A Double-Ended Single Server Queueing System

By

K. Laurie Dolhun

A Thesis
Submitted to the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree of

Master of Science

Department of Mechanical and Industrial Engineering
University of Manitoba
Winnipeg, Manitoba
Canada

0-612-23283-2

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION PAGE

A DOUBLE-ENDED SINGLE SERVER QUEUEING SYSTEM

BY

K. LAURIE DOLHUN

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirements of the degree

of

MASTER OF SCIENCE

The University of Manitoba requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

Automatic storage and retrieval systems (AS/RS) are widely used in warehousing due to such benefits as condensed storage and fast retrieval of goods. An end-of-aisle type of AS/RS is one of the most effective automatic storage and retrieval systems for warehousing and distribution operations.

This thesis presents a stochastic analysis of a simplified end-of-aisle AS/RS using a queueing model with two linked queues. A simplified AS/RS with one storage rack, and one storage/retrieval (S/R) device of unit-load is considered. There are two queues, one of infinite capacity for the items waiting for storage and the other of finite capacity for the requests for items to be removed from storage based on the size of the storage rack. The S/R machine places them into storage and retrieves items on an alternating basis.

Arrivals in both queues are assumed to follow a Poisson distribution, where the arrivals in the second queue are linked to the first queue. Service times of both queues follow an exponential distribution.

A double-ended queueing model is developed and is studied as a Markov process. The resulting Markov chain is of the quasi-birth-and-death type. The Matrix-geometric approach is used to analyze this system and efficient algorithmic procedures for the computation of the rate matrix, steady state vector and important performance measures have been developed.

Numerical examples are presented that show the behavior of the system for various rack sizes. As the rack size increases, the queue length and waiting time both decrease and system performance improves. However, it is shown that under certain conditions increasing the rack size gains minor improvements in system performance. The behavior of the system when jamming occurs is also discussed.

The queueing model presented in this thesis is only for a basic AS/RS. It is limited to Poisson arrivals and exponential service. Extensions of this work might be to non-Poisson arrivals and non-exponential service times.

Acknowledgements

I would like to express my sincere thanks to Dr. A.S. Alfa for his encouragement and support.

I would like to thank Dr. Y. Zhao and Dr. E. Rosenbloom for serving as examiners of this thesis.

I would also like to express my thanks and appreciation to my family and friends for all their support.

Table of Contents

List of Figures

Figure Page

Figure **Page**

List of Tables

Table **Page**

Chapter 1

Introduction

Automatic storage and retrieval systems (AS/RS) are widely used in warehousing and often found in manufacturing. An automatic storage and retrieval (AS/R) system has many benefits such as faster retrieval and condensed storage over conventional warehousing systems and are ideal for integrating with existing automation. These systems have the capability to integrate storage with inventory control and material resource planning. As a result, AS/R systems provide savings in labor costs, improved material flow and inventory control. These systems effectively utilize floor space since they are able to reduce storage requirements by using random storage assignment and minimize the wasted space due to aisles (Allen [1], Bafna [3] and [4], Rygh [16], Hill [11]).

The basic components of an automatic storage and retrieval system are storage racks and storage/retrieval (S/R) machines that transport the items into storage and retrieve them when a request for removal arises. There are several types of AS/R systems available:

- horizontal carousel - consists of shelving mounted on a motor driven horizontal track,
- vertical carousel - consists of shelving and a mechanism that rotates the shelving

on a vertical track,

- vertical lift - consists of modular column storage units with a storage/retrieval mechanism that moves vertically to the location of the desired item,
- end of aisle - consists of rows of storage units and a storage/retrieval mechanism(s) that moves the length of the aisles.

The best known type of AS/RS is the end-of-aisle type. An end-of-aisle AS/RS consists of opposed racks on either side of an aisle. Each aisle contains a stacker crane that is able to move the full length of the aisle and can travel vertically and side to side. There are many benefits to an end-of-aisle system. It is extremely flexible and only limited by the available floor space. The most effective applications are high volume with low to moderate activity, making it one of the most cost effective AS/R systems for warehousing and distribution operations (Allen [1]).

AS/R systems are generally designed to meet the specific needs of the individual user. Systems will differ in many respects such as the number of storage racks, the height, length and depth of the storage racks, and the number and size of the storage spaces. They will also vary in the number of S/R machines used, the number of items that can be carried at one time by the S/R machine (unit load or more), the manner in which the system carries out storage and retrievals of items, and whether the S/R machine is dedicated to one aisle or has the capability of transferring between aisles to service multiple storage racks. Other parameters include arrival and retrieval request rates, speed of service, and acceleration/deceleration rates of the S/R machines.

It is therefore important to have a model of an AS/RS that will provide information on the system performance. The use of a model as a tool in the design process of an AS/RS can help to ascertain the behavior of the system with respect to the

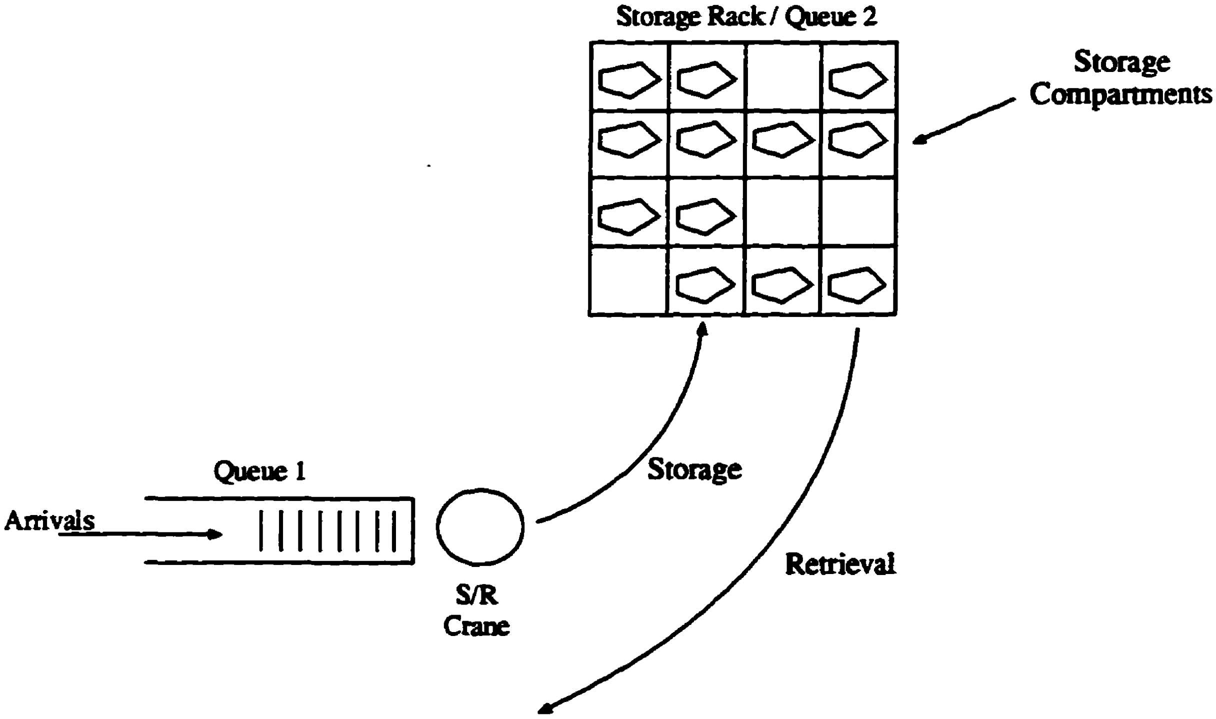

Figure 1.1: Simplified Model of an Automatic Storage and Retrieval System

capacity of the storage racks, the arrival and service rates and the many other parameters, before the system is installed. The model will aid in the decision process of an appropriate rack size resulting in the design of an efficient and cost effective system.

A very complex AS/RS can be decomposed into various simplified systems which can be easily modeled. A model can then be used to gain an understanding of the system behavior. This thesis presents a stochastic model that can be used to evaluate the performance of an AS/RS. The analytical model will be based on a simplified AS/R system as shown in Figure 1.1.

The simplified AS/R system is of the end-of-aisle type and will have one storage rack of capacity (size) M. There is assumed to be one stacker crane (server) that is capable of only serving a single item (customer) at one time. The loads may be packaged goods on a pallet or a single item. A pallet of packaged goods will be considered to be one item and cannot be split. All spaces available for storage in the rack are of the same size, and it is assumed that any item can fit into any storage space. Speed of service, directional movements of the crane, acceleration and deceleration of the crane, transfer time of an item in/out of the rack, etc., are assumed to be part of the service process and included in the service time. Additional assumptions made include that each pallet contains only one type of item and any pallet may be stored in any storage location.

This simplified system is modeled as a double-ended single server queueing system. The first queue, Queue 1, is considered to be the arrival queue where items wait to be placed into storage. Queue 2 is considered to be the queue in which the requests for retrieval arrive. Queue 2 is dependent on the number in the rack.

The remainder of this thesis is organized as follows. Chapter 2 reviews the previous research in the area of AS/RS and stochastic models and introduces the relevancy of the model presented in this thesis. Chapter 3 presents the model and corresponding Markov chain with the performance measures discussed in Chapter 4. Numerical examples and discussion are presented in Chapter 5 with a summary of this work and conclusions presented in Chapter 6.

Chapter 2

Literature

Many researchers have studied automated warehousing systems and AS/RS. Research done in these areas includes simulation, the development of heuristics and analytical models. However, little emphasis has been placed on analytical models. All three approaches have their advantages and disadvantages, and all have different circumstances in which they are best applied.

Many simulation models developed are for complex situations. As pointed out by many researchers, including Lee [12], simulation models are time-consuming to develop and validate even though they can capture most aspects of an AS/R system. The advantage of simulation is that it can represent any level of detail. However, simulation cannot be properly used for optimization purposes. Several people have used simulation such as Seidmann [18], Chow [6], Schwarz et al. [17], Lynn and Wysk [13], and Egbelu and Wu [7] and some have used simulation as a component of an optimization model such as Rosenblatt et al. [15].

Heuristic approaches have been used to aid in the scheduling of AS/R systems. However, with heuristic approaches, the performance of an AS/RS is based on deterministic parameters (often referred to as static models). The weakness of these models is that they lack the true operating aspect of AS/RS such as the stochastic aspect. Heuristics have been used by researchers such as Han et al. [9].

Deterministic models that use network models such as traveling salesman problem and vehicle routing have been used to determine optimal retrieval policies. However, this is not of interest in this thesis.

It appears that most researchers have used hybrids of various methods such as simulation, optimization, heuristics and network theory to study AS/RS. In addition existing queueing models have been used to study AS/RS. However, these applications have not been able to capture the true operating aspect of the system. As a result, queueing models have been used together with simulation and heuristics by people such as Chow [6].

There have been very few models that capture the true stochastic aspect of AS/RS. This is partly due to an attempt to model complex systems that are difficult to model stochastically. However, even for the simple cases, the number of models is very limited. Analytic methods have been studied by Schwarz et al. [17], Graves et al. [8] and Hausman et al. [10]. Bozer and White [5] present a stochastic analysis for a mini-load AS/RS. Stochastic optimization was studied by Azadivar [2].

Lee [12] is the first to present a stochastic analysis of a unit-load AS/RS using an analytical method. He considers a system which has two queues. The first queue is for the incoming items to be placed into storage and the second queue for the retrieval requests. Both queues have limited capacity. The arrivals are assumed to be Poisson and services in both queues are exponential. Arrivals into the second queue are independent of the storage rack and come from an external, unlimited source.

The model does not capture the state of the storage rack and therefore is not able to capture the impact of a full rack and empty second queue on the first queue, when the system is jammed. The server is idle only when both queues are empty.

The model presented in this thesis captures the state of the storage rack and links the two queues. Arrivals into the second queue are dependent on the number of items in the rack which in turn is dependent on the first queue. This type of model has not been previously studied in the literature, to the best of our knowledge. By keeping track of the state of the storage rack, the model is able to capture jamming. Linking of the two queues captures the operating aspect of an AS/R system where items cannot be removed from storage unless they have previously been placed into storage.

Chapter 3

Model

3.1 Double-Ended Queueing Model

The simplified model as described in Chapter 1 is now modeled as a double-ended queueing system in continuous time. A double-ended queueing system is a system that contains two queues that are linked with respect to their customers.

Consider the simplified AS/RS in Figure 1.1. As previously discussed, Queue 1 is considered to be the arrival queue where items wait to be placed into storage and Queue 2 is considered to be the queue in which the requests for retrieval arrive. Retrievals can only be made if items are available in storage. As such, a request for retrieval can only arrive if an item is in the rack. Figure 3.1 provides a simple illustration of the double-ended queueing model.

Queue 1 has an infinite buffer and arrivals into Queue 1 are from an external source. Customers can only arrive into Queue 2 after they have first received service in Queue 1. The storage rack is considered to be the source for arrivals into Queue 2 and has a fixed size, M. Arrival into Queue 2 depends on the output of Queue 1. Items are placed into and retrieved from storage on an ongoing basis, resulting in a source for arrivals that is constantly in flux. Therefore, the source for arrivals into Queue 2 is of variable size between 0 and M.

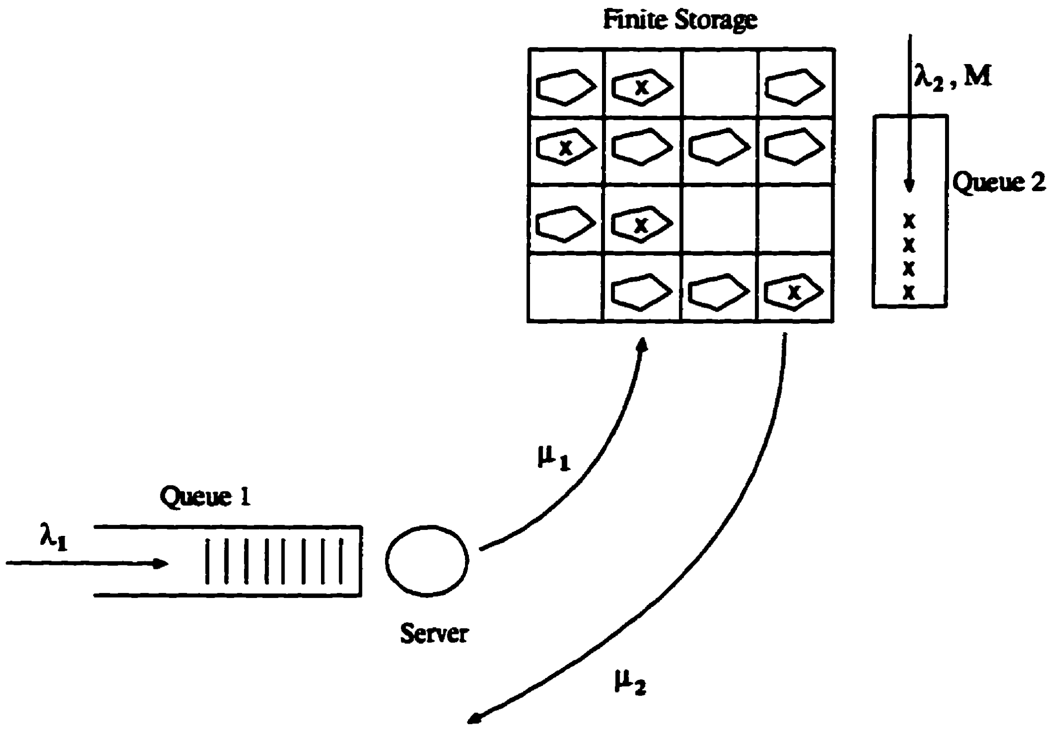

Figure 3.1: Double-Ended Single Server Queueing Model

The server is assumed to only serve one customer at a time (unit-load). Service of both queues follows a first-in, first-out (FIFO) rule and service times are independent.

Customers arrive into Queue 1 and wait for service. Service in this case means the placement of an item into the storage rack. Service of Queue 1 takes place if there is space available in rack. The location in the storage rack where the item is placed is chosen randomly from all open rack locations.

Upon a service completion of a customer in Queue 1, the server will serve the second queue. Items in the rack do not physically enter the second queue. Only a request for removal of the item arrives into Queue 2. In Figure 3.1 the items in the storage rack marked with an "x" are the items that have requests for removal in Queue 2.

If a request for removal of an item from the storage rack is waiting in Queue 2, the server will remove the item from the rack, which then leaves the system. If a request is not waiting, Queue 2 is empty but there may be items in the rack, the server will return to Queue 1 and serve the next customer if space is available.

The server will alternate between Queue 1 and 2 after each service completion. If either queue is empty, the server will continue service of the queue with customers waiting until:

- a customer arrives in the empty queue, then the server will begin alternate service at the completion of the current service, or
- the customers in the current queue being served are exhausted, or
- the system becomes jammed.

If both Queue 1 and 2 are empty, the server is idle and will begin service on the first customer that arrives.

Jamming of the system is defined as the state when the storage rack is full, there are customers waiting in Queue 1 for placement into storage, and Queue 2 is empty (there are no requests for items to be removed from storage). If the system becomes jammed, the server is forced to remain idle until a request for retrieval arrives in Queue 2. If jamming occurs, the system can become unstable if Queue 1 grows out of bounds.

Arrivals in Queue 1 are external and are assumed to follow a Poisson distribution with parameter (λ_1). Service of customers in Queue 1 is essentially the transportation of an item from the Queue 1 to a location in the storage rack and assumed to follow an exponential distribution with parameter (μ_1).

Customers served from Queue 1 are put into storage and become the source from which customers arrive into the second queue. As a result, the maximum number of customers that can be waiting in Queue 2 is M, the capacity of the rack. The actual number of customers in the rack is constantly changing, from 0 to M, as items are placed into and removed from storage. Arrivals in Queue 2 are therefore considered to be from a bounded source and are assumed to follow a Poisson distribution with parameter (λ_2). The arrival rate per item is λ_2. If there are N items in the rack, of which K are waiting to be served in Queue 2, then the actual arrival rate for Queue 2 would be $(N - K)\lambda_2$.

Service of customers in Queue 2 is essentially the transportation of the item from the storage location in the rack to the place of disposal. It is assumed to follow an exponential distribution with parameter (μ_2). Once a customer from Queue 2 is served it leaves the system.

3.2 Markov Process

The system that has just been described in Section 3.1 can be studied as a Markov Chain. Consider

$$\begin{aligned}
\Delta_0 &= \{(0,j,k);\ 0 \le j \le M,\ 0 \le k \le j\} \\
\Delta_1 &= \{(i,j,0);\ i \ge 1,\ 0 \le j \le M\} \\
\Delta_2 &= \{(i,j,1,k);\ i \ge 1,\ 1 \le j \le (M-1),\ 1 \le k \le j\} \\
\Delta_3 &= \{(i,j,2,k);\ i \ge 1,\ 1 \le j \le M,\ 1 \le k \le j\}.
\end{aligned}$$

The state space for this system can be described as Δ where:

$$\Delta = \Delta_0 \cup \Delta_1 \cup \Delta_2 \cup \Delta_3.$$

Δ_0 represents the states of the system when there are no customers waiting in the first queue, $i = 0$; with j items in the rack, $0 \le j \le M$, of which k items are waiting in the second queue, $0 \le k \le j$.

Δ_1 represents the states of the system when there is at least one customer in the first queue, $i \ge 1$; with j items in the rack, $0 \le j \le M$, of which no items are waiting for service in the second queue, $k = 0$.

Δ_2 represents the states of the system when there is at least one customer in the first queue, $i \ge 1$; with j items in the rack, $1 \le j \le M - 1$, of which k items are waiting in the second queue, $1 \le k \le j$, and the server is currently serving the first queue, $l = 1$.

Δ_3 represents the states of the system when there is at least one customer in the first queue, $i \ge 1$; with j items in the rack, $1 \le j \le M$, of which k items are waiting in the second queue, $1 \le k \le j$, and the server is currently serving the second queue, $l = 2$.

The generator matrix Q is of block tridiagonal form with infinite dimension, where the states of the Markov chain Q are arranged in lexicographic order.

$$Q = \begin{bmatrix} B_0 & C_0 & 0 & 0 & \dots & \dots & \dots \\ D_0 & A_1 & A_0 & 0 & \dots & \dots & \dots \\ 0 & A_2 & A_1 & A_0 & \dots & \dots & \dots \\ 0 & 0 & A_2 & A_1 & A_0 & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}. \tag{3.1}$$

Matrix Q is a generator matrix with negative diagonals and everything else non-negative. It can be shown that B_0^{-1} and A_1^{-1} exist, and both are non-positive, hence Q is irreducible. As a generator, the sum of the rows of matrix Q equal zero. Similarly the matrix $A = A_0 + A_1 + A_2$ is also a generator.

In what follows, the subblocks in Q are explained in detail. I_j is defined to be an identity matrix of dimension j. In addition, $\boldsymbol{e}$ is defined to be a column vector of ones, $\boldsymbol{e}(i)$ a column vector of ones of size (i), and $\boldsymbol{e}_i$ a column vector of zeros with the value one at position i.

3.2.1 Matrix B_0

The matrix B_0 is of dimension $\frac{1}{2}(M+1)(M+2) \times \frac{1}{2}(M+1)(M+2)$ and describes the transitions from boundary states in Δ_0 to boundary states in Δ_0, implying no change in the state of the first queue. Clearly B_0 represents the transition of the system when there are no arrivals or service completions in the first queue, however, there could be changes in the state of the rack and Queue 2.

$$B_0 = \begin{bmatrix} B_0^{00} & 0 & 0 & \dots & \dots \\ B_0^{10} & B_0^{11} & 0 & 0 & \dots \\ 0 & B_0^{21} & B_0^{22} & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & B_0^{M,M-1} & B_0^{M,M} \end{bmatrix}. \tag{3.2}$$

Within the matrix B_0 there are subblocks describing the state of the rack and the second queue.

1. The subblocks B_0^{jj} are of dimension $(j+1) \times (j+1)$ for $j = 0, \dots, M$ and represent no change in the total number of items in the rack. The elements, $B_0^{jj}(k_1, k_2)$, represent a change in the number of customers in the second queue from k_1 $(k_1 = 0, \dots, j)$ to k_2 $(k_2 = 0, \dots, j)$.

$$B_0^{jj}(k_1,k_2) = \begin{cases} a & 0 \le k_1 = k_2 \le j,\ 0 \le j \le M \\ b & 0 \le k_1 \le j-1,\ k_2 = k_1 + 1,\ 1 \le j \le M \\ 0 & \text{otherwise,} \end{cases} \tag{3.3}$$

where a and b are given in Equations (3.4) and (3.5).

$$a = -(\lambda_1 + (j - k_1)\lambda_2 + \delta_{k_1}\mu_2) \tag{3.4}$$

$$\delta_{k_1} = \begin{cases} 0 & k_1 = 0 \\ 1 & \text{otherwise,} \end{cases}$$

$$b = (j - k_1)\lambda_2. \tag{3.5}$$

2. The subblocks B_0^{jj-1} are of dimension $(j+1) \times j$ for $j = 1, \dots, M$ and represent a change from j to $j-1$ items in the rack. The elements of B_0^{jj-1} represent a change from k to $k-1$ customers in the second queue, i.e. a service completion in Queue 2.

$$B_0^{j,j-1} = \begin{bmatrix} 0 & 0 & 0 & \dots & \dots \\ \mu_2 & 0 & 0 & 0 & \dots \\ 0 & \mu_2 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \mu_2 & 0 \end{bmatrix}. \tag{3.6}$$

3.2.2 Matrix C_0

The matrix C_0 is of dimension $\frac{1}{2}(M+1)(M+2) \times (M^2+M+1)$ and describes the transitions from boundary states in Δ_0 to states in $\Delta_1 \cup \Delta_3$, implying an arrival in Queue 1.

$$C_0 = \begin{bmatrix} C_0^{00} & 0 & \dots & 0 \\ 0 & C_0^{11} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & C_0^{M,M} \end{bmatrix}. \tag{3.7}$$

The subblocks $C_0^{j,j}$, of dimension $(j+1) \times (2j+1)$ for $j = 0, \dots, M-1$, and $C_0^{M,M}$, of dimension $(j+1) \times (j+1)$, represent no change in the number of items in the rack and in the number of customers in the second queue.

$$C_0^{j,j} = \begin{bmatrix} \lambda_1 e_1' & 0 \\ 0 & I_j \lambda_1 \end{bmatrix} \tag{3.8}$$

.

3.2.3 Matrix D_0

The matrix D_0 is of dimension $(M^2+M+1) \times \frac{1}{2}(M+1)(M+2)$ and describes the transitions from states in $\Delta_1 \cup \Delta_2$ to boundary states in Δ_0, implying a service completion of the only waiting customer in Queue 1.

$$D_0 = \begin{bmatrix} 0 & D_0^{01} & 0 & \dots & 0 \\ 0 & 0 & D_0^{12} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & D_0^{M-1,M} \\ 0 & 0 & \dots & \dots & 0 \end{bmatrix}. \tag{3.9}$$

The subblocks $D_0^{j,j+1}$ are of dimension $(2j+1) \times (j+2)$ for $j = 1, \ldots, M-1$ and represent a change from j to $j+1$ items in the rack with no change in the number of customers in the second queue.

$$D_0^{0,1} = [\mu_1 \; 0], \tag{3.10}$$

$$D_0^{j,j+1} = \begin{bmatrix} I_j\mu_1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.11}$$

3.2.4 Matrix A_2

The matrix A_2 is of dimension $(M^2 + M + 1) \times (M^2 + M + 1)$ and describes the transitions from states in $\Delta_1 \cup \Delta_2 \cup \Delta_3, i \geq 2$ to states in $\Delta_1 \cup \Delta_2 \cup \Delta_3, i \geq 2$, with a change from i to $i-1$ customers in the first queue . This indicates that there has been a service completion in Queue 1.

$$A_2 = \begin{bmatrix} 0 & A_2^{01} & 0 & \dots & 0 \\ 0 & 0 & A_2^{12} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & A_2^{M-1,M} \\ 0 & 0 & \dots & \dots & 0 \end{bmatrix}. \tag{3.12}$$

The subblocks $A_2^{j,j+1}$, of dimension $(2j+1) \times (2j+3)$ for $j = 1, \ldots, M-2$, $A_2^{M-1,M-1}$, of dimension $(2j+1) \times (j+2)$, and $A_2^{M-1,M}$ of dimension $M \times 1$

represent a change from j to $j+1$ items in the rack with no change in the number of customers in the second queue.

$$A_2^{0,1} = [\mu_1 \; 0 \; 0] = \mu_1 \mathbf{e}_1', \tag{3.13}$$

$$A_2^{j,j+1} = \begin{bmatrix} (A_2^{j,j+1})_0 & (A_2^{j,j+1})_1 \\ 0 & 0 \end{bmatrix}, \tag{3.14}$$

where

$$(A_2^{j,j+1})_0 = \mu_1 (e_1 \otimes e_1'), \tag{3.15}$$

$$(A_2^{j,j+1})_1 = \begin{bmatrix} 0 & 0 \\ I_j\mu_1 & 0 \end{bmatrix}, \tag{3.16}$$

and $\otimes$ is the Kronecker product. $(A_2^{j,j+1})_0$ is of dimension $(j+1) \times (j+2)$ for $j = 1, \ldots, M-2$. $(A_2^{j,j+1})_1$ is of dimension $(j+1) \times (j+1)$ for $j = 1, \ldots, M-1$.

3.2.5 Matrix A_1

The matrix A_1 is of dimension $(M^2 + M + 1) \times (M^2 + M + 1)$ and describes the transitions from states in $\Delta_1 \cup \Delta_2 \cup \Delta_3, i \geq 1$ to states in $\Delta_1 \cup \Delta_2 \cup \Delta_3, i \geq 1$, with no change in the first queue. Clearly A_1 is the transition of the system when there are i customers in the first queue and there remains i customers in the first queue;

$$A_1 = \begin{bmatrix} A_1^{00} & 0 & 0 & \ldots & 0 \\ A_1^{10} & A_1^{11} & 0 & \ldots & 0 \\ 0 & A_1^{21} & A_1^{22} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & A_1^{M,M-1} & A_1^{M,M} \end{bmatrix}. \tag{3.17}$$

Within the matrix A_1 there are subblocks describing the state of the rack and the second queue.

1. The subblocks $A_1^{j,j}$, of dimension $(2j+1) \times (2j+1)$ for $j = 1, \ldots, M-1$, and $A_1^{M,M}$, of dimension $(j+1) \times (j+1)$, represent no change in the number of items in the rack. The elements represent and a change in the number of customer in the second queue from k_1 $(k_1 = 0, \ldots, j)$ to k_2 $(k_2 = 0, \ldots, j)$.

$$A_1^{0,0} = a_0, \tag{3.18}$$

$$A_1^{j,j} = \begin{bmatrix} (A_1^{j,j})_0 & 0 \\ 0 & (A_1^{j,j})_1 \end{bmatrix}, \tag{3.19}$$

$$A_1^{M,M} = \begin{bmatrix} (A_1^{M,M})_0 & (A_1^{M,M})_1 \\ 0 & (A_1^{M,M})_2 \end{bmatrix}. \tag{3.20}$$

$(A_1^{j,j})_0$ is of dimension $(j+1) \times (j+1)$ for $j = 1, \ldots, M-1$, $(A_1^{j,j})_1$ is of dimension $j \times j$ for $j = 1, \ldots, M-1$, $(A_1^{M,M})_1$ is of dimension $1 \times M$, and $(A_1^{M,M})_2$ is of dimension $M \times M$, where

$$(A_1^{M,M})_0 = -(\lambda_1 + M\lambda_2), \tag{3.21}$$

$$(A_1^{M,M})_1 = \begin{bmatrix} M\lambda_2 & 0 & \ldots & \ldots & 0 \end{bmatrix}, \tag{3.22}$$

$$(A_1^{j,j})_i(k_1,k_2) = \begin{cases} a_0 & i=0,\ 0 \le k_1 = k_2 \le j,\ 0 \le j \le M-1, \\ a_1 & i=1,\ 1 \le k_1 = k_2 \le j,\ 0 \le j \le M-1, \\ b & i=0,1,\ 0 \le k_2 = k_1+1 \le j-1,\ 1 \le j \le M, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$(A_1^{M,M})_2(k_1,k_2) = \begin{cases} a_1 & 1 \le k_1 = k_2 \le M, \\ b & 0 \le k_2 = k_1+1 \le M-1, \\ 0 & \text{otherwise.} \end{cases}$$

Expressions for a_0, a_1 and b are given in Equations (3.23) to (3.25).

$$a_0 = -(\lambda_1 + (j-k)\lambda_2 + \mu_1), \tag{3.23}$$

$$a_1 = -(\lambda_1 + (j-k)\lambda_2 + \mu_2), \tag{3.24}$$

$$b = (j-k)\lambda_2. \tag{3.25}$$

2. The subblocks $A_1^{j,j-1}$, of dimension $(2j+1) \times (2j-1)$ for $j = 2, \ldots, M-1$, and $A_1^{M,M-1}$, of dimension $(M+1) \times (2M-1)$ represent a change from j to $j-1$ items in the rack. The elements represent a change in the number of customers in the second queue from k to $k-1$, i.e. a service completion in Queue 2.

$$A_1^{1,0} = [0\ 0\ \mu_2]', \tag{3.26}$$

$$A_1^{j,j-1} = \begin{bmatrix} 0 & 0 \\ I_{j-1}\mu_2 & 0 \end{bmatrix}. \tag{3.27}$$

3.2.6 Matrix A_0

The matrix A_0 is of dimension $(M^2+M+1) \times (M^2+M+1)$ and describes the transitions from states in $\Delta_1 \cup \Delta_2 \cup \Delta_3, i \geq 1$ to states in $\Delta_1 \cup \Delta_2 \cup \Delta_3, i \geq 1$, with a change from i to $i+1$ customers in the first queue . This indicates that there has been an arrival in Queue 1.

$$A_0 = \begin{bmatrix} A_0^{00} & 0 & \dots & 0 \\ 0 & A_0^{11} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & A_0^{M,M} \end{bmatrix}. \tag{3.28}$$

The subblocks $A_0^{j,j}$, of dimension $(2j+1) \times (2j+1)$ for $j = 0, \dots, M-1$, and $A_0^{M,M}$, of dimension $(M+1) \times (M+1)$, represent no change in the number of items in the rack and no change in the number of customers in the second queue.

$$A_0^{j,j} = I_j\lambda_1. \tag{3.29}$$

3.3 Stationary Distribution

Our interest is to find the steady state vector $\boldsymbol{x} = [\boldsymbol{x}_0\ \boldsymbol{x}_1\ \boldsymbol{x}_2\ \dots\ \dots]$ by solving $\boldsymbol{x}Q = 0$ and $\boldsymbol{x}\boldsymbol{e} = 1$, where $\boldsymbol{x}_i$ for $i \geq 0$ is the stationary probability vector of being in state i at any particular time, i.e. of having i customers in Queue 1 including the one being served. Given the $\boldsymbol{x}_i$'s, we can easily calculate other performance measures.

The matrix Q is of the quasi-birth-and-death type. There exists a matrix R which is the minimal nonnegative solution to the matrix-quadratic equation,

$$R^2A_2 + RA_1 + A_0 = 0. \tag{3.30}$$

If Q is positive recurrent, then we know that this matrix R has all its eigenvalues inside the unit disk ($sp(R) < 1$). Throughout this thesis the term positive recurrence and stability are used interchangeably. The stability condition will be discussed in Section 3.5.

Given R, then

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i R, \quad \text{for} \quad i \geq 1. \tag{3.31}$$

Since we only know $\boldsymbol{x}_{i+1}$ in terms of $\boldsymbol{x}_i R$, for $i \geq 1$, we still need to obtain the boundary vectors $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$. The boundary behavior can be determined from Equations (3.32) and (3.33). The vector $[\boldsymbol{x}_0\ \boldsymbol{x}_1]$ can be obtained by

$$0 = [\boldsymbol{x}_0\ \boldsymbol{x}_1]B[R], \tag{3.32}$$

where

$$B[R] = \begin{bmatrix} B_0 & C_0 \\ D_0 & A_1 + RA_2 \end{bmatrix}. \tag{3.33}$$

The resulting vector $[\boldsymbol{x}_0\ \boldsymbol{x}_1]$ is normalized using Equation (3.34).

$$\boldsymbol{x}_0\boldsymbol{e} + \boldsymbol{x}_1[I - R]^{-1}\boldsymbol{e} = 1. \tag{3.34}$$

3.4 Matrix R

The matrix R is a very important part of getting performance measures of the system. It is often called the rate matrix. Given that the Markov chain is in level i, R is the expected number of times that level $i+1$ is visited before returning to level i.

An efficient method for calculating R is:

$$R(n+1) = -(R(n)^2 A_2 + A_0) A_1^{-1}, \tag{3.35}$$

where $R(n)$ is the value of matrix R at the n^{th} iteration with $R(0) = 0$.

The matrix R is of dimension $(M^2 + M + 1) \times (M^2 + M + 1)$ which could be very large if M is large, and can be partitioned as in Equation (3.36). In the block element $R_{v,w}$, v represents the number in the rack when the system starts from level i, and w represents the number in the rack when the system returns to level i.

$$R = \begin{bmatrix} R_{00} & R_{01} & \cdots & \cdots & R_{0M} \\ R_{10} & R_{11} & \cdots & \cdots & R_{0M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{M0} & R_{M1} & \cdots & \cdots & R_{MM} \end{bmatrix}. \tag{3.36}$$

However, we noticed that the matrices A_0, A_1 and A_2 are very sparse. We therefore take advantage of the sparsity of these matrices when calculating R. Equation (3.30) in Section 3.3 can be written in block form as

$$0 = \delta_{i,j} A_0^{j,j} + R_{i,j} A_1^{j,j} + (1 - \delta_{M,j}) R_{i,j+1} A_1^{j+1,j} + (1 - \delta_{0,j}) \sum_{v=0}^{M} R_{i,v} R_{v,j-1} A_2^{j-1,j}, \tag{3.37}$$

where

$$0 \leq j \leq M, \quad 0 \leq i \leq M \quad \text{and} \quad \delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & \text{otherwise.} \end{cases}$$

Equation 3.37 is now written for computational purposes as

$$R_{i,j} = -\left[\delta_{i,j} A_0^{j,j} + (1 - \delta_{M,j}) R_{i,j+1} A_1^{j+1,j} + (1 - \delta_{0,j}) \sum_{v=0}^{M} R_{i,v} R_{v,j-1} A_2^{j-1,j}\right] (A_1^{j,j})^{-1}. \tag{3.38}$$

3.5 Stability Conditions

To determine the steady state solution to the model presented, the system must be stable. In addition, to use Neuts' [14] matrix geometric method to solve the model, the system must be stable. Based on Neuts, the stability condition can be stated as

$$\boldsymbol{\pi} A_0 \mathbf{e} < \boldsymbol{\pi} A_2 \boldsymbol{e}, \tag{3.39}$$

where $\boldsymbol{\pi}$ is the invariant vector of matrix A, ie. $\boldsymbol{\pi} A = 0$ and $\boldsymbol{\pi e} = 1$.

First, let us consider the stability condition for a rack size of 1, $M = 1$. For this case

$$A_0 = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_1 \end{bmatrix}, \tag{3.40}$$

$$A_1 = \begin{bmatrix} -(\mu_1 + \lambda_1) & 0 & 0 \\ 0 & -(\lambda_2 + \lambda_1) & \lambda_2 \\ \mu_2 & 0 & -(\mu_2 + \lambda_1) \end{bmatrix}, \tag{3.41}$$

$$A_2 = \begin{bmatrix} 0 & \mu_1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{3.42}$$

and

$$\boldsymbol{\pi} = [\pi_{0,0}, \quad \pi_{1,0}, \quad \pi_{1,1,2}]. \tag{3.43}$$

After routine algebraic operations, the condition of Equation (3.39) for $M = 1$ results in the necessary and sufficient condition for stability:

$$\frac{1}{\lambda_1} > \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\lambda_2}. \tag{3.44}$$

The condition in Equation (3.44) implies that the mean interarrival time into the first queue, $\frac{1}{\lambda_1}$, must be greater than the sum of mean service time in Queue 1, the mean interarrival time in Queue 2 and the mean service time in Queue 2 for this system to be stable for $M = 1$.

Following this same procedure for $M = 2$, and applying Equation (3.39) the necessary and sufficient condition for stability in the first queue is given as

$$\frac{1}{\lambda_1} > \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{2\lambda_2}\left[\frac{\mu_1}{\mu_1 + \lambda_2}\right]\left[\frac{\mu_2}{\mu_2 + \lambda_2}\right]. \tag{3.45}$$

However, for $M = 3$, this procedure becomes very cumbersome, as such all we can obtain is the sufficient condition for stability. It is clear from the development that the stability condition for the general case is

$$\frac{1}{\lambda_1} > \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{M\lambda_2}\theta_M, \quad M \geq 1, \tag{3.46}$$

with $\theta_1 = 1$ and $\theta_2 = \left[\frac{\mu_1}{\mu_1+\lambda_2}\right]\left[\frac{\mu_2}{\mu_2+\lambda_2}\right]$. If we set $\theta_M = 1$ for all M in Equation (3.46), a sufficient condition for stability (but a conservative one) is;

$$\frac{1}{\lambda_1} > \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{M\lambda_2}, \quad M \geq 1. \tag{3.47}$$

Based on the patterns observed in Equation (3.44) and (3.45) it is *"conjectured"* that the expression θ_M of the general model for stability can be represented as

$$\theta_M = \prod_{j=0}^{M-1}\left(\frac{\mu_1}{\mu_1 + j\lambda_2}\right)\prod_{i=0}^{M-1}\left(\frac{\mu_2}{\mu_2 + i\lambda_2}\right). \tag{3.48}$$

3.6 Special Case of $M = 1$

A special case arises when the storage rack size is one, $M = 1$. Since there is space in the rack for only one item, only one item can be placed into the rack before the system becomes jammed. Jamming of the system forces the server to remain idle until a request for retrieval arrives in the second queue and then the server can remove the item from storage. Upon removal of the item the server can then place another item into storage.

After a closer examination of the system when $M = 1$, we discover that the server must serve only one customer from entry into the first queue until that customer leaves the system. Only after that customer leaves the system can the server serve another customer.

If we consider phase service, where service is provided to a customer in several phases, the double-ended queueing system, where $M = 1$, can be redefined into a single queue system which behaves like an $M/PH/1$ queue except at the boundary.

This enables the R matrix to be determined explicitly and to interpret the stability conditions.

The A_i matrices ($i = 0, 1, 2$) for $M = 1$ are as stated in Equations (3.40) to (3.42). The arrival rate is λ_1 and the service time is (β, S), where:

$$\beta = [0\ 1\ 0], \tag{3.49}$$

$$S = \begin{bmatrix} -\mu_1 & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 \\ \mu_2 & 0 & -\mu_2 \end{bmatrix}. \tag{3.50}$$

The first phase of service is placement of the item into storage, the second phase of service is that a retrieval request enters the second queue, and the third phase of service is retrieving the item from storage.

The A_i matrices ($i = 0, 1, 2$) for $M = 1$ simplify to give the following expressions:

$$A_0 = \lambda_1 I, \tag{3.51}$$

$$A_1 = S - \lambda_1 I, \tag{3.52}$$

$$A_2 = S^0 \beta. \tag{3.53}$$

The R matrix of this system can be obtained analytically as

$$R = \lambda_1(\lambda_1 I - \lambda_1(e\beta) - S)^{-1}, \tag{3.54}$$

and

$$\rho = \frac{\lambda_1}{\mu}, \tag{3.55}$$

$$\frac{1}{\mu} = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\lambda_2}. \tag{3.56}$$

Chapter 4

Performance Measures

Obtaining performance measures is essential in order to utilize the queueing model described in Chapter 3 as a tool in the design of an AS/R system. These performance measures allow us to determine how the system is behaving for a specific size of storage rack given the arrival and service rates. Comparison on these performance measures for various storage rack sizes will provide insight into what size of an AS/R system would be most appropriate for the give n conditions. The model can also be used to determine the behavior of the system and ultimately help decide the best operating conditions of an AS/RS (arrival and service rates) for a given storage rack size.

The stationary distribution is given as

$$\boldsymbol{x} = [\boldsymbol{x}_0,\ \boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \ldots\ldots]. \tag{4.1}$$

Each term of Equation (4.1) can be written as

$$\boldsymbol{x}_i = [x_{i,0},\ x_{i,1},\ x_{i,2},\ \ldots\ldots,\ x_{i,M}] \quad \text{for} \quad i \geq 0\ , \tag{4.2}$$

where

$$\boldsymbol{x}_{0,j} = [x_{0,j,0},\ x_{0,j,1},\ \ldots\ldots,\ x_{0,j,k}] \quad \text{for} \quad j=0,\ldots,M; \ \ k=0,\ldots,j \ , \tag{4.3}$$

and

$$\boldsymbol{x}_{i,j} = [x_{i,j,0},\ x_{i,j,l}] \quad \text{for} \quad i \geq 1; \ \ j=0,\ldots,M \ . \tag{4.4}$$

The term $\boldsymbol{x}_{i,j,l}$ of Equation (4.4) can be written as

$$\boldsymbol{x}_{i,j,l} = [x_{i,j,1},\ x_{i,j,2}], \tag{4.5}$$

where

$$\boldsymbol{x}_{i,j,1} = [x_{i,j,1,1},\ x_{i,j,1,2},\ \ldots\ldots,\ x_{i,j,1,k}] \quad \text{for} \quad k=1,\ldots,j \ , \tag{4.6}$$

and

$$\boldsymbol{x}_{i,j,2} = [x_{i,j,2,1},\ x_{i,j,2,2},\ \ldots\ldots,\ x_{i,j,2,k}] \quad \text{for} \quad k=1,\ldots,j \ . \tag{4.7}$$

Using the probabilities obtained in the stationary distribution we are then able to obtain performance measures such as queue length, average waiting time, idleness and jamming probability for the system. These performance measures enable us to analyze the behavior of the system.

4.1 Queue Length

4.1.1 Mean Number in the System

Queue 1

The mean number of customers in the system for the Queue 1 ($\mu_{L_{1s}}$) is given as

$$\mu_{L_{1s}} = \sum_{i \geq 1} \sum_{j=0}^{M} i\boldsymbol{x}_{i,j,0} + \sum_{i \geq 1} \sum_{j=1}^{M-1} \sum_{l=1}^{2} \sum_{k=1}^{j} i\boldsymbol{x}_{i,j,l,k} + \sum_{i \geq 1} \sum_{k=1}^{M} i\boldsymbol{x}_{i,M,2,k}. \tag{4.8}$$

This indicates the number of customers waiting for service in the first queue including the customer being served. The mean number in the system for Queue 1 is important since this queue can grow out of bounds if the system is unstable, either from a violation of the stability conditions or from the storage rack becoming full and the system jammed.

If we reorganize the terms in Equation (4.8) according to stationary distribution for $\boldsymbol{x}_i$ shown in Equation (4.2) we obtain the following expression:

$$\mu_{L_{1s}} = (1\boldsymbol{x}_1 + 2\boldsymbol{x}_2 + 3\boldsymbol{x}_3 + 4\boldsymbol{x}_4 + \ldots)\boldsymbol{e}. \tag{4.9}$$

Since $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i R^i$ Equation (4.9) becomes:

$$\mu_{L_{1s}} = (1\boldsymbol{x}_1 + 2\boldsymbol{x}_1 R + 3\boldsymbol{x}_1 R^2 + 4\boldsymbol{x}_1 R^3 + \ldots)\boldsymbol{e}. \tag{4.10}$$

Equation (4.10) can be reduced to the following:

$$\mu_{L_{1s}} = \boldsymbol{x}_1 \frac{d}{dR}(I + R + R^2 + R^3 + R^4 + \ldots)\boldsymbol{e}. \tag{4.11}$$

Equation (4.11) can be rewritten using $\boldsymbol{x}_1$ and R as

$$\mu_{L_{1s}} = \boldsymbol{x}_1(I - R)^{-2}\boldsymbol{e}. \tag{4.12}$$

Queue 2

The mean number of customers in the system for the Queue 2 ($\mu_{L_{2s}}$) is given as

$$\mu_{L_{2s}} = \sum_{j=1}^{M}\sum_{k=1}^{j} kx_{0,j,k} + \sum_{i\geq 1}\sum_{j=1}^{M-1}\sum_{l=1}^{2}\sum_{k=1}^{j} kx_{i,j,l,k} + \sum_{i\geq 1}\sum_{k=1}^{M} kx_{i,M,2,k}. \tag{4.13}$$

This indicates the number of customers waiting for service in the second queue, including the customer currently being served. Expressing Equation (4.13) in terms of $\boldsymbol{x}_1$ and R gives us the following expression.

$$\mu_{L_{2s}} = \boldsymbol{x}_0\boldsymbol{b}_1 + \boldsymbol{x}_1(I - R)^{-1}\boldsymbol{b}_2, \tag{4.14}$$

where

$$\boldsymbol{b}_1 = [c_0\ c_1\ \ldots\ldots\ c_i]' \text{ for } i = 0, \ldots M, \tag{4.15}$$

$$c_i = [0\ 1\ \ldots\ldots\ i] \text{ for } i = 0, \ldots M,$$

and

$$\boldsymbol{b}_2 = [d_0\ d_1\ \ldots\ldots\ d_i]' \text{ for } i = 0, \ldots M, \tag{4.16}$$

$$d_0 = [0],$$

$$d_i = [d_{i0}\; d_{i1}] \;\; \text{for} \;\; i = 1, \ldots M,$$

$$d_{i0} = [0\; 1\; 2\; \ldots\ldots\; i], \;\; d_{i1} = [1\; 2\; \ldots\ldots\; i] \;\; \text{for} \;\; i = 1, \ldots M.$$

4.1.2 Mean Number in the Queue

Queue 1

The mean number of customers in the queue for the Queue 1 ($\mu_{L_{1q}}$) can also be written using $\boldsymbol{x}_1$ and R and is given as

$$\mu_{L_{1q}} = \sum_{i\geq 1} \sum_{j=0}^{M-1} (i-1)x_{i,j,0} + \sum_{i\geq 1} \sum_{j=1}^{M-1} \sum_{k=1}^{j} (i-1)x_{i,j,1,k} + \sum_{i\geq 1} \sum_{j=1}^{M} \sum_{k=1}^{j} i x_{i,j,2,k} + \sum_{i\geq 1} i x_{i,M,0}. \tag{4.17}$$

Equation (4.17) can be rewritten as:

$$\mu_{L_{1q}} = \mu_{L_{1s}} - \sum_{i\geq 1} \sum_{j=0}^{M-1} x_{i,j,0} - \sum_{i\geq 1} \sum_{j=1}^{M-1} \sum_{k=1}^{j} x_{i,j,1,k}. \tag{4.18}$$

Equation (4.18) can be written in terms of $\boldsymbol{x}_1$ and R as

$$\mu_{L_{1q}} = \mu_{L_{1s}} - \boldsymbol{x}_1 (I - R)^{-1} \boldsymbol{b}_3, \tag{4.19}$$

where

$$\boldsymbol{b}_3 = [f_0\; f_1\; \ldots\ldots\; f_i]' \;\; \text{for} \;\; i = 0, \ldots M, \tag{4.20}$$

$$f_0 = [1],$$

$$f_i = [f_{i0}\; 0] \;\; \text{for} \;\; i = 1, \ldots M - 1,$$

$$f_{i0} = \boldsymbol{e}(i+1)'.$$

$$f_M = [0].$$

f_i is of size $2i+1$ for $i = 1, \ldots, M-1$ and f_M is of size $i+1$. This indicates the number of customers waiting for service in the first queue not including the customer currently being served.

Queue 2

The mean number of customers in the queue for the Queue 2 ($\mu_{L_{2q}}$) is given as

$$\mu_{L_{2q}} = \sum_{j=1}^{M} \sum_{k=1}^{j} (k-1) x_{0,j,k} + \sum_{i \geq 1} \sum_{j=1}^{M-1} \sum_{k=1}^{j} k x_{i,j,1,k} + \sum_{i \geq 1} \sum_{j=1}^{M} \sum_{k=1}^{j} (k-1) x_{i,j,2,k}. \qquad (4.21)$$

Equation (4.21) can be rewritten as:

$$\mu_{L_{2q}} = \mu_{L_{2s}} - \sum_{j=1}^{M} \sum_{k=1}^{j} x_{0,j,k} - \sum_{i \geq 1} \sum_{j=1}^{M} \sum_{k=1}^{j} x_{i,j,2,k}. \qquad (4.22)$$

Expressing Equation (4.22) in terms of $\boldsymbol{x}_1$ and R gives

$$\mu_{L_{2q}} = \mu_{L_{2s}} - \boldsymbol{x}_0 \boldsymbol{b}_4 - \boldsymbol{x}_1 (I - R)^{-1} \boldsymbol{b}_5, \qquad (4.23)$$

where

$$\boldsymbol{b}_4 = [g_0 \; g_1 \; \ldots\ldots \; g_i]' \;\; \text{for} \;\; i = 0, \ldots M, \qquad (4.24)$$

$$g_0 = [0],$$

$$g_i = [0 \; g_{i1}] \text{ for } i = 1, \dots M,$$

$$g_{i1} = \boldsymbol{e}(i)',$$

and

$$\boldsymbol{b}_5 = [h_0 \; h_1 \; \dots\dots \; h_i]' \text{ for } i = 0, \dots M, \tag{4.25}$$

$$h_0 = [0],$$

$$h_i = [0 \; h_{i1}] \text{ for } i = 1, \dots M,$$

$$h_{i1} = \boldsymbol{e}(i)'.$$

g_i is of size $i+1$ for $i = 1, \dots, M$, h_i is of size $2i+1$ for $i = 1, \dots, M-1$ and h_M is of size $i+1$.

This indicates the number of customers waiting for service in the second queue not including the customer currently being served.

4.2 Probability of Server Being Idle

The probability that the server is idle, y_0, is given as

$$y_0 = \sum_{j=0}^{M} x_{0,j,0} + \sum_{i \geq 1} x_{i,M,0}. \tag{4.26}$$

Equation (4.26) can be expressed in terms of $\boldsymbol{x}_1$ and R as

$$y_0 = \boldsymbol{x}_0 \boldsymbol{b}_6 + \boldsymbol{x}_1 (I - R)^{-1} \boldsymbol{b}_7, \tag{4.27}$$

where

$$\boldsymbol{b}_6 = [\boldsymbol{n}_0\ \boldsymbol{n}_1\ \ldots\ldots\ \boldsymbol{n}_i]' \text{ for } i = 0, \ldots M, \tag{4.28}$$

$$\boldsymbol{n}_0 = [0],$$

$$\boldsymbol{n}_i = [\boldsymbol{n}_{i0}\ 0] \text{ for } i = 1, \ldots M,$$

$$\boldsymbol{n}_{i0} = [1],$$

and

$$\boldsymbol{b}_7 = [\boldsymbol{p}_0\ \boldsymbol{p}_1\ \ldots\ldots\ \boldsymbol{p}_i]' \text{ for } i = 0, \ldots M, \tag{4.29}$$

$$\boldsymbol{p}_i = [0] \text{ for } i = 1, \ldots M - 1,$$

$$\boldsymbol{p}_M = [\boldsymbol{p}_{M0}\ 0]$$

$$\boldsymbol{p}_{M0} = [1].$$

$\boldsymbol{n}_i$ is of size $i + 1$ for $i = 1, \ldots, M$, $\boldsymbol{p}_i$ is of size $2i + 1$ for $i = 1, \ldots, M - 1$ and $\boldsymbol{p}_M$ is of size $i + 1$.

The probability that the server is idle indicates that the server is not serving any customers. Server idleness can arise from two different states. Either there are no customers waiting for service in both Queue 1 and 2 or the system is jammed and the server must wait for a customer to arrive in Queue 2 before service can resume.

4.3 Mean Number of Items in the Rack

The mean number of items in the rack μ_{N_r} is given as

$$\mu_{N_r} = \sum_{j=1}^{M}\sum_{k=1}^{j} jx_{0,j,k} + \sum_{i\geq 1}\sum_{j=1}^{M} jx_{i,j,0} + \sum_{i\geq 1}\sum_{j=1}^{M-1}\sum_{l=1}^{2}\sum_{k=1}^{j} jx_{i,j,l,k} + \sum_{i\geq 1}\sum_{k=1}^{M} Mx_{i,M,2,k}. \quad (4.30)$$

Expressing Equation (4.30) in terms of $\boldsymbol{x}_1$ and R gives

$$\mu_{N_r} = \boldsymbol{x}_0\boldsymbol{b}_8 + \boldsymbol{x}_1(I - R)^{-1}\boldsymbol{b}_9, \quad (4.31)$$

where

$$\boldsymbol{b}_8 = [s_0\ s_1\ \ldots\ldots\ s_i]' \text{ for } i = 0, \ldots M, \quad (4.32)$$

$$s_0 = [0],$$

$$s_i = \boldsymbol{e}(i+1)'i \text{ for } i = 1, \ldots M,$$

and

$$\boldsymbol{b}_9 = [t_0\ t_1\ \ldots\ldots\ t_i]' \text{ for } i = 0, \ldots M, \quad (4.33)$$

$$t_0 = [0],$$

$$t_i = \boldsymbol{e}(2i+1)'i \text{ for } i = 1, \ldots M.$$

The mean number in the rack indicates the average number of items stored in the rack at any given time. Utilization of the rack is given as $\frac{\mu_{N_r}}{M}$. This performance measure is important for design purposes as it is an indication of storage rack utilization.

A small value indicates low utilization. If utilization is low, the AS/R system may not require such a large rack and for cost reasons a smaller rack size may be more suitable. A high utilization value does not necessarily mean the rack is too small but rather that the system is getting good performance from the particular rack size, unless the jamming probability is high. A high jamming probability would mean that the rack is filling up too quickly and items are not leaving fast enough.

An AS/R system with a high utilization value and a high jamming probability may require a larger rack size if the waiting time and queue length are becoming too large. An efficient AS/R system should have a short waiting time or small queue length, yet be completely utilized.

4.4 Probability of Rack Being Full

The probability that the rack is full, y_F, is given as

$$y_F = \sum_{k=0}^{M} x_{0,M,k} + \sum_{i\geq 1} x_{i,M,0} + \sum_{i\geq 1} \sum_{k=1}^{M} x_{i,M,2,k}. \tag{4.34}$$

Expressing Equation (4.34) in terms of $\boldsymbol{x}_1$ and R gives

$$y_F = \boldsymbol{x}_0 \boldsymbol{b}_{10} + \boldsymbol{x}_1 (I - R)^{-1} \boldsymbol{b}_{11}, \tag{4.35}$$

where

$$\boldsymbol{b}_{10} = [v_0 \; v_1 \; \ldots\ldots \; v_i]' \text{ for } i = 0, \ldots M, \tag{4.36}$$

$$v_i = [0] \text{ for } i = 0, \ldots M-1,$$

$$v_M = e(M+1)',$$

and

$$b_{11} = [w_0 \; w_1 \; \ldots\ldots \; w_i]' \text{ for } i = 0, \ldots M, \tag{4.37}$$

$$w_i = [0] \text{ for } i = 0, \ldots M-1,$$

$$w_M = e(M+1)'.$$

v_i is of size $i+1$ for $i = 1, \ldots, M$, w_i is of size $2i+1$ for $i = 1, \ldots, M-1$ and w_M is of size $i+1$.

4.5 Jamming Probability

The jamming probability, y_J, is the probability that the rack is full with no customers waiting for service in the second queue.

$$y_J = \sum_{i \geq 1} x_{i,M,0}. \tag{4.38}$$

Expressing Equation (4.38) in terms of x_1 and R gives

$$y_J = x_1(I-R)^{-1} e_{M^2+1}. \tag{4.39}$$

Jamming occurs when the storage rack is full, there are no requests for retrieval waiting in Queue 2 and there are items waiting in Queue 1 to be placed into storage. The server is forced to be idle until a request for retrieval arrives into Queue 2 and the server can then resume service. Once an item has been removed from the storage rack the system is no longer jammed and the server can serve Queue 1.

A high jamming value indicates that the rack is frequently filled to its maximum capacity and items are not being removed from storage fast enough to service items arriving to be placed into storage. This suggests that a larger rack size would give better system performance.

4.6 Waiting Time

It is known by Little's law that

$$L = \lambda W, \tag{4.40}$$

where L is the mean number in the system, λ is the arrival rate into the queue, and W is the mean waiting time in the system. Using Little's law, we can derive the equations for the mean waiting time in the system and in the queue for Queue 1.

It is also known from Little's law that the mean number in the system, where the arrival source is finite, is defined as

$$L = \bar{\lambda} W, \tag{4.41}$$

where $\bar{\lambda}$ is the effective arrival rate into the queue. Using Little's law for a finite source, we can derive the equations for the mean waiting time in the system and in the queue for Queue 2.

4.6.1 Mean Waiting Time in the System

Queue 1

For Queue 1, the mean waiting time in the system, W_{1_s} is as follows:

$$W_{1_s} = \frac{\mu_{L_{1_s}}}{\lambda_1}. \tag{4.42}$$

The mean waiting time in the system for Queue 1 indicates the average length of time a customer must wait for service to start from the time the customer arrives in Queue 1.

Queue 2

The effective arrival rate $\bar{\lambda_2}$ is given as

$$\bar{\lambda_2} = \sum_{j=1}^{M}\sum_{k=0}^{j} z_{jk}\lambda_2(j-k), \tag{4.43}$$

$$z_{jk} = \sum_{i}\sum_{j} x_{ijlk}, \tag{4.44}$$

where z_{jk} is the probability that j items are waiting in the rack, of which k are waiting to be served in Queue 2. Since Queue 2 has a bounded source, the expression $\bar{\lambda_2}$ for the effective arrival rate is now used in place of $\bar{\lambda}$ in Equation (4.41). As a result, the mean waiting time in the system for Queue 2, W_{2_s}, is as follows.

$$W_{2_s} = \frac{\mu_{L_{2_s}}}{\bar{\lambda_2}}. \tag{4.45}$$

The mean waiting time in the system for Queue 2 indicates the average length of time a customer must wait for service to start from the time the customer arrives in Queue 2.

4.6.2 Mean Waiting Time in the Queue

Queue 1

From Equation (4.40), the mean waiting time in the queue for Queue 1, W_{1q}, can be obtained as

$$W_{1q} = \frac{\mu_{L_{1q}}}{\lambda_1}. \tag{4.46}$$

The mean waiting time in the queue for Queue 1 indicates the average time a customer will wait in Queue 1 before receiving service, not including the time it takes to serve the current customer.

Queue 2

From Equations (4.41) and (4.43), the mean waiting time in the queue for Queue 2, W_{2q}, is given as

$$W_{2q} = \frac{\mu_{L_{2q}}}{\lambda_2}. \tag{4.47}$$

The mean waiting time in the queue for Queue 2 indicates the average time a customer will wait in Queue 2 before receiving service, not including the time it takes to serve the current customer.

4.7 Effective Arrival Rate into Queue 2

As stated previously, items are placed into and retrieved from storage on an ongoing basis, resulting in a source for arrivals into Queue 2 that is variable of size between 0 and M. The expression for the effective arrival rate $\bar{\lambda}_2$, for a finite source, is given in Equation (4.43). Rewriting Equation (4.43) in detail gives the effective arrival rate, $\bar{\lambda}_2$, as

$$\begin{aligned}\bar{\lambda}_2 = & \sum_{j=0}^{M}\sum_{k=0}^{j} x_{0,j,k}\,\lambda_2\,(j-k) \\ & + \sum_{i\geq 1}\sum_{j=0}^{M} x_{i,j,0}\,\lambda_2\,j \\ & + \sum_{i\geq 1}\sum_{j=1}^{M-1}\sum_{l=1}^{2}\sum_{k=1}^{j} x_{i,j,l,k}\,\lambda_2\,(j-k) \\ & + \sum_{i\geq 1}\sum_{k=1}^{M} x_{i,M,2,k}\,\lambda_2\,(M-k).\end{aligned} \tag{4.48}$$

The expression in Equation (4.48) can be reduced to

$$\bar{\lambda}_2 = \lambda_2(\mu_{N_r} - \mu_{L_{2s}}). \tag{4.49}$$

From numerical computation, it was observed that for a stable system the effective arrival rate, $\bar{\lambda}_2$, was equivalent to the arrival rate into Q1.

$$\bar{\lambda}_2 = \lambda_1. \tag{4.50}$$

For a general system, whether the system is stable or unstable, it is conjectured that the effective arrival rate, $\bar{\lambda}_2$, would not necessarily be equivalent, hence

$$\bar{\lambda}_2 \leq \lambda_1. \tag{4.51}$$

Evidently an item must arrive in Queue 1 before it can arrive in Queue 2. Therefore, the arrival rate into Queue 2 cannot be larger than the arrival rate into Queue 1 for a general system. In the case of a stable system, the arrival rate into Queue 1 equals the effective arrival rate into Queue 2.

Chapter 5

Numerical Examples

This chapter presents numerical examples for hypothetical situations that will be used to assess the behavior of the system. Different rack sizes of 1, 4, 7 and 10 are considered and the arrival and service rates are varied. The numerical examples will provide answers to such questions as:

- What is the performance of the system for various rack sizes?
- What impact on the performance of the system does varying the arrival and service rates have?
- Is it more efficient to increase the rack size versus increasing the service rate of the S/R machine?
- At what point does increasing the size of the rack have very little benefit on system performance?

Several experiments have been conducted to determine the behavior of the model and will be discussed in this chapter. However, before the numerical examples are presented it is important to discuss the validation of the model.

5.1 Validation of Model

Based on the equations presented in Chapter 3, models were developed for rack sizes of 1, 4, 7 and 10 units. For each of the rack sizes the model was validated in the following manner.

The first method for validation of the model is to increase λ_2 and μ_2. As λ_2 and μ_2 are increased towards infinity the limiting behavior of the system can be determined.

For an M/M/1 queueing system the following equations hold:

$$L = \frac{\rho}{1-\rho} = \mu_{L(M/M/1)}, \tag{5.1}$$

$$W = \frac{L}{\lambda}, \tag{5.2}$$

where L is the mean number of customers in the system, W is the mean waiting time in the system, λ is the arrival rate in the queue, μ is the service rate and $\rho = \frac{\lambda}{\mu}$.

For the double-ended queueing system, as both λ_2, and μ_2 tend towards infinity,

$$\lambda_2 \rightarrow \infty \quad \text{and} \quad \mu_2 \rightarrow \infty,$$

the behavior of the double-ended queue tends towards the behavior of an $M/M/1$ system.

$$\mu_{L(DE)} \rightarrow \mu_{L(M/M/1)},$$

where $\mu_{L(DE)}$ is the mean number of customers in the system for the double-ended queue and $\mu_{L(M/M/1)}$ is the mean number of customers in the system for an $M/M/1$ queue.

Let $\mu_{L(DE)}(M)$ be the mean number of customers in the system for Queue 1 for a rack size of M. Also, let $W_{DE}(M)$ be the waiting time for the double-ended queue for rack size M and $W_{M/M/1}$ the waiting time for an $M/M/1$ queue.

Using arrival and service rates of $\lambda_1 = 0.2$, $\mu_1 = 1.0$, $\rho_1 = 0.2$ and $\lambda_2 = 2.0$, the mean number of customers in the system for the double-ended queue was calculated. Tables 5.1 and 5.2 give the queue length and the waiting time for the double-ended system as the service rate in Queue 2 is increased. It is evident that the limiting behavior of the double-ended queueing system is that of an $M/M/1$ queueing system. For a rack size of one, the values of λ_2 and μ_2 must be considerably high before the system behaves as an $M/M/1$ queue.

The second method for validation of the model is to verify that the Lemma for the effective arrival rate, $\bar{\lambda_2}$, discussed in Chapter 4 holds. For all models, the effective arrival rate $\bar{\lambda_2}$ does equal λ_1.

$\lambda_1 = 0.2, \qquad \mu_1 = 1.0$

Queue Length		$\lambda_2 : \mu_2$				
		2 : 6.67	2 : 20	2 : 40	2 : 200	2 : 400
$M/M/1$	$\mu_{L(M/M/1)}$	0.25				
Double-Ended Queue	$\mu_{L(DE)}(1)$	0.4492	0.4159	0.4079	0.4016	0.4008
	$\mu_{L(DE)}(4)$	0.2545	0.2512	0.2506	0.2501	0.2501
	$\mu_{L(DE)}(7)$	0.2545	0.2512	0.2506	0.2501	0.2500
	$\mu_{L(DE)}(10)$	0.2545	0.2512	0.2506	0.2501	0.2500

		$\lambda_2 : \mu_2$				
		2 : 800	20 : 8000	200 : 8000	2000 : 8000	5000 : 8000
	$\mu_{L(DE)}(1)$	0.4004	0.2633	0.2513	0.2502	0.2501

Table 5.1: Queue Length Validation of Model

$\lambda_1 = 0.2, \qquad \mu_1 = 1.0$

Queue Length		$\lambda_2 : \mu_2$				
		2 : 6.67	2 : 20	2 : 40	2 : 200	2 : 400
$M/M/1$	$W_{M/M/1}$	1.25				
Double-Ended Queue	$W_{DE}(1)$	2.2461	2.0797	2.0396	2.0079	2.0039
	$W_{DE}(4)$	1.2725	1.2560	1.2528	1.2505	1.2503
	$W_{DE}(7)$	1.2724	1.2560	1.2528	1.2505	1.2502
	$W_{DE}(10)$	1.2725	1.2560	1.2528	1.2505	1.2502

		$\lambda_2 : \mu_2$				
		2 : 800	20 : 8000	200 : 8000	2000 : 8000	5000 : 8000
	$W_{DE}(1)$	2.0020	1.3166	1.2567	1.2508	1.2504

Table 5.2: Waiting Time Validation of Model

5.2 Experiments

The first set of experiments determines the behavior of the system as a function of the arrival rate in Queue 1 where λ_1 and μ_2 were varied. Since the behavior for each rack size modeled is similar, only the behavior for a rack size of $M = 4$ will be discussed.

The second set of experiments fixes the arrival and service rates and determines the behavior of the system as the rack size increases.

For both sets of experiments the performance measures discussed in Chapter 4 were obtained including the mean number of customers in the system, the waiting time, and the jamming probability.

5.2.1 Varying Arrival and Service Rates

This section presents the results for the experiments in which both λ_1 and μ_2 have been varied. The service rate for Queue 1 and the arrival rate into Queue 2 have been fixed at $\mu_1 = 1.1$ and $\lambda_2 = 2.0$.

Queue Length

The mean queue length in the system for Queue 1 is shown in Figure 5.1. The mean queue length in the system for Queue 1 is as expected. The graph shows that as traffic into Queue 1 increases, the queue length in Queue 1 increases exponentially. As the arrival rate increases in Queue 1, the queue length grows quickly. Notice that the asymptote of the curve shifts to the left, indicating that the system will become unstable faster, as the service rate decreases in Queue 2.

The mean queue length in the system for Queue 2 is shown in Figure 5.2. The graph shows that as traffic into Queue 1 increases, the queue length in Queue 2 increases exponentially. The same type of behavior that was observed for Queue 1 is

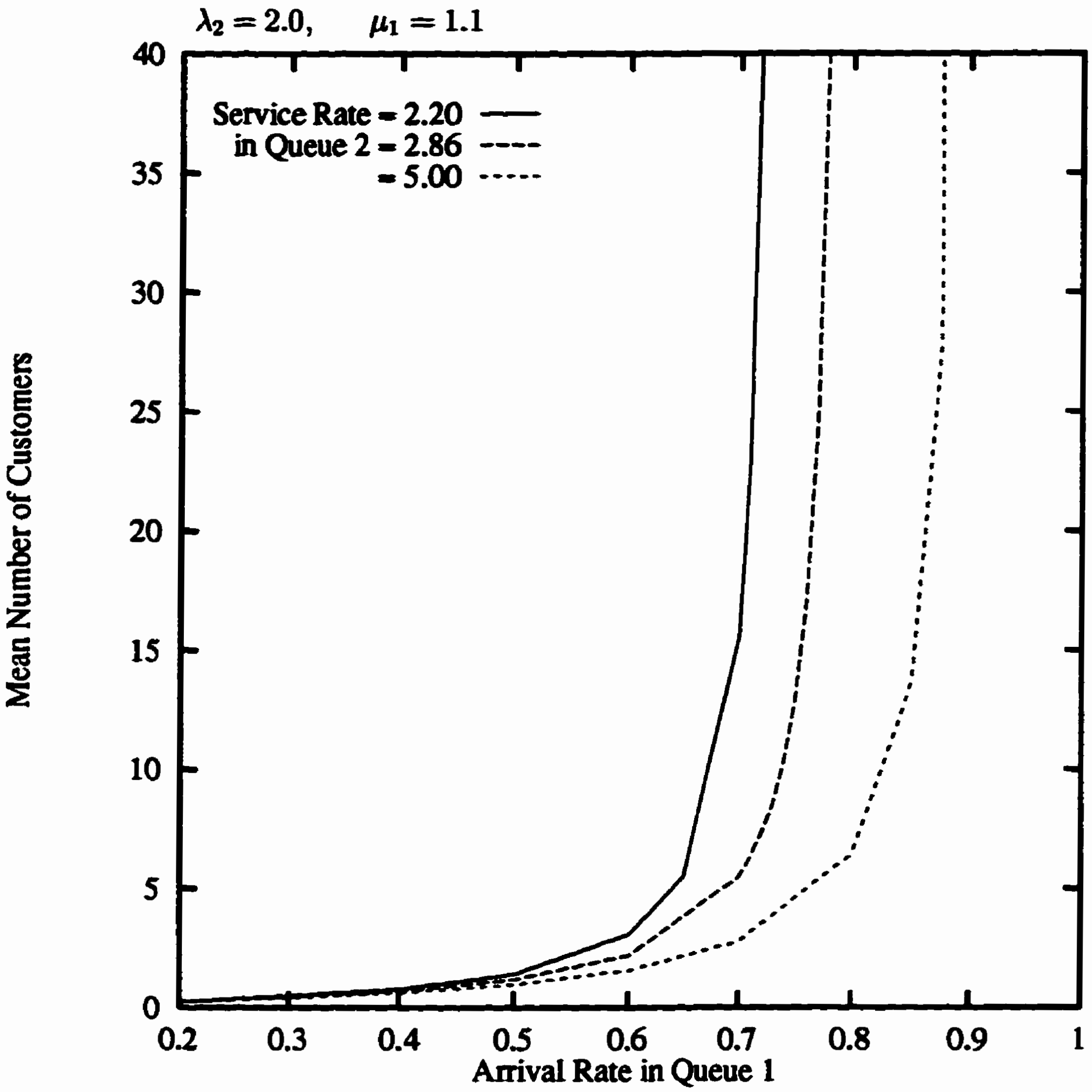

Figure 5.1: Mean Queue Length in Queue 1 - $M = 4$ ($\mu_1 < \mu_2$)

also observed for Queue 2. The key difference is that in the case of Queue 2 the slope of the curve is much gentler and approaches infinity slower.

Waiting Time

The mean waiting time in the system for Queue 1 is shown in Figure 5.3. The graph shows that as traffic into Queue 1 increases, the waiting time in Queue 1 increases exponentially. The behavior of this graph is similar to that of the queue length in both Queue 1 and 2 since the asymptote of this curve also shifts to the left, indicating that the system will become unstable faster, as the service rate decreases in Queue 2.

The mean waiting time in the system for Queue 2 is shown in Figure 5.4. The graph shows that as traffic into Queue 1 increases, the waiting time in Queue 2 increases exponentially. The behavior of this graph is of the same type since the asymptote of this curve shifts to the left, indicating that the system will become unstable faster, as the service rate decreases in Queue 2.

Mean Number of Items in the Rack

The mean number of items in the rack is shown in Figure 5.5. The graph shows that as traffic into Queue 1 increases, the mean number of items in the rack increases. Notice that for a rack size of $M = 4$, the mean number of customers in the rack does not reach capacity.

Utilization of the rack as previously discussed is given as $\frac{\mu N_r}{M}$. Low utilization would indicate that there is room for increasing the use of the storage facility, or that a smaller rack would be more cost efficient. A high utilization would indicate that the system is being used efficiently, however, there is little room for increasing the utilization. If the system requires extra capacity for future expansion, a larger rack

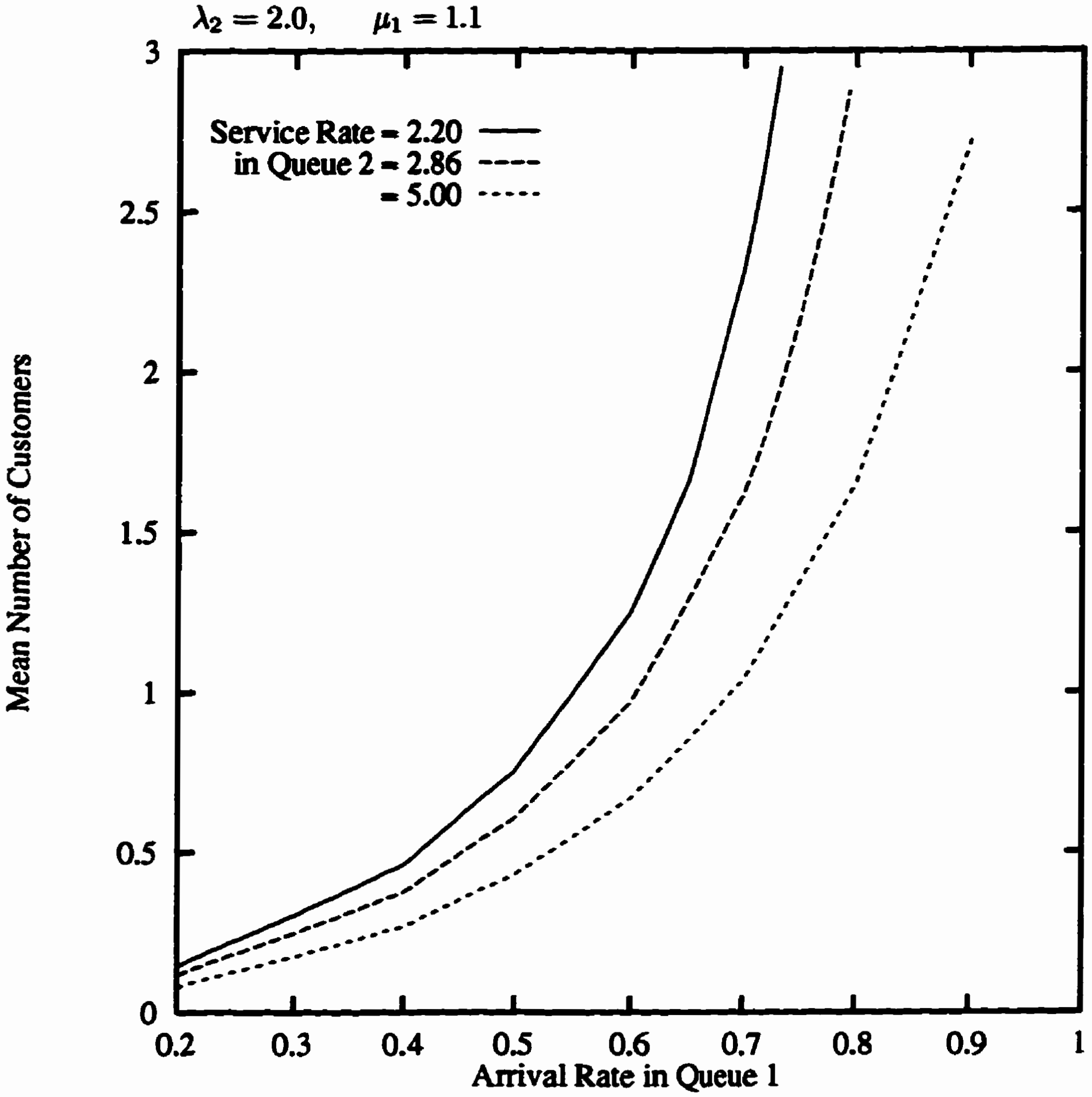

Figure 5.2: Mean Queue Length in Queue 2 - $M = 4$ ($\mu_1 < \mu_2$)

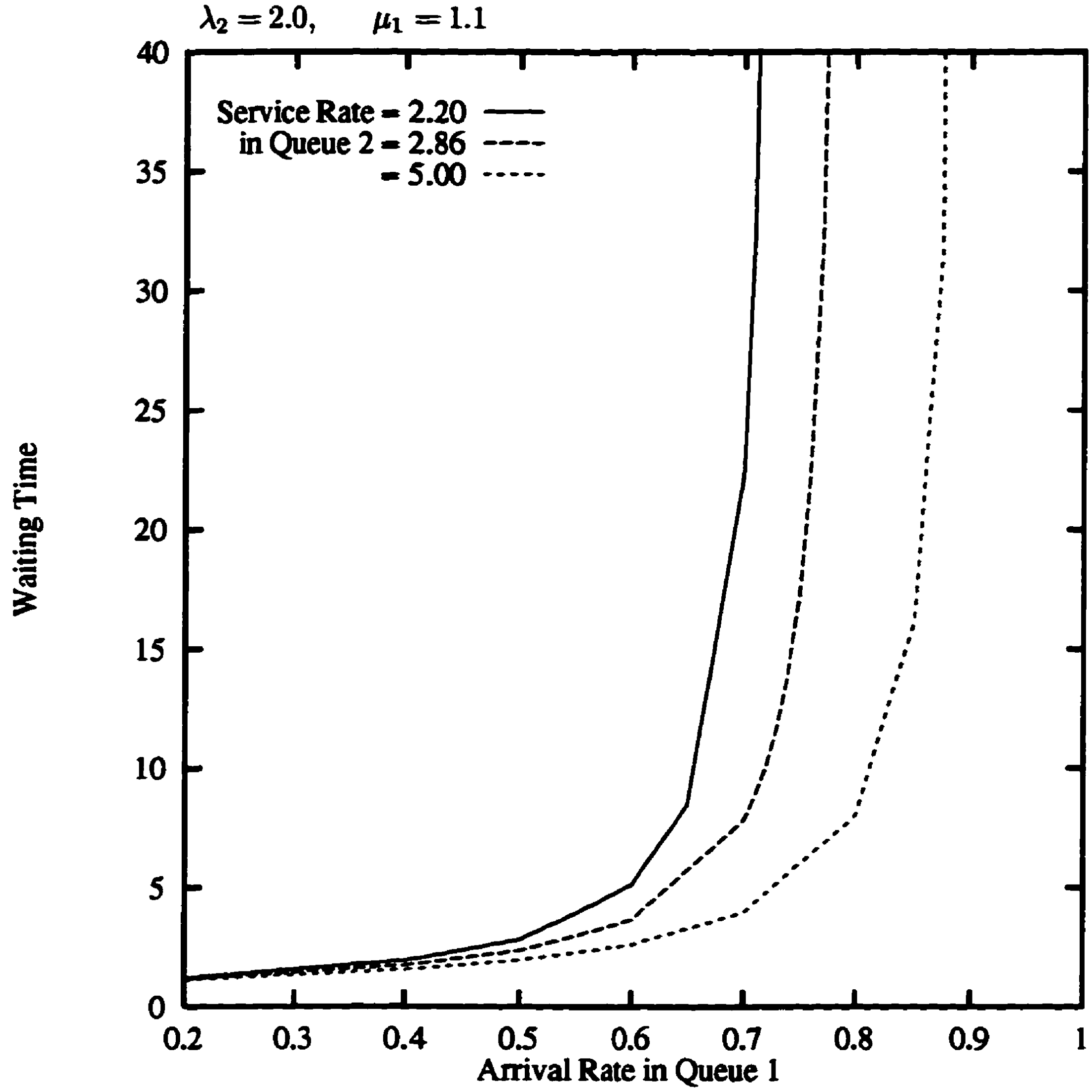

Figure 5.3: Mean Waiting Time in Queue 1 - $M = 4$ ($\mu_1 < \mu_2$)

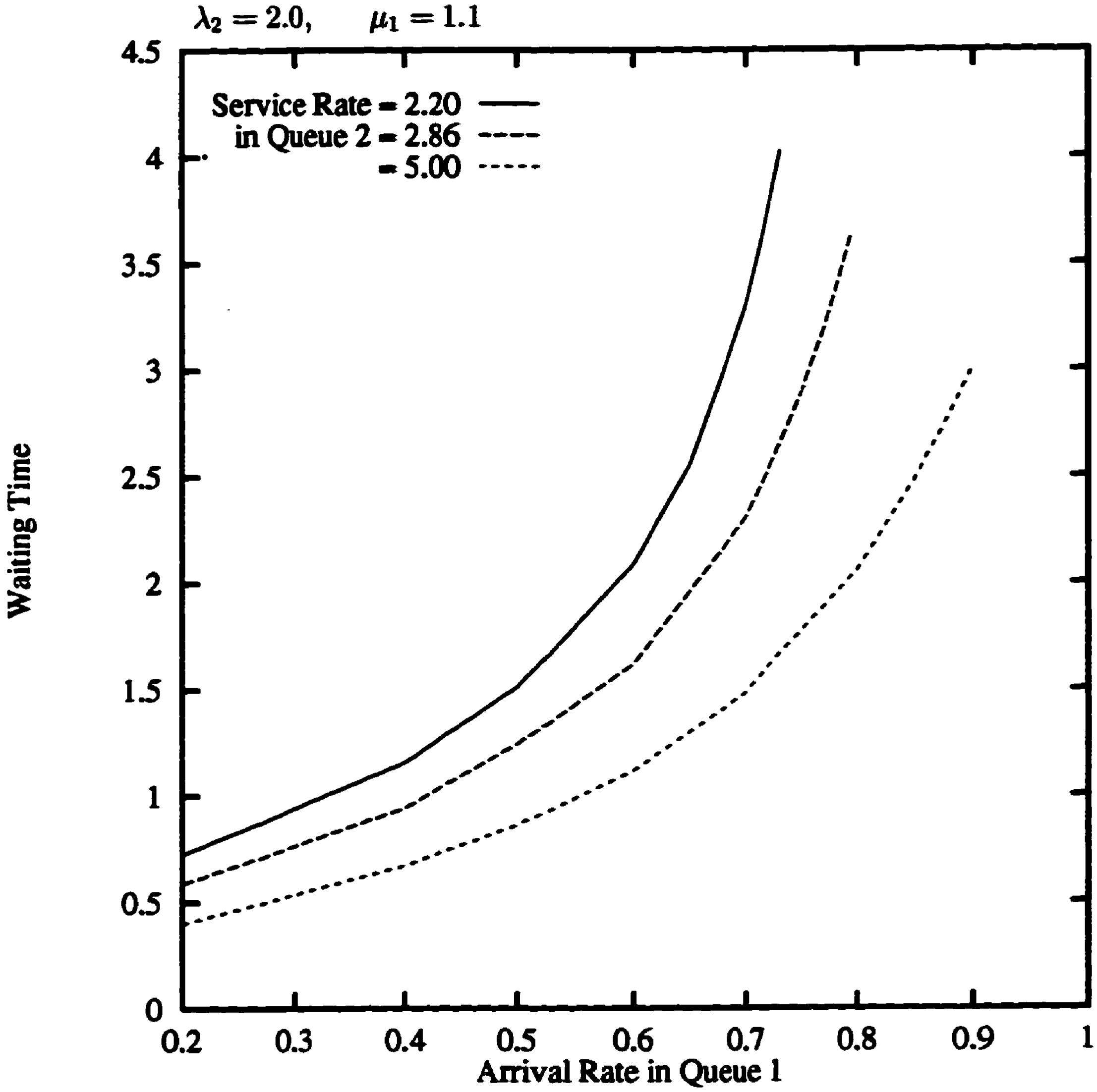

Figure 5.4: Mean Waiting Time in Queue 2 - $M = 4$ ($\mu_1 < \mu_2$)

size might be better.

Probability of Server Being Idle

The probability that the server is idle is shown in Figure 5.6. The graph shows that as traffic into Queue 1 increases, the probability that the server is idle decreases. The probability that the server is idle decreases in a straight line. We also observe that as the service rate in Queue 2 increases, the probability that the server is idle increases.

In the case of $M/M/1$, the probability that the server is idle is given as

$$1 - \rho = 1 - \frac{\lambda}{\mu}. \tag{5.3}$$

As λ changes, the curve will actually be a straight line. However, even with jamming included, the probability that the server is idle is still a straight line. Although it is not evident from Figure 5.3, the probability that the server is idle never reaches zero.

Probability of Jamming

The probability that the system becomes jammed has been studied and several graphs are presented to show the behavior of the system for different parameters.

Figure 5.7 shows the probability of jamming for a rack size of one, $M = 1$, which behaves as expected. It is observed that as the arrival rate into Queue 1 increases, the probability that the system becomes jammed increases. As the service rate in Queue 2 increases, the probability that the system becomes jammed decreases. This graph is for the case where $\mu_1 < \mu_2$. However, for the cases where $\mu_1 = \mu_2$ and $\mu_1 > \mu_2$ the behavior is similar. For rack sizes of $M = 4, 7$ and 10, this is also true in some instances and then the curves cross over each other. Once this occurs,

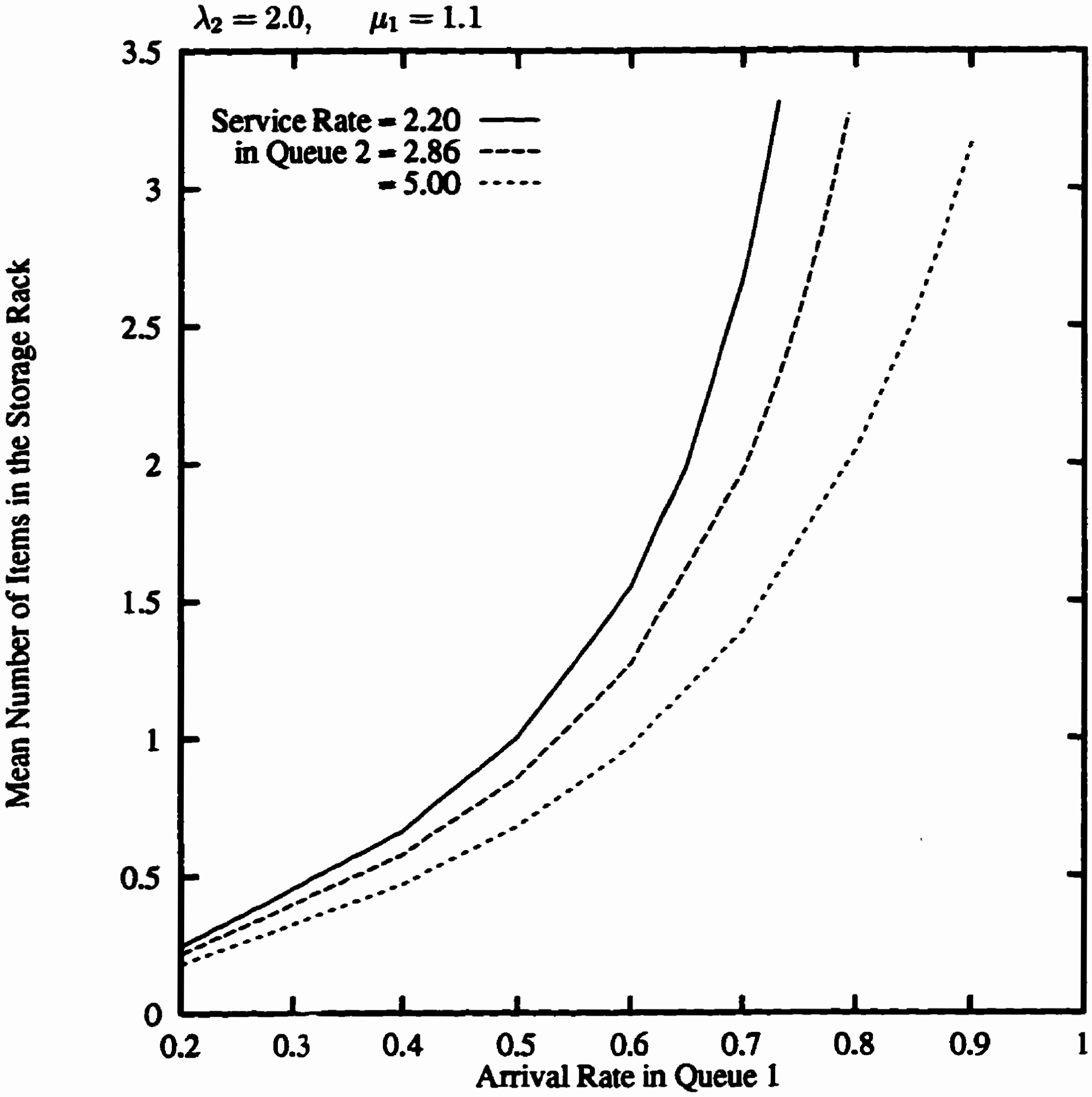

Figure 5.5: Mean Number of Items in the Storage Rack - $M = 4$ ($\mu_1 < \mu_2$)

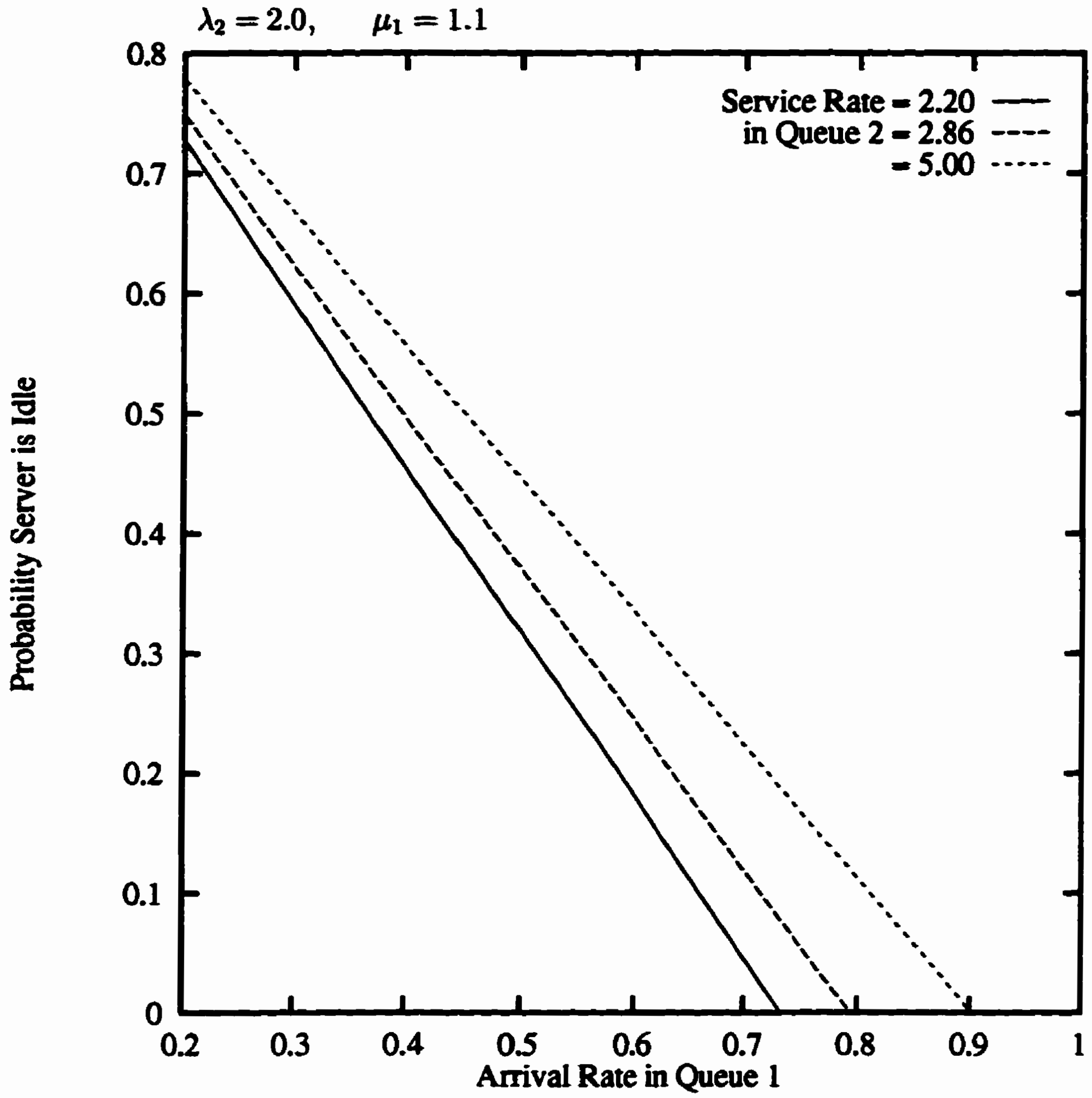

Figure 5.6: Probability the Server is Idle - $M = 4$ ($\mu_1 < \mu_2$)

the jamming probability increases as the service rate in Queue 2 increases. This is counter-intuitive. It is also noted that the cross over points shift further to the right as the rack size increases.

Figures 5.8, 5.9 and 5.10 show the probability of jamming for rack sizes of $M = 4, 7$ and 10 when $\mu_1 < \mu_2$. Figures 5.11, 5.12 and 5.13 show the probability of jamming when $\mu_1 > \mu_2$. Figures 5.14, 5.15 and 5.16 show the probability of jamming when $\mu_1 = \mu_2$. These graphs show cross over points in the curves that change as the rack size increases.

It is also observed that in some instances before the system becomes unstable, the jamming probability decreases as λ_1 increases. This decrease is also counter-intuitive. As pointed out, the increase in jamming probability as μ_2 increases and the reduction in the jamming probability as λ_1 increases are counter-intuitive. The best explanation that can be offered at the present has to do with the relation to ρ_2. ρ_2, defined as λ_2/μ_2, is the level of traffic intensity of Queue 2 and is a representation of how traffic is leaving the rack compared to the service rate. When ρ_2 is low, there are few items waiting to be removed from storage. A low ρ_2 will cause the jamming probability to increase in some instances. Alternately, when ρ_2 is high, there are many customers waiting to be removed from the rack, and the jamming probability is low. The amount of customers in the storage rack is dependent on λ_1 and μ_1. When λ_1 is small there is not necessarily a high probability of jamming as there are fewer items being placed into the rack. However, when λ_1 is large there is a high probability that there is jamming. Note that a decrease in jamming probability does not necessarily mean a better system as the queue length and waiting time are now large. This is only one possible explanation.

There are several factors that affect the jamming probability of a double-ended queueing system. Further study is required to fully understand how the system behaves.

5.2.2 Varying Size of Storage Rack (M)

This section presents results for experiments where μ_1, λ_2 and μ_2 have been fixed at $\mu_1 = 1.1$, $\lambda_2 = 2.0$ and $\mu_2 = 2.2$ and λ_1 has been varied for different rack sizes.

Figure 5.17 shows that for storage sizes of 4, 7 and 10 the difference in the behavior of the system is negligible. Under certain operating parameters the performance of rack sizes $M = 4, 7$ and 10 will be the same. However, under different operating parameters, there will be a difference in the performance of the system for the various rack sizes. Figure 5.18 shows a difference in the behavior of the system for rack sizes of 7 and 10. The performance measures obtained for these experiments behave in the same manner as discussed previously.

In regards to Figure 5.18, even though the system may have an improvement in performance with the increased storage rack size, the benefits of the additional size may not outweigh the additional investment required for the increased capacity.

Figure 5.19 shows the mean number of customers in the system for rack sizes of $M = 4, 7$ and 10. Notice that the queue length in Queue 1 decreases and eventually levels off indicating that further increases in rack size will not bring additional performance improvements.

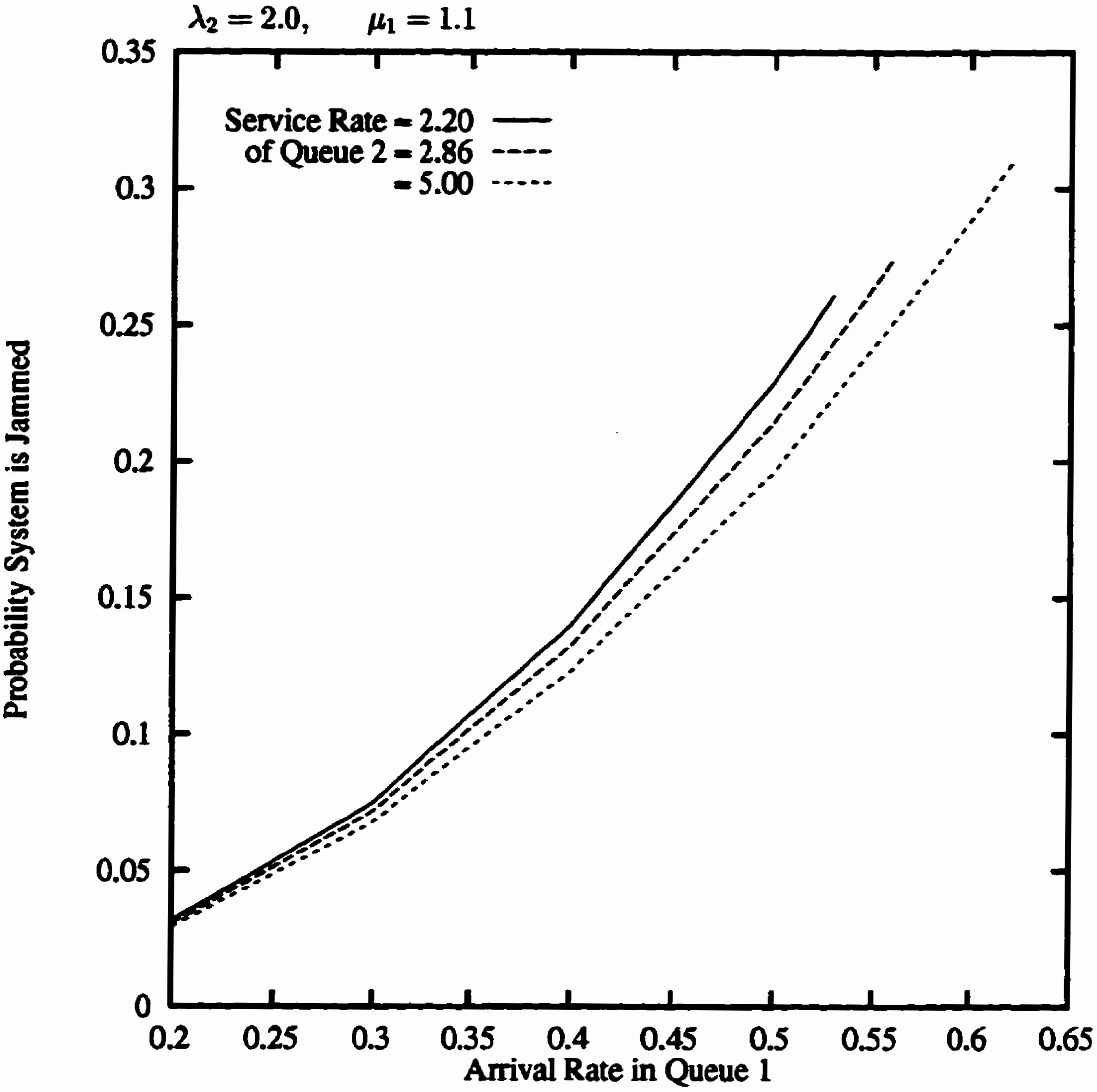

Figure 5.7: Probability the System is Jammed - $M = 1$ ($\mu_1 < \mu_2$)

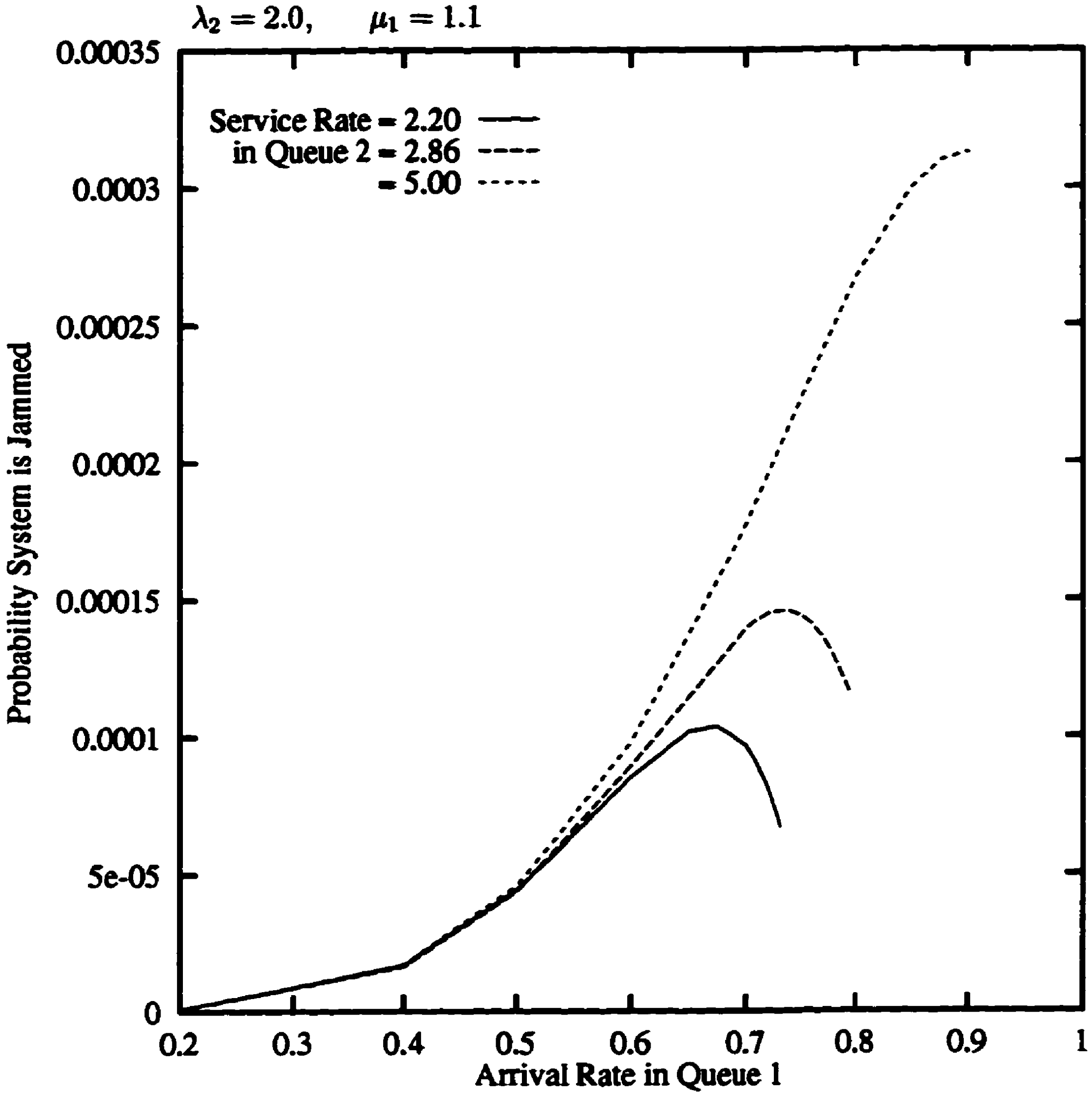

Figure 5.8: Probability the System is Jammed - $M = 4$ ($\mu_1 < \mu_2$)

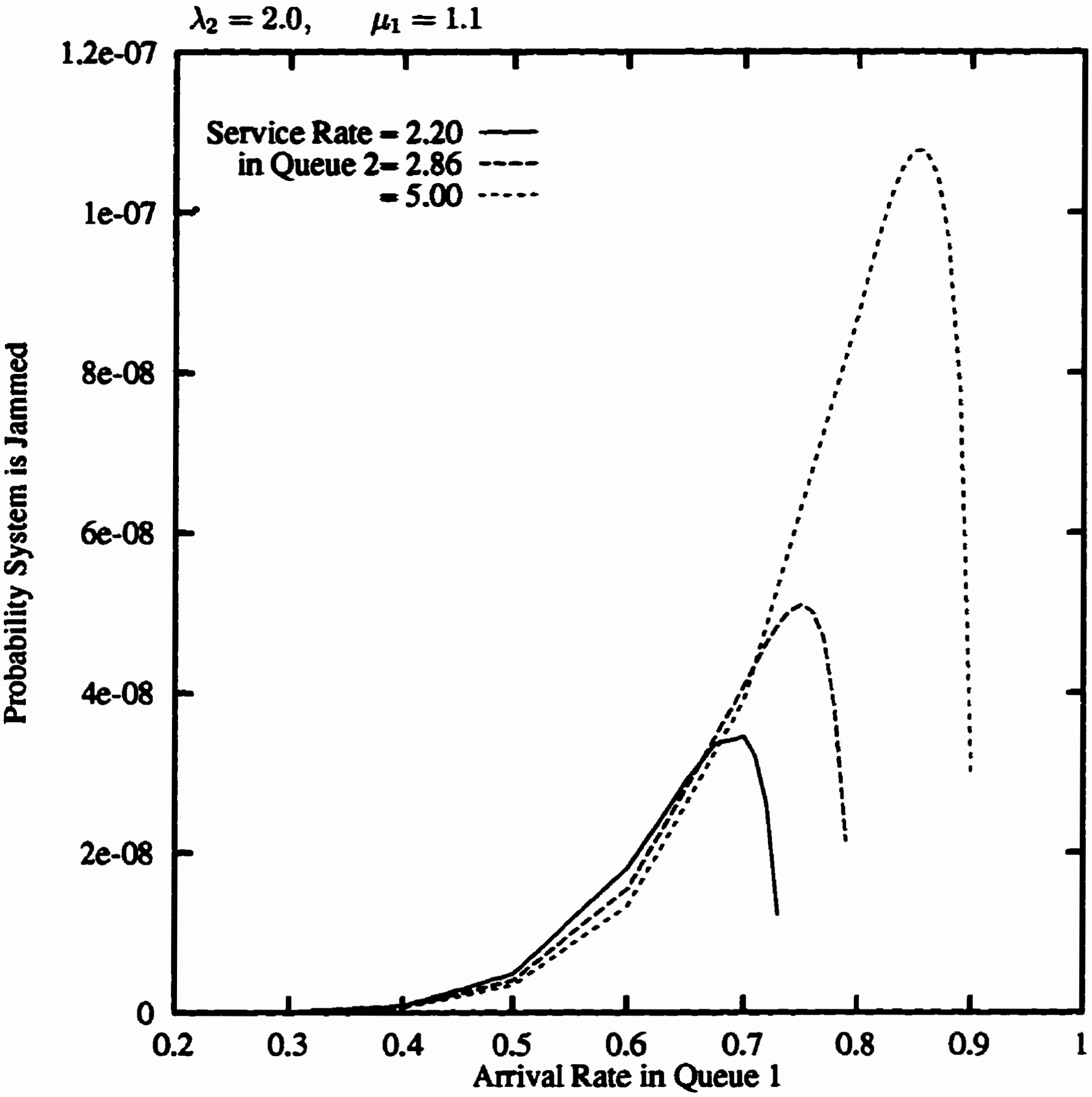

Figure 5.9: Probability the System is Jammed - $M = 7$ ($\mu_1 < \mu_2$)

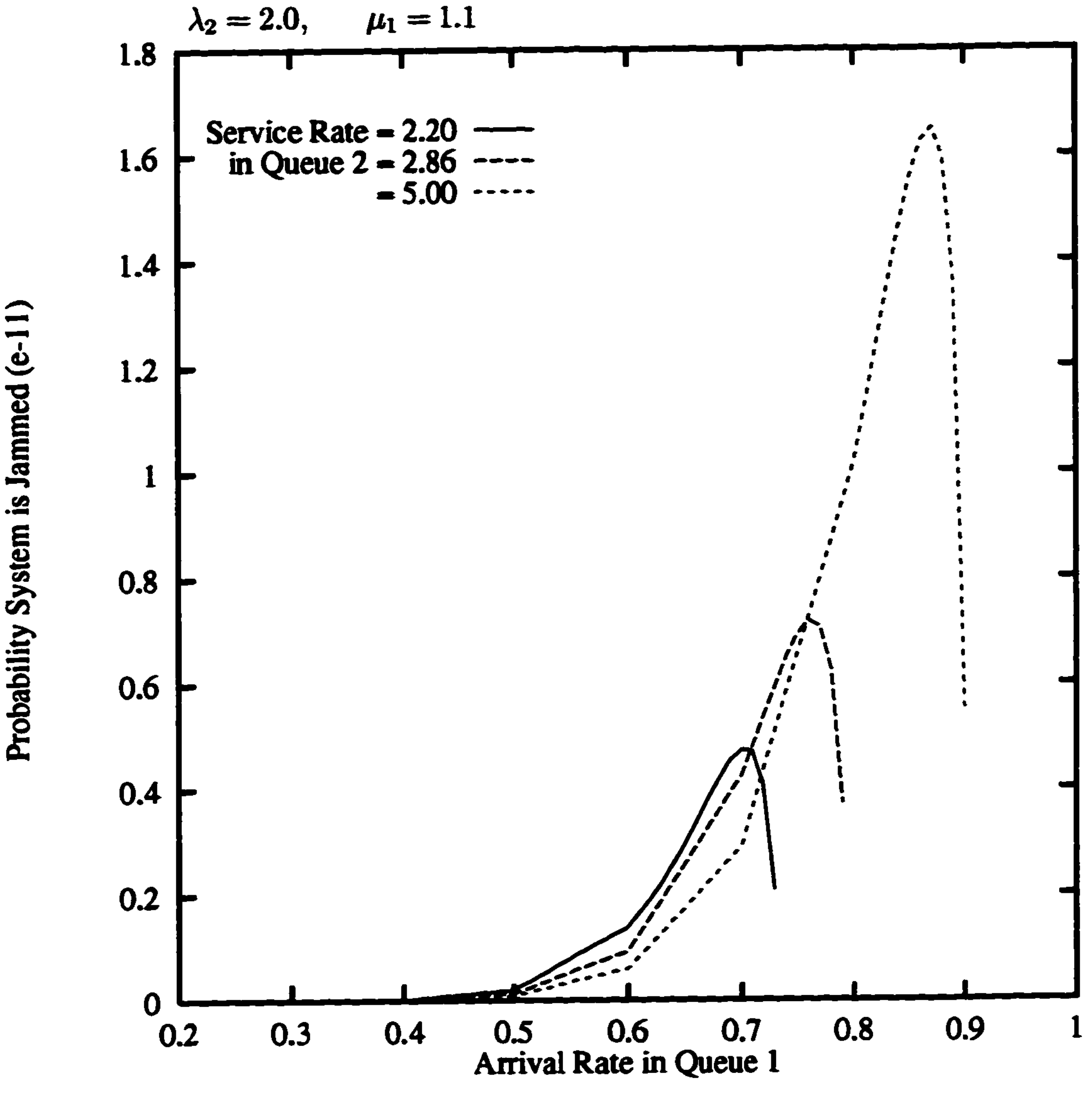

Figure 5.10: Probability the System is Jammed - $M = 10$ $(\mu_1 < \mu_2)$

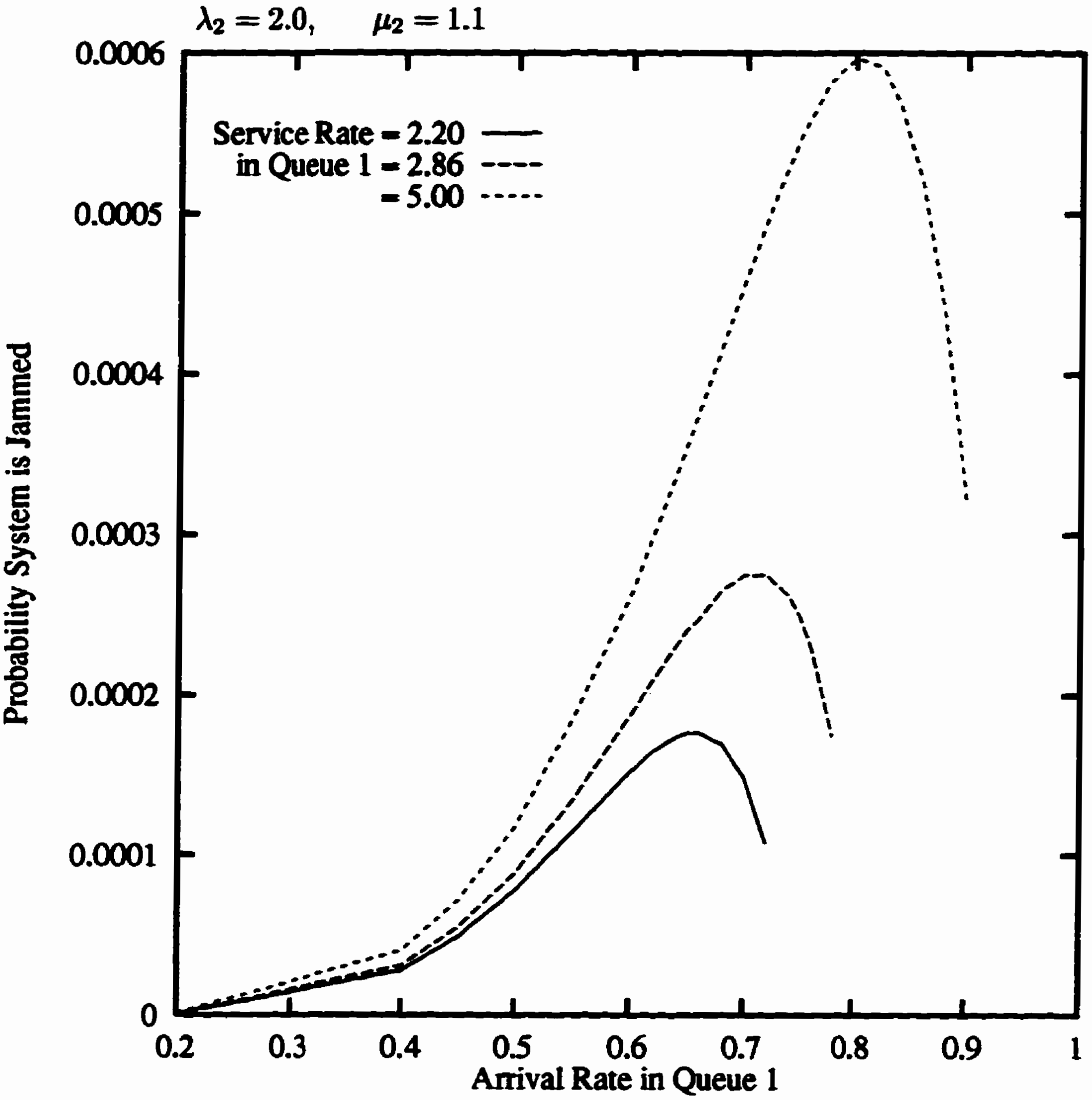

Figure 5.11: Probability the System is Jammed - $M = 4$ ($\mu_1 > \mu_2$)

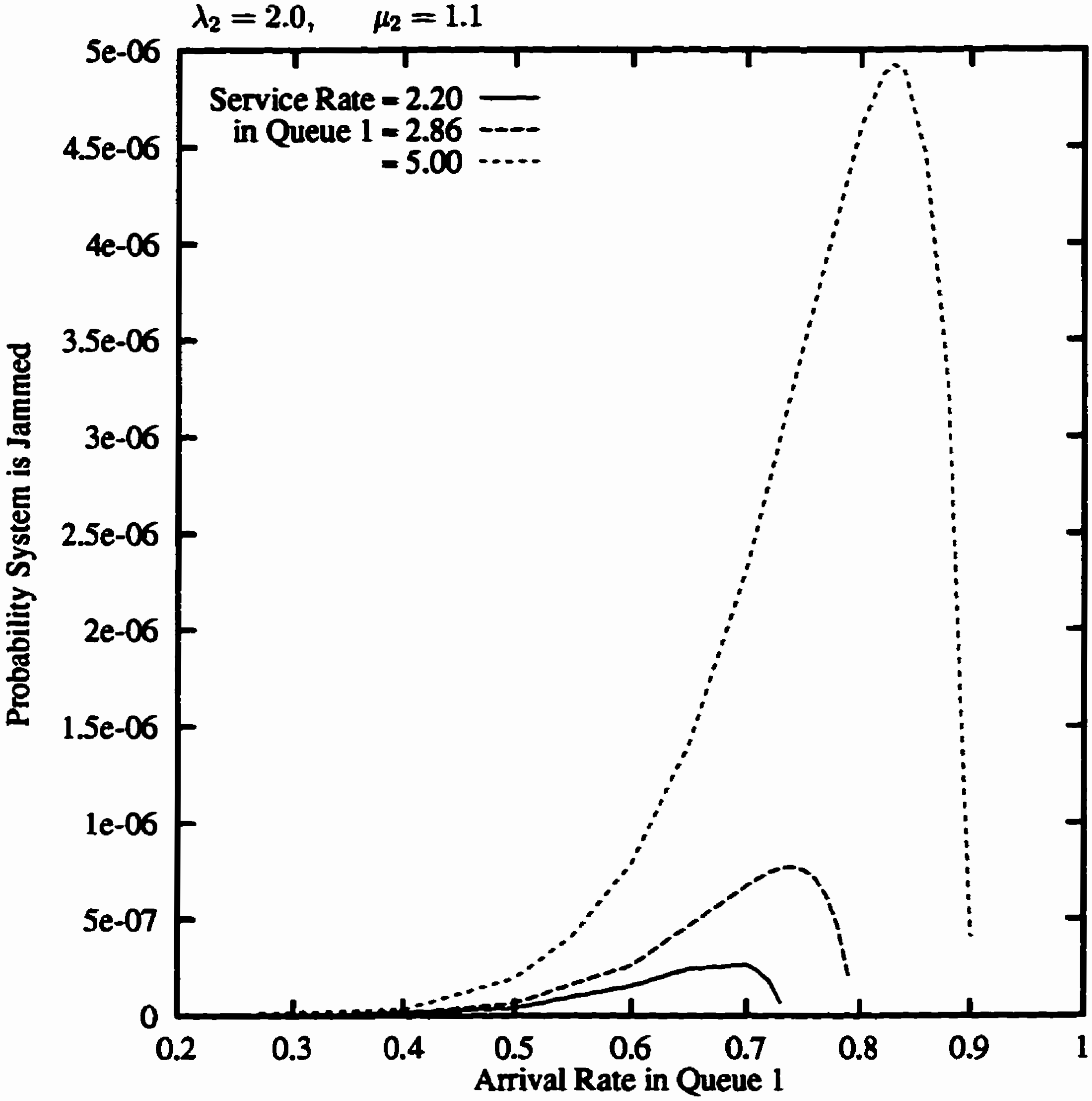

Figure 5.12: Probability the System is Jammed - $M = 7$ ($\mu_1 > \mu_2$)

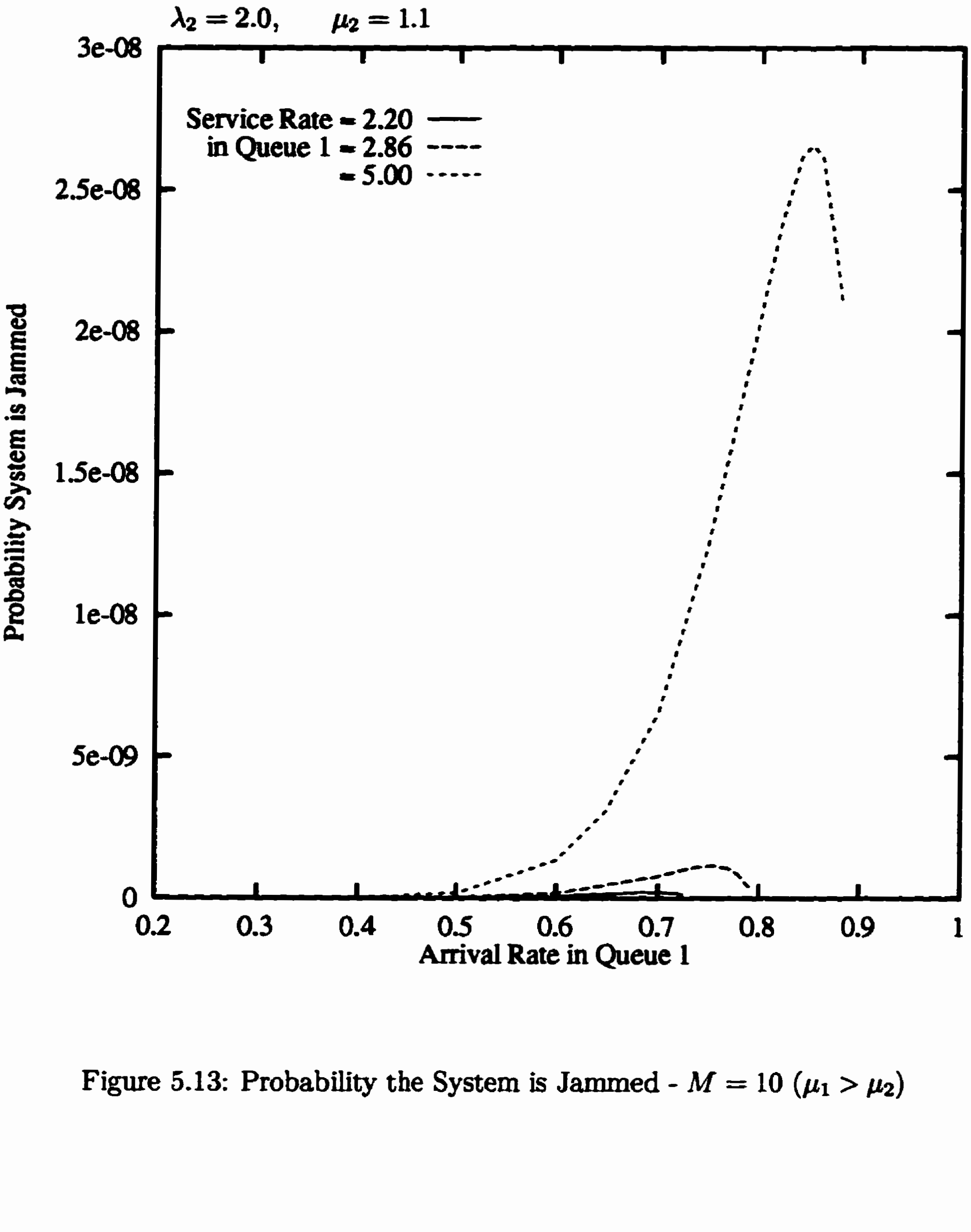

Figure 5.13: Probability the System is Jammed - $M = 10$ ($\mu_1 > \mu_2$)

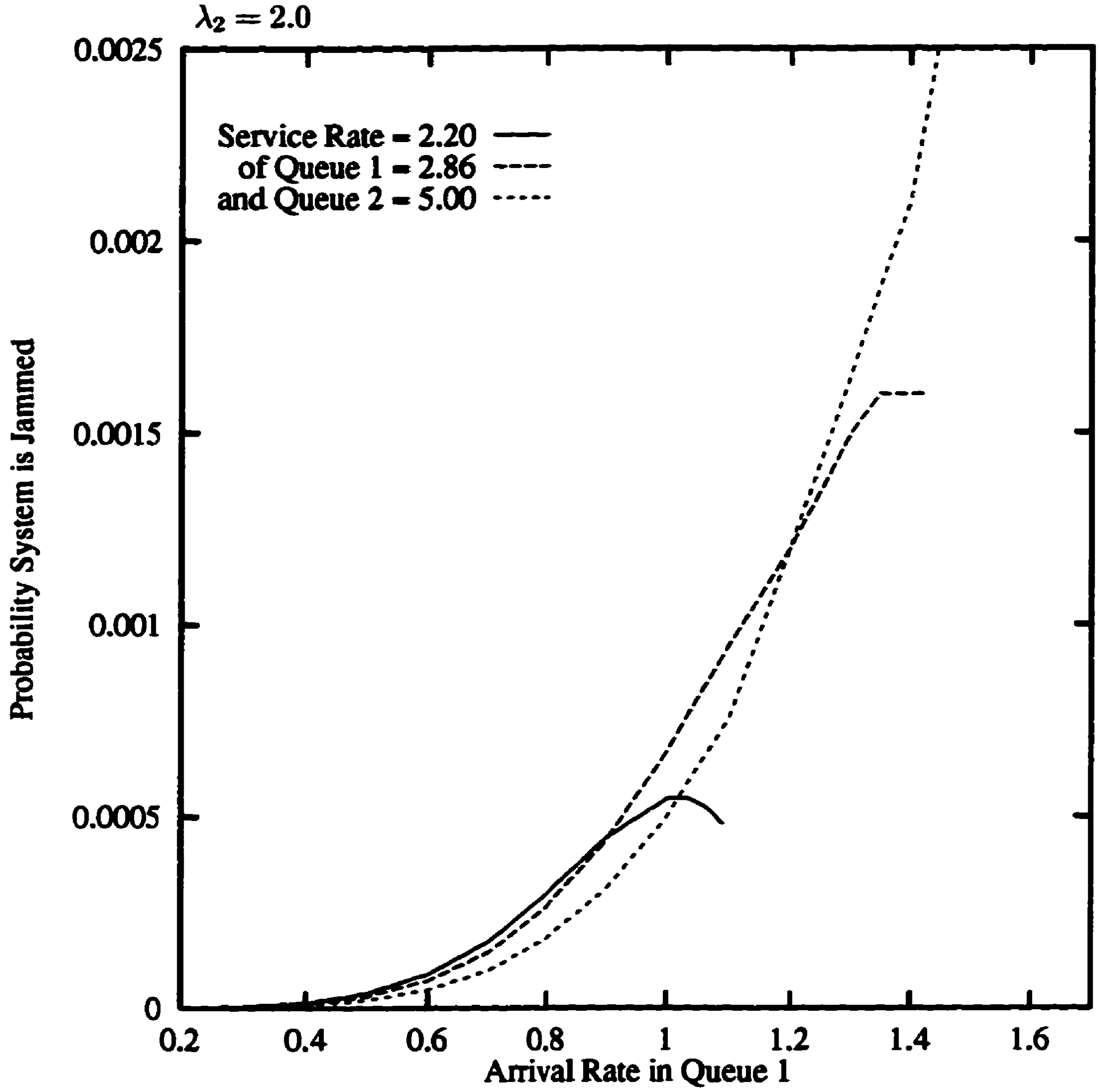

Figure 5.14: Probability the System is Jammed - $M = 4$ $(\mu_1 = \mu_2)$

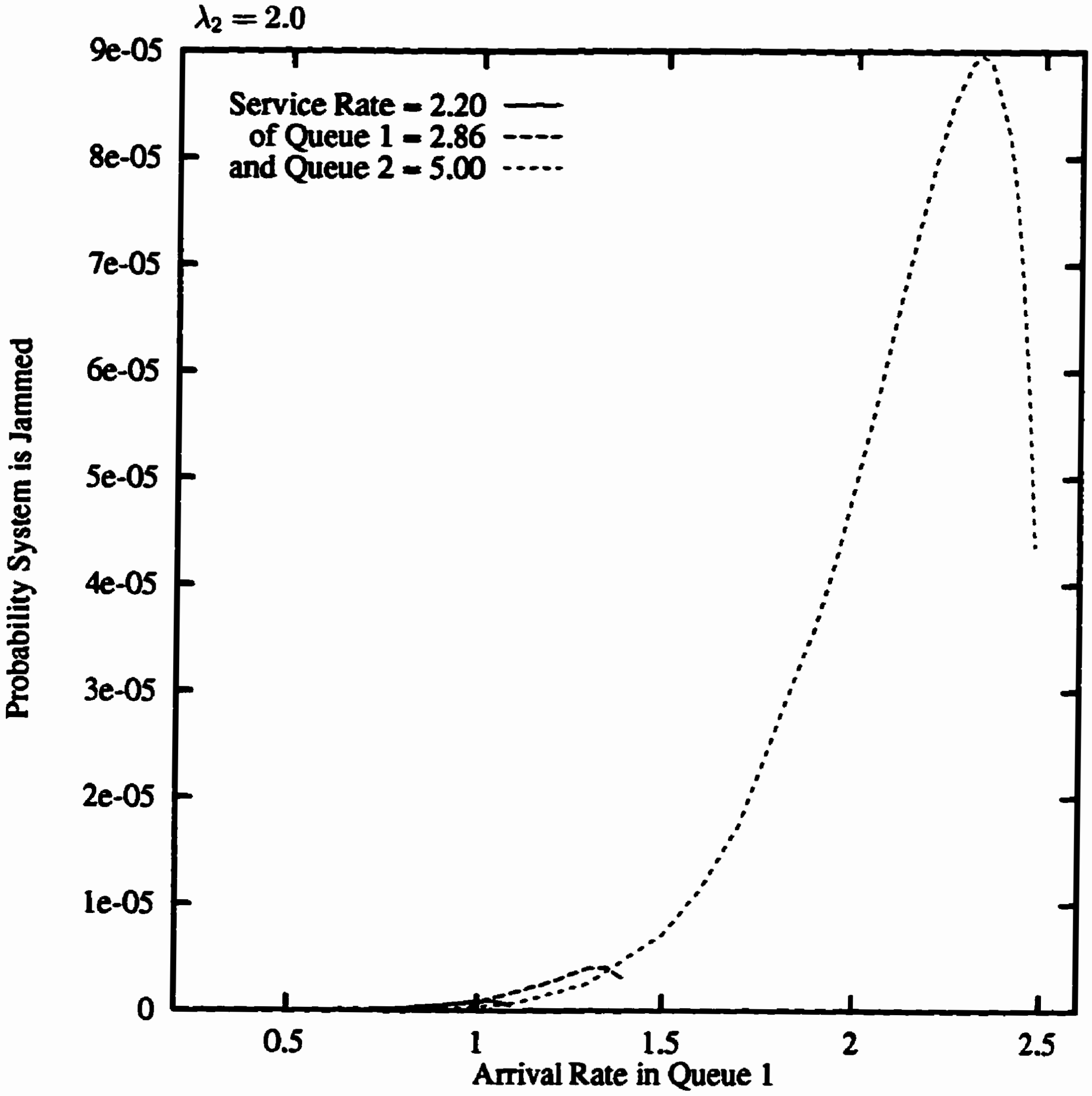

Figure 5.15: Probability the System is Jammed - $M = 7$ ($\mu_1 = \mu_2$)

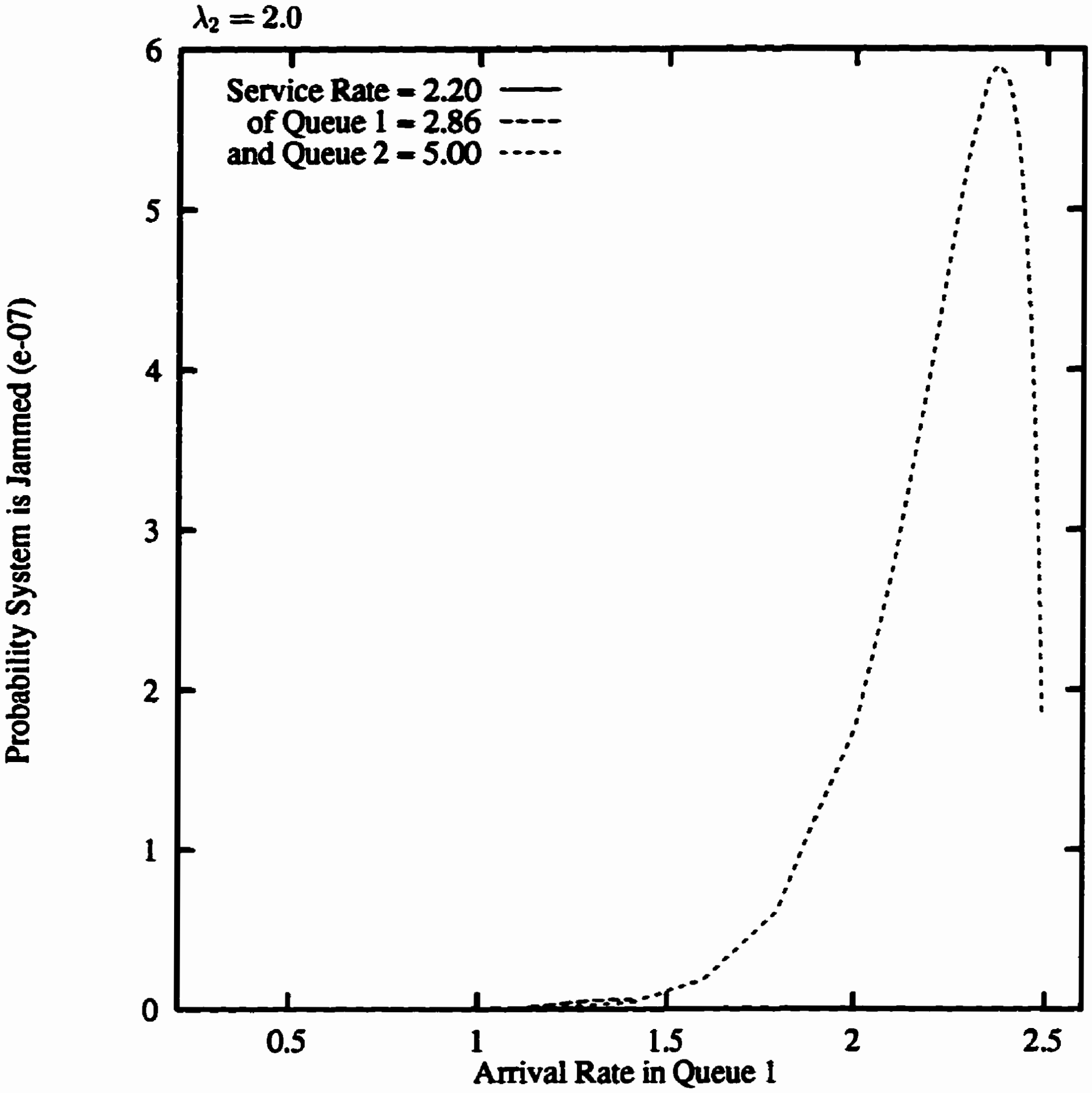

Figure 5.16: Probability the System is Jammed - $M = 10$ ($\mu_1 = \mu_2$)

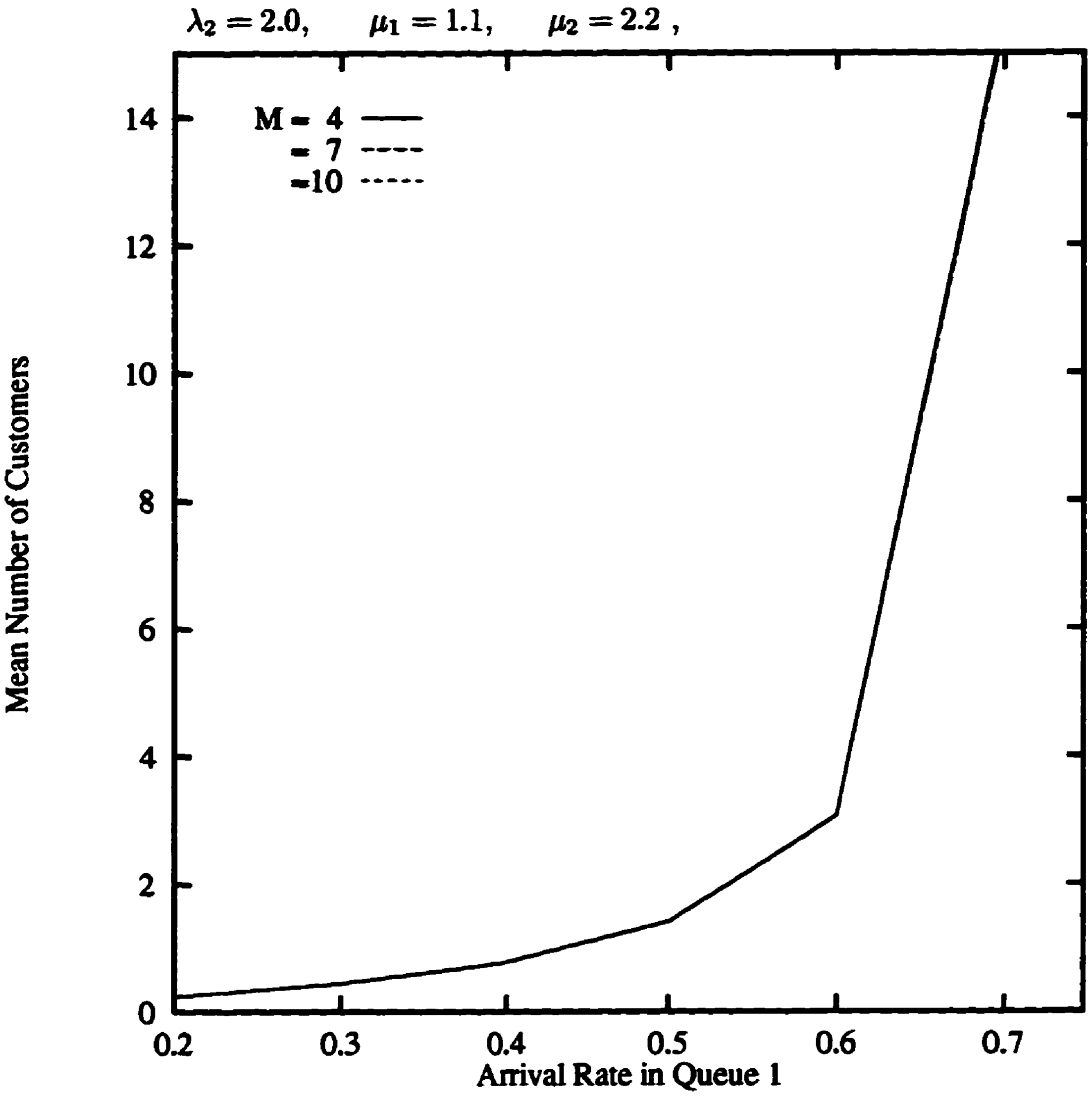

Figure 5.17: Mean Queue Length in Queue 1 - $M = 4, 7, 10$ (μ_1 and μ_2 low)

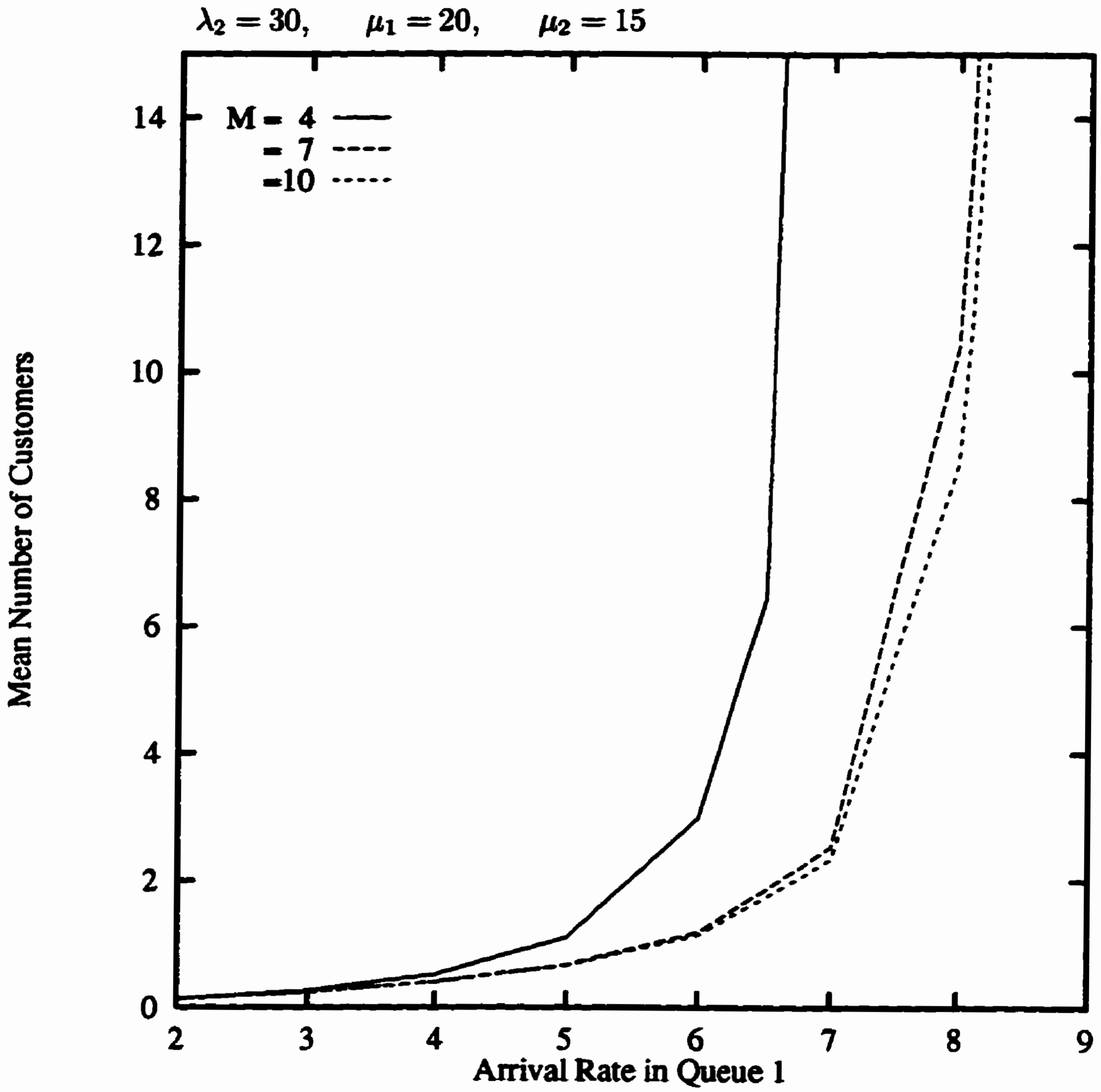

Figure 5.18: Mean Queue Length in Queue 1- $M = 4, 7, 10$ (μ_1 and μ_2 high)

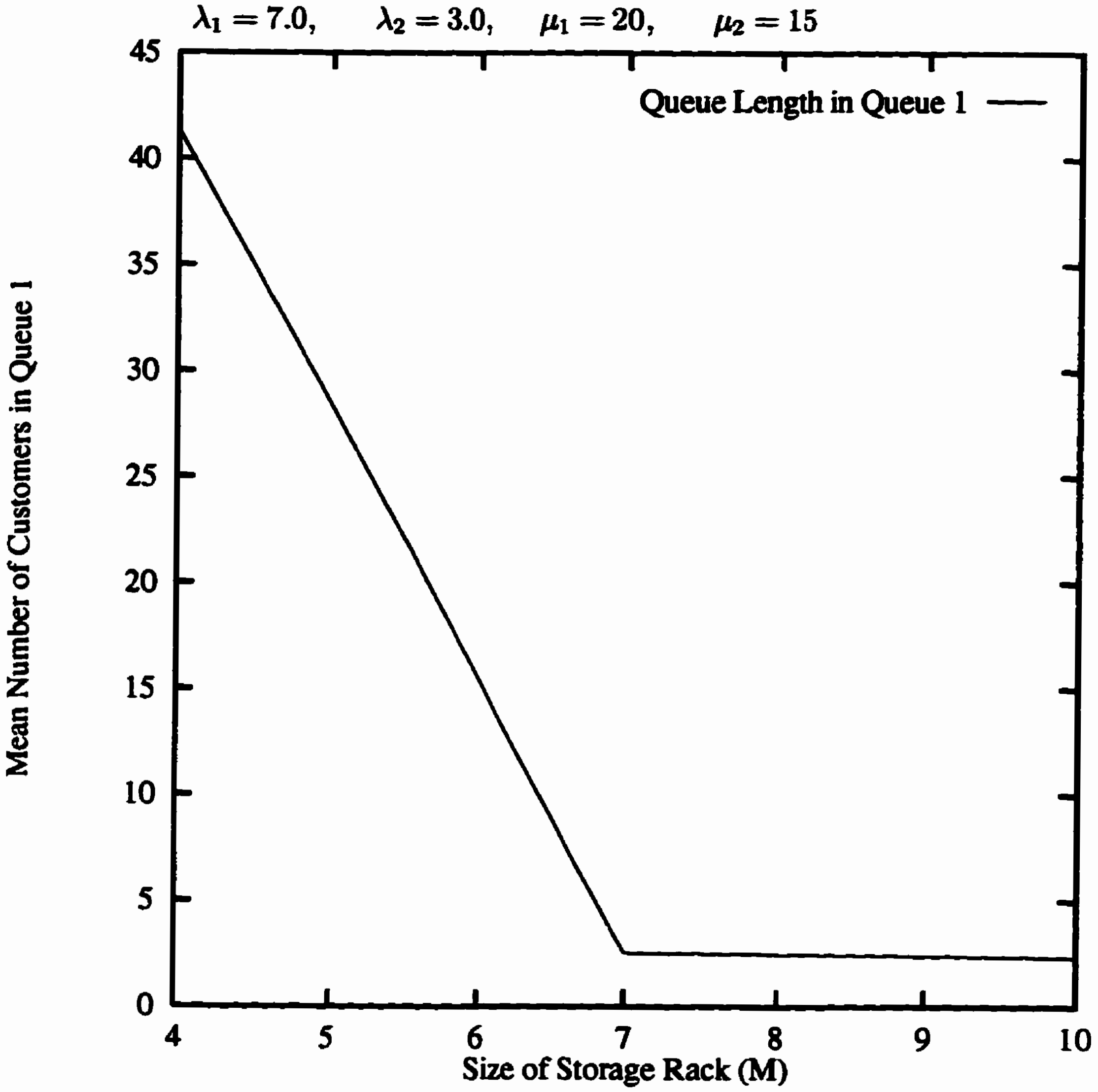

Figure 5.19: Mean Queue Length in Queue 1 for Size of Rack

5.3 How to Use The Results for Design Purposes

In most AS/R systems the arrival rates into the first and second queue will be fixed and not usually controllable. Also, in most cases the service rate for both Queue 1 and 2 will be the same, $\mu_1 = \mu_2 = \mu$, but unknown. The designer will have to select the appropriate values of μ_1 and μ_2. The interest is to find the service rate which best suits the rack size of M, and the rack size of M which is most appropriate given specific service requirements.

Let $C_w(\cdot)$ be the cost associated with delay of an item to be placed into and removed from storage, $C_M(\cdot)$ be the cost associated with a specific rack size M, and $C_\mu(\cdot)$ be the cost of an S/R machine that provides service at a rate of μ.

Given a certain waiting time in Queue 1, $W_{1_s}(\mu, M)$, for a specific rack size M and service rate μ, the minimum total cost of an AS/R system can be determined by

$$\text{Minimize} \quad Z = C_w(W_{1_s}(\mu, M)) + C_M(M) + C_\mu(\mu) \tag{5.4}$$

$$\text{Subject to:} \quad \mu > 0 \tag{5.5}$$

$$M \geq 1 \tag{5.6}$$

$$f(s) > 0 \tag{5.7}$$

where

$$f(s) = \boldsymbol{\pi} A_2 \mathbf{e} - \boldsymbol{\pi} A_0 \mathbf{e}. \tag{5.8}$$

Stability is conjectured to be

$$f(s) = \left(\frac{1}{\lambda_1}\right) - \left(\frac{1}{\mu_1}\right) - \left(\frac{1}{\mu_2}\right) - \left(\frac{1}{M\lambda_2}\right) \prod_{j=0}^{M-1} \left(\frac{\mu_1}{\mu_1 + j\lambda_2}\right) \prod_{i=0}^{M-1} \left(\frac{\mu_2}{\mu_2 + i\lambda_2}\right), \quad (5.9)$$

and can be determined numerically.

Chapter 6

Summary and Conclusions

The contribution of this thesis is the development of a queueing model for an automatic storage and retrieval system that can be used for analysis of system performance and behavior. This model can be used as a tool, in the design of these systems, for determining rack size, arrival and service rates for best performance.

In most cases the model generally behaves as expected. However, some results for the jamming probability do not behave as expected. Also, depending on the system parameters utilized, the rack size may or may not be a factor in the manner in which the system behaves.

The model is of a simplified AS/RS. Its use is therefore limited and can only model a basic AS/RS. For a more complex system one may have to resort to simulation, whose limitations are also well known. However, the current model can still be used as a component of a larger and more complex system. In addition, it can also be used to approximate most AS/RS.

The current model assumes Poisson arrivals and exponential service, but we know that most arrivals in industry are hardly of Poisson type and service is also hardly exponential. The arrival process of an internal system such as this one is often correlated and it is well known that the Markovian Arrival Process (MAP) is a good representation of correlated arrivals.

It is known that because our service is a combination of travel time and placement or removal from the rack, a more general service time, which probably consists of a fixed travel time and a random variable for placement or removal, would be more appropriated. It is also well known that most general services can be represented (very well approximated) by phase type service. Future work should attempt to extend this model to MAP arrivals and phase services.

This model is also limited to a single server of unit-load. Future work in this area might be to extend the model to group services and include other popular service policies.

References

[1] S.L. Allen. A Selection Guide to AS/R Systems. *Industrial Engineering*, 24(3):28–31, 1992.

[2] F. Azadivar. Maximization of the Throughput of a Computerized Automated Warehousing System Under System Constraints. *Intl. Journal of Production Research*, 24(3):551–566, 1986.

[3] K.M. Bafna. Procedures for Investigating AS/RS Feasibility, Suitability are Outlined for IEs. *Industrial Engineering*, 15(8):60–66, 1983.

[4] K.M. Bafna. Procedures Given for Determining AS/RS System Size and Preparing Specs. *Industrial Engineering*, 15(6):76–81, 1983.

[5] Y.A. Bozer and J.A. White. Design and Performance Models for End-of-Aisle Order Picking Systems. *Management Science*, 36(7):852–866, 1990.

[6] W. Chow. An Analysis of Automated Storage and Retrieval Systems in Manufacturing Assembly Lines. *IIE Transactions*, 18(6):204–214, 1986.

[7] P.J. Egbelu and C.-T. Wu. A Comparison of Dwell Point Rules in an Automated Storage/Retrieval System. *Intl. Journal of Production Research*, 31(11):2515–2530, 1993.

[8] S.C. Graves, W.H. Hausman, and L.B. Schwarz. Storage-Retrieval Interleaving in Automatic Warehousing Systems. *Management Science*, 23(9):935–945, 1977.

[9] M.H. Han, L.F. McGinnis, J.S. Shieh, and J.A. White. On Sequencing Retrievals in an Automated Storage/Retrieval System. *IIE Transactions*, 19(1):56–66, 1987.

[10] W.H. Hausman, L.B. Schwarz, and S.C. Graves. Optimal Storage Assignment in Automatic Warehousing Systems. *Management Science*, 22(6):629–638, 1976.

[11] J.M. Hill. Computers and AS/RS Revolutionize Warehousing. *Industrial Engineering*, 12(6):34–45, 1980.

[12] H.F. Lee. Performance Analysis for Automated Storage and Retrieval Systems. *IIE Transactions*, 29:15–28, 1997.

[13] R.J. Lynn and R.A. Wysk. An Expert System Based Controller for an Automated Storage/Retrieval System. *Intl. Journal of Production Research*, 28(4):735–756, 1990.

[14] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models - An algorithmic approach.* Johns Hopkins University Press, Baltimore, MD, 1981.

[15] M.J. Rosenblatt, Y. Roll, and V. Zyser. A Combined Optimization and Simulation Approach for Designing Automated Storage/Retrieval Systems. *IIE Transactions*, 25(1):40–50, 1993.

[16] O.B. Rygh. Justify An Automated Storage & Retrieval System. *Industrial Engineering*, 13(7):20–24, 1981.

[17] L.B. Schwarz, S.C. Graves, and W.H. Hausman. Scheduling Policies for Automatic Warehousing Systems: Simulation Results. *AIIE Transactions*, 10(3):260–270, 1978.

[18] A. Seidmann. Intelligent Control Schemes for Automated Storage and Retrieval Systems. *Intl. Journal of Production Research*, 26(5):931–952, 1988.