).

13. Jacobson, P. A. and Lazowska, E. D. 'The Method of Surrogate Delays: Simultaneous Resource Possession in Analytic Models of Computer Systems,' Proceedings ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems, Las Vegas, September 1981.

VITA

Darral J. Freund was born 28 March 1955 in Washington, D. C. He graduated from Crossland High School, Temple Hills, Maryland in 1973 and attended the University of Maryland from which he received a Bachelor of Science degree in computer science in May 1977. Upon graduation, he received a commission in the USAF through the ROTC program. He was a computer systems design engineer at the Defense Intelligence Agency, Washington, D.C. until entering the School of Engineering, Air Force Institute of Technology in May 1981.

Permanent Address: 4703 Teak Court
Temple Hills, Maryland 20748

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFIT/CCS/MA/82D-2	2. GOVT ACCESSION NO. AD-A124877	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* A NEW APPROACH FOR SOLVING SIMULTANEOUS RESOURCE POSSESSION PROBLEMS IN CLOSED QUEUEING NETWORKS		5. TYPE OF REPORT & PERIOD COVERED MS Thesis 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Darral J. Freund		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Air Force Institute of Technology Wright-Patterson AFB, OH 45433		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE December, 1982 13. NUMBER OF PAGES 206
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

Approved for public release: IAW AFR 190-17.

LYNN E. WOLAVER
Dean for Research and Professional Development
Air Force Institute of Technology (ATC)
Wright-Patterson AFB OH 45433

4 JAN 1983

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Analytical modeling
Simultaneous resource possession
Closed queueing networks
Queueing
Computer performance evaluation

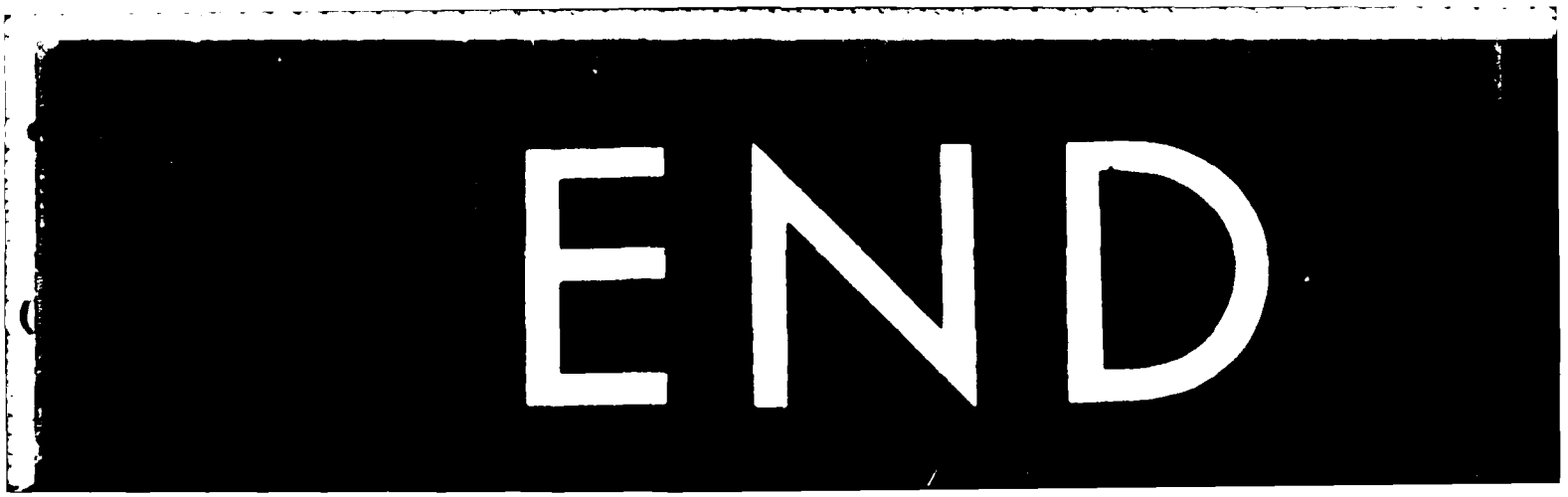
20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

A new approximate technique is presented for analyzing closed queueing networks with simultaneous resource possession. It is an analytic non-iterative solution procedure that is suitable for multiple entry systems such as I/O models. It relies on solving a closed queueing network consisting only of the subsystem with simultaneous resource possession, where queues can have service rates that are a function of the utilization of all the queues (number of busy servers). The submodel is solved using a newly discovered product form solution. Results are compared with simul-

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

ation models and other available analytic techniques.


END