

AD-A108 010

AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH
AN ANALYSIS OF RECOVERABLE ITEM INVENTORY SYSTEMS WITH SERVICE --ETC(U)
NOV 78 P L KNEPELL

F/G 15/5

UNCLASSIFIED

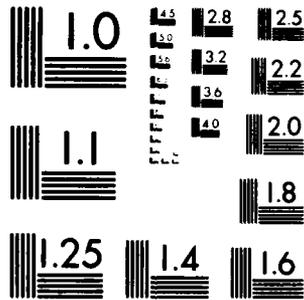
AFIT-CI-79-313T-5

NL

1-2

OF
COPIES

A large grid of approximately 15 columns and 15 rows. The grid is mostly filled with black, suggesting that the data has been redacted or is otherwise obscured. The grid lines are visible, forming a clear structure for data entry or display.



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

UNCLASS

SECURITY CLASSIFICATION OF THIS PAGE (When Data F)

AD A108010

DTIC FILE COPY

1 REPORT NUMBER 79-313T-S 6		2 GOVT ACCESSION NO AD A108010		3 RECIPIENT'S CATALOG NUMBER	
4 TITLE (and Subtitle) An Analysis of Recoverable Item Inventory Systems with Service Facilities Subject to Breakdown				5 TYPE OF REPORT & PERIOD COVERED THESIS/DISSERTATION	
6 AUTHOR (last name, first) Peter L. Knepell				6 PERFORMING ORG REPORT NUMBER	
9 PERFORMING ORGANIZATION NAME AND ADDRESS AFIT STUDENT AT: Cornell University				7 CONTRACT OR GRANT NUMBER(s)	
11 CONTROLLING OFFICE NAME AND ADDRESS AFIT/NR WPAFB OH 45433				12 REPORT DATE Nov 78	
14 MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) LEVEL II				13 NUMBER OF PAGES 119	
16 DISTRIBUTION STATEMENT (of this Report) APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED				15 SECURITY CLASS. (of this report) UNCLASS	
18 SUPPLEMENTARY NOTES APPROVED FOR PUBLIC RELEASE: IAW AFR 190-17				15a DECLASSIFICATION/DOWNGRADING SCHEDULE	
19 KEY WORDS (Continue on reverse side if necessary and identify by block number) ATTACHED				DTIC ELECTED DEC 2 1981 H 23 NOV 1981	
20 ABSTRACT (Continue on reverse side if necessary and identify by block number)				FREDERICK C. LYNCH Director Air Force Institute of Technology (ATIC) Wright-Patterson AFB, OH 45433	

AN ANALYSIS OF RECOVERABLE ITEM INVENTORY SYSTEMS
WITH SERVICE FACILITIES SUBJECT TO BREAKDOWN

Peter L. Knepell, Ph.D.
Cornell University 1979

The purpose of this study is to analyze an inventory/maintenance system for recoverable items, that is, items which are subject to repair when they fail. The repair of items is performed by a maintenance facility which has a fixed number of service stations or channels which are also subject to failure. When an item fails, a demand is immediately placed for a like replacement from a spare pool. The failed part is sent to the repair facility to be serviced on a first-come, first-served basis. The spare pool is replenished when repair on the item is completed. When a service station fails, repair is initiated immediately and the failed server is replaced by an operative spare server if one is available. This analysis is limited to a single-echelon system with no outside sources of supply or repair.

The objective of this study is to model the system described in order to observe the relationship of system performance to spare stock levels and service facility design. Specifically, the model is used to minimize the total expected unit backorders given an investment constraint on the number of spare items, service channels and spare servers in the system. For long range planning purposes, this is accomplished for a system with demands which are stochastic and stationary in nature. An extension is provided, to consider the case where demands are non-stationary and/or the time dependent behavior of the system needs to be described.

In order to express the total expected unit backorders, a representation for the distribution of the number of units requiring repair is needed. Approximations are developed using diffusion techniques since the actual distributions are difficult to express. The diffusion approximation is applied to an optimization problem to provide the best allocation of investments in the system. A simple solution algorithm is given.

Finally, a view of the time-dependent behavior of the system is provided. The problem is decomposed into finding the distributions for (1) the number of units in requiring repair given no service channel failures and (2) the time between service channel failures. We provide a brief review of the literature for the first distribution and an in-depth study of the latter distribution.

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/ _____	
Availability Codes	
Avail and/or	
Dist	Special
A	

AN ANALYSIS OF RECOVERABLE ITEM INVENTORY SYSTEMS
WITH SERVICE FACILITIES SUBJECT TO BREAKDOWN

Peter L. Knepell, Ph.D.
Cornell University 1979

The purpose of this study is to analyze an inventory/maintenance system for recoverable items, that is, items which are subject to repair when they fail. The repair of items is performed by a maintenance facility which has a fixed number of service stations or channels which are also subject to failure. When an item fails, a demand is immediately placed for a like replacement from a spare pool. The failed part is sent to the repair facility to be serviced on a first-come, first-served basis. The spare pool is replenished when repair on the item is completed. When a service station fails, repair is initiated immediately and the failed server is replaced by an operative spare server if one is available. This analysis is limited to a single-echelon system with no outside sources of supply or repair.

The objective of this study is to model the system described in order to observe the relationship of system performance to spare stock levels and service facility design. Specifically, the model is used to minimize the total expected unit backorders given an investment constraint on the number of spare items, service channels and spare servers in the system. For long range planning purposes, this is accomplished for a system with demands which are stochastic and stationary in nature. An extension is provided, to consider the case where demands are non-stationary and/or the time dependent behavior of the system needs to be described.

-iv-

In order to express the total expected unit backorders, a representation for the distribution of the number of units requiring repair is needed. Approximations are developed using diffusion techniques since the actual distributions are difficult to express. The diffusion approximation is applied to an optimization problem to provide the best allocation of investments in the system. A simple solution algorithm is given.

Finally, a view of the time-dependent behavior of the system is provided. The problem is decomposed into finding the distributions for (1) the number of units in requiring repair given no service channel failures and (2) the time between service channel failures. We provide a brief review of the literature for the first distribution and an in-depth study of the latter distribution.

TABLE OF CONTENTS

	Page
1. INTRODUCTION	1
2. THE MODEL	5
2.1 Purpose of the Model	5
2.2 The Expected Backorder Objective Function	6
2.3 Basic Assumptions	8
2.4 Mathematical Statement of the Model	10
3. STATIONARY DISTRIBUTION ANALYSIS	13
3.1 The Queueing Model	13
3.2 Analytic Results	17
3.2.1 Previous Results	17
3.2.2 The Single Server, Single Spare System	21
3.3 Diffusion Approximations for Queueing Systems	26
3.3.1 Derivation of the Diffusion Equation	29
3.3.2 Previous Approximations for Queueing Systems	35
3.3.2.1 Stationary distributions for the GI/M/C queue	36
3.3.2.2 Transient distributions for the GI/G/C queue	42
3.3.3 Approximation for the GI/M/C Queue Subject to Server Breakdown	44
3.3.3.1 Derivation of infinitesimal moments	45
3.3.3.2 Definition of an approximation region	47
3.3.3.3 Solution of the diffusion equation	53
3.3.3.4 Approximation for p_0	58
3.3.3.5 Comparative Analysis	61
4. AN OPTIMIZATION METHOD	75
4.1 Approximating the Backorder Function	76
4.2 An Algorithm for Determining Unit Stocklevels and Service System Design	83

5. NON-STATIONARY ANALYSIS	87
5.1 Introduction and Motivation	87
5.2 Time Dependent Distribution Analysis	89
5.2.1 Closed Form Solutions	90
5.2.2 Approximation Methods	91
5.3 Passage Time Distributions	93
5.3.1 Previous Results	94
5.3.2 Extension to Erlang Distributed Repair Times	100
5.3.3 Extension to Deterministic Repair Times	106
6. CONCLUDING REMARKS	116
BIBLIOGRAPHY	119

LIST OF ILLUSTRATIONS

	Page
3.1 Queueing System Model	14
3.2 State Space Transitions	22
3.3 Relative Errors for Single Server, No Spares	49
3.4 Relative Errors for Two Servers, No Spares	50
3.5 Relative Errors for Single Server, One Spare	51
3.6 Effects of Relative Size of $K\xi$ and $K\eta$	52
3.7 $C=1, L=0, \rho=.55$	64
3.8 $C=1, L=0, \rho=.75$	64
3.9 $C=1, L=0, \rho=.95$	65
3.10 $C=1, L=1, \rho=.55$	66
3.11 $C=1, L=1, \rho=.75$	66
3.12 $C=1, L=1, \rho=.95$	67
3.13 $C=2, L=0, \rho=.55$	68
3.14 $C=2, L=0, \rho=.75$	68
3.15 $C=2, L=0, \rho=.95$	69
3.16 $C=2, L=2, \rho=.75$	70
3.17 $C=2, L=2, \rho=.95$	70
3.18 $C=3, L=0, \rho=.75$	71
3.19 $C=3, L=0, \rho=.95$	71
3.20 $C=3, L=3, \rho=.75$	72
3.21 $C=3, L=3, \rho=.90$	72
3.22 $C=5, L=0, \rho=.75$	73
3.23 $C=5, L=0, \rho=.95$	73

3.24	C=5, L=3, $\rho=.75$	74
3.25	C=5, L=3, $\rho=.95$	74
4.1	Parameters K, r, and b_1 ($\lambda > \mu \bar{C}(0)$)	82
4.2	Parameters K, r, and b_1 ($\lambda < \mu \bar{C}(0)$)	82
4.3	Performance and Tradeoff Curves (C Fixed)	85
5.1	System Performance for Different Service States	88
5.2	Transition Flows for Number of Inoperative Servers	93
5.3	Transitions From GOOD to BAD States	95
5.4	Log Survival Function Comparison [7]	100
5.5	Service State Changes Between Arrival Epochs	101
5.6	Service State Changes Between Service Epochs	107
5.7	GOOD to BAD to GOOD State Transitions	111
5.8	Transitions When Repairs Initiated Immediately	114

LIST OF BASIC SYMBOLS

λ_i	is the failure rate of item i
λ	$= \sum_{i=1}^m \lambda_i$
μ	is the service rate of a single service channel
ξ	is the failure rate of a single service channel
η	is the rate of repair of a failed server
ρ	is the system traffic intensity
$\sigma^2(x)$	is the infinitesimal variance of the diffusion process given state x
σ_a^2	is the variance of interarrival times
σ_d^2	is the variance of service times
$\sigma_k^+(s)$	is the Laplace transform of the p.d.f. $S_k^+(t)$
a_i	is the i^{th} row of submatrix A , $i=0,1,\dots$
$A(t)$	is the arrival process for a queue
B	is the budget available for purchase of spares and servers; also, the BAD state and a submatrix
b_i	is the i^{th} row of submatrix B , $i=0,1,\dots$
$B_i(s_i, L)$	is the expected backorders for item i given a spare stock level of s_i
$B(S, L)$	$= \sum_{i=1}^m B_i(s_i, L)$
C	is the number of service channels
\bar{C}	is the expected number of service channels operational at any point in time
C_C	is the cost of a service channel

C_L is the cost of a spare server
 C_i is the cost of a spare item i
 $D(t)$ is the departure process for a queue
 E_i is the essentiality of item i
 $E_n(x,t)$ is the infinitesimal n^{th} moment of a diffusion process in state x at time t
 $f(w,\tau;x,t)$ is the probability density of a transition from state w at time τ to state x at time t
 $F_k(x,t)$ is the probability of x or less units in the system at time t given k service channels operational
 $\bar{F}_P(t) = \int_t^\infty S_P(y)dy =$ the survival function of T_P
 $\bar{F}_G(t) = \int_t^\infty S_G(y)dy =$ the survival function of T_G
 $F_{ij}(t)$ is the probability that the j^{th} failure occurs before time t given i servers are being repaired at time 0
 G is the GOOD state
 L is the number of spare servers provided
 $m(x)$ is the infinitesimal mean of the diffusion process given state x
 $N(t) = A(t)-D(t) =$ continuous-time, discrete valued random variable
 $p_n = p(n|\lambda,\mu,\xi,\eta,C,L) =$ the stationary probability of n units in the system
 \tilde{p}_n is the approximation of p_n
 $P_k(a,b,t)$ is the probability of a transition from a to b in time t given k service channels operational
 q_k is the stationary probability of k servers are in operational order
 q_k' is the stationary probability k servers are in operational order just prior to a service channel failure

R is a fixed repair time
 s_i is the spare stock level for item i
 $S = (s_1, s_2, \dots, s_m)$ = the set of spare stock levels for all items
 $S_k^+(t)$ is the p.d.f. of the transition time from state k to $k + 1$
 $S_G(t)$ is the p.d.f. for the variable T_G
 $S_P(t)$ is the p.d.f. for the variable T_P
 T_G is the post recovery failure time
 T_B is the post failure recovery time
 T_P is the time from the perfect state to the BAD state
 w_i is the i^{th} row of submatrix w , $i=0,1,\dots$
 $X(t)$ is a continuous-time, continuous valued random variable
 $Y_k(t)$ is an Erlang distribution with shape parameter k
 $\underline{1}$ is the column vector $(1,1,\dots,1)^T$

CHAPTER I
INTRODUCTION

The purpose of this study is to analyze an inventory/maintenance system for recoverable items, that is, items which are subject to repair when they fail. The repair of items is performed by a maintenance facility which has a fixed number of service stations or channels which are also subject to failure. When an item fails, a demand is immediately placed for a like replacement from a spare pool. The failed part is sent to the repair facility to be serviced on a first-come, first-served basis. The spare pool is replenished when repair on the item is completed. When a service station fails, repair is initiated immediately and the failed server is replaced by an operative spare server if one is available. This analysis is limited to a single-echelon system with no outside sources of supply or repair.

The objective of this study is to model the system described in order to observe the relationship of system performance to spare stock levels and service facility design. Specifically, the model is used to minimize the total expected unit backorders given an investment constraint on the number of spare items, service channels and spare servers in the system. For long range planning purposes, this is accomplished for a system with demands which are stochastic and stationary in nature. An extension is provided, to consider the case where demands are non-stationary and/or the time dependent behavior of the system needs to be described.

These inventory/maintenance systems involve vast capital invest-

ments; hence, the design and control of these systems are a great concern for managers. In large scale industrial and military activities, a majority of the inventory items are inexpensive consumable (non-recoverable) units; however, a large proportion of the inventory investment is for spare stock levels of recoverable items. Sherbrooke [32] states that recoverable item spares in the Air Force account for 78 percent of the total investment, amounting to approximately five billion dollars. Currently, automated repair stations, costing up to 16 million dollars each, are being purchased by military organizations to repair units which cost an average of 100 thousand dollars each.

To provide a better understanding of the structure of this system, we will describe a specific example where the units which fail are sophisticated aircraft electronic (avionics) components, such as radar, navigation instruments and radios. Spare units are stocked at the airfield where the demands occur. The unit failure rates are usually low and failures occur independently. Occasionally, a rash of breakdowns occur in a particular item or the failure of one item may induce the failure of different units. These sources of dependent demand are infrequent and difficult to predict and, therefore, are not considered in the analysis. In practice, a small number of units are sent to another facility or higher echelon level for repair or replacement.

The repair stations are situated at the airfield. They also involve sophisticated electronic components and their failure characteristics are similar to those of the recoverable items. In addition, these stations are periodically out of service for modification, preventive maintenance, and calibration. If a service station fails, a high priority is placed on its repair since service interruptions

ultimately affect the number of operational aircraft. A supply of spare components for the repair stations is usually provided.

Thus the objective in designing the system in this example is to provide the optimal investment allocation for spare units and service facilities at the airfield so that the maximum number of aircraft are operationally ready.

We begin the study in Chapter II with a brief description of some previously developed models of recoverable item inventory systems. This is followed by a justification for the use of the expected number of unit backorders as a measure of system performance. After some basic assumptions are listed, a mathematical statement of our model is provided.

Chapters III and IV provide a planning tool for managers to use when designing a recoverable item inventory system. Chapter III develops methods for obtaining the stationary probability distribution of the number of units being repaired. We use this distribution to compute the expected number of unit backorders. Approximations are developed using diffusion techniques since the actual distributions are difficult to express. In Chapter IV, the diffusion approximation is applied to an optimization problem to decide on the allocation of investments in the system. A simple solution algorithm is given.

A view of the time dependent behavior of the system is given in Chapter V. The problem is decomposed into finding the distributions for (1) the number of units requiring repair given no service channel failures and (2) the time between service channel failures. We provide a brief review of the literature for the first distribution and an in-

depth study of the latter distribution. The final chapter contains some closing remarks and suggestions for future research.

CHAPTER II

THE MODEL

Much attention has been focused on attempts to model inventory systems like the one described earlier. Feeney and Sherbrooke [5] examined single-echelon recoverable item systems where demand was generated by a compound Poisson process. Sherbrooke [32] extended the results to a two echelon system in a model he called METRIC (Multi-Echelon Technique for Recoverable Item Control). Muckstadt [23] extended the METRIC model to include part hierarchies. All of these papers assumed that the service facility has adequate capacity to repair all items without delay (i.e., the infinite server assumption). Gross, et. al., [2,9] considered a recoverable item system with finite service capacity. They modeled their system as a classical machine-repairman problem. However, they too assumed the servers were reliable.

Typically, service facilities are constrained in their capacity to a finite number of servers and, in some cases, the servers are subject to failure. Under these considerations it is important to consider the effects of service facility design on overall system performance. Some of the work done in this area will be discussed later; in general, very few results are available in the context of production-inventory control.

2.1 Purpose of the Model

A model for a single-echelon, recoverable item inventory system will be developed in this chapter. The model can then be used to quantify the relationships between (1) service reliability and capacity

and (2) overall system performance. While it will be useful as a design tool for managers, it is not intended to help them make day-to-day decisions in the dynamic environment of the inventory system. Although it is our objective to create a realistic model, some simplifying assumptions will be made to facilitate the analysis. The most important step is to establish a meaningful performance measure.

2.2 The Expected Backorder Objective Function

We shall use the sum of the expected unit backorders as our performance measure. Consider the system's operation for a fixed number of days, and count the total number of days in which units are backordered for that period. The expected value of this number divided by the number of days in the period gives the expected backorders per day. Our goal is to minimize this function. Note that by this definition, a ten day backorder is equivalent to ten backorders for one day.

Other performance measures, such as NORS rate, fill rate, and ready rate, are not as versatile as expected backorders when modeling recoverable item inventory systems. In Air Force parlance, the NORS (not operationally ready, supply) rate is considered an excellent measure of logistics support. This figure represents the minimum number of aircraft which cannot perform a mission due to supply backorders. It has the advantage of measuring the direct impact that the inventory system has on the fleet it supports. Unfortunately, it is a difficult measure to quantify and use in inventory models. For example, if ten different aircraft are grounded, each due to a different component being backordered, then only one aircraft is considered NORS. This is further

complicated by component interchangeability, substitutability and redundancy. Once quantified, the NORS function is not separable and, hence, is difficult to work with.

Another measure, fill rate, is defined as the fraction of demands that are immediately satisfied by supply. This measure ignores the length of time a unit is backordered. Sherbrooke [32] points out that when fill rate is employed in a multi-echelon inventory system, managers are encouraged to concentrate nearly all stock at the lowest echelon. While backorders will be infrequent, they will have long durations. Another disadvantage is that a "fill" is usually defined as an immediate satisfaction of a demand. If a short delay in satisfaction is acceptable, the resulting optimal policy may be considerably different. In fact, as longer delays are accepted, the more closely the results resemble those of the backorder criterion.

Ready rate is defined as the fraction of items which are not backordered. This measure does not reflect the number of units backordered on a particular item. Thus, it is conceivable that inexpensive items will be stocked heavily in favor of the expensive items. Then backorders will accumulate only on the relatively few expensive items and the system performance, measured by ready rate, will appear excellent while large numbers of backorders exist for expensive items.

The expected backorder criterion combines the number of backorders and the length of each backorder as a penalty. It eliminates the need to determine arbitrary backorder and holding costs and provides a direct measure of support that the inventory system provides. Sherbrooke [32] mentions a single-echelon example in which fill rate, ready rate, and expected backorder objective functions provide essentially identical

stockage policies. However, when applied to multi-echelon problems, the expected backorder criterion yields more reasonable results. Additionally, the expected backorder function is convex and separable, properties which are computationally helpful and not necessarily possessed by other criteria.

2.3 Basic Assumptions

A list of the basic assumptions made for this model will be given, followed by an expanded discussion of each.

1. The demand process for each of m different items is a Poisson process. All demands occur independently at a rate λ_i , $i=1, \dots, m$.
2. With each demand, units are exchanged on a one-for-one basis.
3. All units turned in are serviced.
4. Service times are stochastically independent and exponentially distributed at rate μ . There are at most C service channels available. If service is interrupted by a channel failure, then the unit is immediately moved to the next available service channel and service is resumed without delay.
5. There is no batching of items for repair. Items are serviced on a first-in, first-served basis.
6. Service channels fail independently as Poisson events at a rate ξ .
7. Failed servers are repaired immediately with exponentially distributed repair times at a rate η .
8. Failed channels are replaced instantaneously, if a spare is available. If no spare is available, channels are replaced in order of breakdown times when repaired servers become available. L spare servers

are provided.

The first assumption implies that the arrival rate of units does not depend on the size of the population. In a standard application, there is a finite source of demand; however, when looking at the scenarios we are modeling, this assumption is valid. For example, consider a fleet of aircraft as generating units requiring service. As aircraft units are backordered, the number of operational aircraft decreases. Since the flying schedule is fixed, the remaining aircraft will have to fly more to satisfy the schedule. Since we assume an aircraft's failure rate is directly related to its usage, the perceived fleet failure rate is assumed to remain the same. This, of course, neglects the possibility that a large proportion or, for that matter, the entire fleet could be grounded at the same time. In a later chapter we will allow the demand rate to vary over time to reflect changing flying schedules or failure characteristics.

The second assumption reflects an (S,S-1) inventory policy. The next assumption states that the system is conservative (i.e., no condemnations). In practice, the inventory items are expensive, so these assumptions reflect a reasonable policy.

The fourth assumption gives non-preemptive priority to a unit whose service is interrupted. Theoretically, given the memoryless property of the exponential distribution, this assumption does not matter. It will become apparent later that exponentially distributed interarrival times are not necessary for the approximation techniques used. This assumption is made because the approximation method proposed was tested against systems with exponentially distributed service times.

The sixth assumption implies that service channels can fail even

when idle. This is realistic since calibration tests, preventive maintenance and modifications are typical for the systems being modeled. The next assumption implies that there is an adequate number of repairmen to work on the failed channels. This can be altered to specify a limited number of repairmen; however, this adds to the notational and computational complexity, but does not alter the method of analysis.

In Chapter I, it was mentioned that the service channels are large and costly to establish. Spares for these channels are relatively inexpensive. Therefore, it is reasonable to assume that while $C + L$ servers are provided, only C are useable service channels and the remaining L must be held in reserve.

2.4 Mathematical Statement of the Model

The objective of inventory managers is to provide the greatest system performance given a fixed budget. Thus, our goal is to minimize total expected backorders outstanding at any point in time subject to a budget constraint. This model will be different from those mentioned earlier because the investment constraint links the purchase of unit spares, service channels, and spare servers. Thus the system performance will be dependent on the service facility design as well as the allocation of funds to unit spares.

Suppose $p(n| \cdot)$ represents the probability that n units are in the service facility (in service or awaiting service). We know this number will be a function of the input and the output of the service facility. The parameter λ_1 characterizes the input process for item i and the parameters $\mu, \xi, \eta, C,$ and L determine the output process. For each unit of type i we can express the expected number of backorders outstanding

at any time as

$$\sum_{x > s_i} (x - s_i) p(x | \lambda_i, \mu, \xi, \eta, C, L). \quad (2.1)$$

Backorders in some items may be considered more serious than for others.

In this case, we can weight the backorder function in (2.1) with an essentiality factor, E_i .

The costs of unit spares and service channels are considered to be linear. Since the initial setup cost for service channels is very large, there will be different costs for service channels and their spares. An expression for the investment constraint is

$$C \cdot C_c + L \cdot C_s + \sum_{i=1}^m C_i s_i \leq B. \quad (2.2)$$

If we are concerned with a long range planning tool, we must be careful to design a service facility which can provide adequate service over an extended period of time. Clearly, the potential output rate must be at least as large as the input rate:

$$\lambda = \sum_{i=1}^m \lambda_i < \mu \bar{C}, \quad (2.3)$$

where \bar{C} is the expected number of service channels operational at any point in time. The mathematical derivation of this constraint will be provided later.

Combining all the statements above we have a mathematical statement of the model as follows:

$$\text{minimize } \sum_{i=1}^m E_i \sum_{x > s_i} (x - s_i) p(x | \lambda_i, \mu, \xi, \eta, C, L)$$

subject to

$$CC_c + LC_s + \sum_{i=1}^m C_i s_i \leq B, \quad (P)$$

and $\lambda < \mu\bar{C}$,

where C , L , and s_i are non-negative integers, $i=1, \dots, m$.

For future reference, this optimization problem will be denoted as problem P.

CHAPTER III
STATIONARY DISTRIBUTION ANALYSIS

To apply the model developed in Chapter II as a long range planning tool, we need to find an accurate expression for $p(n|\cdot)$, the stationary distribution for the number of units in the system. To do this, we view the single-echelon, recoverable item system as a queueing system in which the service facility has a finite number of servers, each subject to failure. The servers and their repair facility will be another queueing system imbedded in the first one. This structure will be exploited to develop the stationary probability distributions of units in the service system. Some analytic results will be given and a general approximation method will be derived.

3.1 The Queueing Model

This section will develop a queueing model to represent the system introduced in Chapter I. It will become apparent that this system has much in common with the classical machine-repairman problem. The assumptions given previously establish the queue service discipline and allow for the creation of a state space for a continuous-time Markov process.

The system to be modeled can be described schematically (see Figure 3.1). Units demanding service come from an infinite source. They enter a multi-server queue and are served on a first-come, first-served basis. As soon as a unit joins the queue, a replacement unit, if one is available, is immediately returned to the source from the spare pool. If no replacement is available it is backordered.

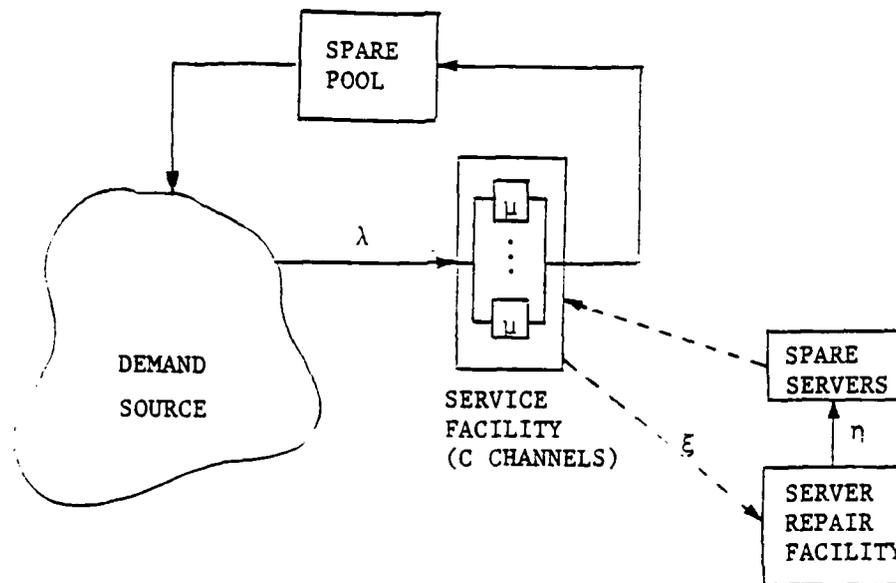


Figure 3.1: Queuing System Model

Servers are subject to failure and are replaced by spare servers, if one is available. Figure 3.1 illustrates why this system could be called "a two-dimensional machine-repairman system."

We must assume this system is non-saturated in order to guarantee the existence of a stationary probability distribution. To assure this, as mentioned earlier, we will require

$$\lambda < \mu \bar{C}, \quad (3.1)$$

where \bar{C} is the expected number of servers available at any point in time. This condition is necessary and sufficient when there are no spare servers ($L=0$); however, it does not seem easy to prove in general [20,33].

For the queuing system developed, the probability function $p(n|\cdot)$ has some familiar forms in special cases. (For notational convenience, we will let $p_n = p(n|\cdot)$, for the remainder of the chapter.) If $C \rightarrow \infty$,

then from Palm's Theorem we know that p_n is Poisson with rate λ/μ . If $C=1$ and there are no service station failures (i.e., $\xi=0$ or $L \rightarrow \infty$), then p_n is geometric with parameter λ/μ , since we have a simple M/M/1 queue. If there are no spare service stations ($L=0$) and C is finite, then p_n can be determined using the results of Mittrany and Avi-Itzhak [20].

Since the servers can fail at any time, the number of operational channels is independent of the number of units in the system. (The reverse, however, is not true.) Thus, the stationary probability distribution of the number of servers operational can be obtained separately. This subsystem can be viewed as a classical machine-repairman problem where the servers and their spares are the "machines" and there are always an adequate number of "repairmen."

Theorem 3.1.

If we have a system with C service channels subject to failure, L spare servers, and $C + L$ repairman where:

- i) the servers fail independently, as Poisson events at rate ξ , and
- ii) the repair times are independent and exponentially distributed with rate η ,

then the stationary probability of having k operational servers, q_k , exists and is given by

$$q_k = \begin{cases} \frac{C^L C!}{(C+L-k)! k!} \left(\frac{\xi}{\eta}\right)^{C+L-k} q_{C+L} & , 0 \leq k \leq C - 1, \\ \frac{1}{(C+L-k)!} \left(\frac{C\xi}{\eta}\right)^{C+L-k} q_{C+L} & , C \leq k \leq C + L, \end{cases} \quad (3.2)$$

where

$$q_{C+L} = \left[\sum_{k=0}^L \frac{C^k}{k!} \left(\frac{\xi}{\eta}\right)^k + \sum_{k=L+1}^{C+L} \frac{C^L C!}{(C+L-k)! k!} \left(\frac{\xi}{\eta}\right)^k \right]^{-1}. \quad (3.3)$$

A simple proof of these results can be found in reference 6.

Thus the stationary probability of having k repair channels operational is

$$\Pr\{k \text{ channels up}\} = \begin{cases} q_k & , 0 \leq k \leq C-1, \\ \sum_{n=C}^{C+L} q_n, & k = C. \end{cases} \quad (3.4)$$

In the special case where $L=0$ (i.e., no spares) we have

$$q_k = \binom{C}{k} \left(\frac{\xi}{\eta}\right)^{C-k} q_{C+L}, \quad 0 \leq k \leq C, \quad (3.5)$$

where

$$q_{C+L} = \frac{1}{\sum_{k=0}^C \binom{C}{k} \left(\frac{\xi}{\eta}\right)^k} = \frac{1}{\left(1 + \frac{\xi}{\eta}\right)^C}. \quad (3.6)$$

Defining $\rho = \frac{\xi}{\eta}$ we have

$$q_k = \frac{\binom{C}{k} \rho^k}{(1 + \rho)^C} = \binom{C}{k} \left(\frac{\rho}{1 + \rho}\right)^k \left(\frac{1}{1 + \rho}\right)^{C-k}, \quad 0 \leq k \leq C. \quad (3.7)$$

So when no spare servers are provided, the stationary probability distribution is binomial.

The conditional probability that k servers are operational given that a failure is just about to occur is not the same as q_k since the failed service units come from a finite source. A derivation for this new distribution, q_k' , can be found in reference 9.

Corollary 3.1. Let the same conditions as Theorem 3.1 hold. Then the stationary probability $q'_k = \text{Pr} \{ k \text{ servers operational} \mid \text{a server failure is about to occur} \}$ exists and is given by

$$q'_k = \begin{cases} \frac{kq_k}{C - \sum_{n=0}^C (C-n)q_n} & , 0 \leq k \leq C \\ \frac{Cq_k}{C - \sum_{n=0}^C (C-n)q_n} & , C+1 \leq k \leq L+C. \end{cases} \quad (3.8)$$

3.2 Analytic Results

The typical procedure used to derive the line length distribution involves probability generating functions. A Markovian state space is designed, balance equations are developed and a generating function is derived. The poles of the generating function provide information which is otherwise difficult to derive. The previous work found in the literature has been limited to cases where no spare service channels are available. This work will be discussed in greater detail and then a derivation for the "simple" case of one server and one spare will be given.

3.2.1 Previous Results

The earliest published work done on queueing systems with service station subject to breakdown was by White and Christie in 1958 [34]. The most comprehensive article on single server queueing systems with Poisson inputs and exponential service times was written in 1963 by Avi-

Itzhak and Naor [1]. They allowed a general distribution for server breakdowns and repair and were only able to derive the expected line length and expected waiting time. Shogan in 1977 provided the line length distribution for a single server system [33]. His results are summarized below.

Theorem 3.2 (Shogan)

Suppose we have a system with:

- (i) independent, exponentially distributed inter-arrival times, rate λ ,
- (ii) independent Erlang distributed service times with mean $1/\mu$ and shape parameter k ,
- (iii) a single server, no spares, with exponential inter-failure times, rate ξ , and
- (iv) Erlang distributed repair times with mean $1/\eta$ and shape parameter m .

If $\lambda/\mu < \eta/(\xi + \eta)$, then the stationary probability distribution

$$P_{ij} = \Pr \{ \begin{array}{l} i \text{ phases of repair required on the server,} \\ j \text{ phases of service in the system} \end{array} \}$$

exists and is defined by the recursive equations (for $j = 0, 1, 2, \dots$):

$$P_{00} = \eta/(\xi + \eta) - \lambda/\mu, \quad (3.9)$$

$$P_{m,j} = (\lambda - m\eta)^{-1} (\xi P_{0,j} + \lambda P_{m,j-k}), \quad (3.10)$$

$$P_{i,j} = (\lambda - m\eta)^{-1} (m\eta P_{i+1,j} + \lambda P_{i,j-k}), \quad i = m-1, m-2, \dots, 1, \quad (3.11)$$

$$P_{0,j+1} = (k\mu)^{-1} \left(\lambda \sum_{n=j+1-k}^j P_{0,n} + \xi \sum_{n=0}^j P_{0,n} - m\eta \sum_{n=0}^j P_{1,n} \right), \quad (3.12)$$

where P_{ij} is zero if it has a negative subscript and summations are zero if the lower limit of summation is negative.

The expressions in Theorem 3.2 are typical. A closed form solution for the line length probabilities has not been found. The recursive relations are a direct result of the balance equations and P_{00} was obtained from the pole of a lengthy generating function. After solving for all the state probabilities, the steady state probability of having n customers in the system is given by:

$$P_n = \begin{cases} \sum_{i=0}^m P_{i,0} & \text{for } n=0 \\ \sum_{i=0}^m \sum_{j=(n-1)k+1}^{nk} P_{ij} & \text{for } n > 0. \end{cases} \quad (3.13)$$

The groundwork for multi-server queues with server breakdowns and no spares has been established in an article by Mitrany and Avi-Itzhak in 1968 [20]. Their results for a two server system are summarized below.

Theorem 3.3 (Mitrany and Avi-Itzhak)

Suppose we have a system with:

- (i) independent, exponentially distributed inter-arrival times, rate λ ,
- (ii) independent, exponentially distributed service times with mean $1/\mu$,
- (iii) two servers, no spares, with independent, exponentially

distributed inter-failure times, rate ξ , and

(iv) independent, exponentially distributed repair times, with mean $1/\eta$.

Define

$$z = [(\lambda + \mu + \xi + \eta) - \sqrt{(\lambda + \mu + \xi + \eta)^2 - 2\lambda\mu}] / 2\lambda, \quad (3.14)$$

$$N = 2\mu\eta - \lambda\eta - \lambda\xi, \quad \text{and} \quad (3.15)$$

$$D = \mu(\xi + \eta) [2(\xi + \eta)(2\mu + \lambda + 2\xi z) + \lambda(1 - z)(2\mu + \lambda)] . \quad (3.16)$$

If $\lambda/\mu < 2\eta / (\xi + \eta)$, then the steady state probability

$P_{ij} = \Pr \{ i \text{ operating servers, } j \text{ units in the system} \}$

exists and is defined by the recursive equations (for $i = 0, 1, 2$):

$$P_{00} = \xi^2 [4\mu + (2\lambda + 4\xi)z] \cdot N / (\lambda + \eta)D, \quad (3.17)$$

$$P_{10} = (\lambda + \eta) P_{00} / \xi, \quad (3.18)$$

$$P_{20} = [(\lambda + 2\eta)\mu + (2\xi\eta - \lambda\mu)z] \cdot N / D, \quad (3.19)$$

$$[j\mu + i\xi + \lambda + (2-i)\eta] P_{ij} = (3-i)\eta P_{i-1,j} + (i+1)\xi P_{i+1,j} + \lambda P_{i,j-1} \\ + (j+1)\mu P_{i,j+1}, \quad \text{for } j < i, \quad (3.20)$$

$$[i\mu + i\xi + \lambda + (2-i)\eta] P_{ij} = (3-i)\eta P_{i-1,j} + (i+1)\xi P_{i+1,j} + \lambda P_{i,j-1} \\ + i\mu P_{i,j+1}, \quad \text{for } j \geq i, \quad (3.21)$$

where P_{ij} is zero if $i > 2$ or any subscript is negative.

The steady state probability of having n customers in the system is simply

$$p_n = \sum_{i=0}^2 P_{i,n} \quad \text{for } n \geq 0.$$

The work required for $n > 2$ is computationally and notationally intractable and Mitraný and Avi-Itzhak suggest numerical methods. In a related article published in 1973, Yechiali was also unable to derive closed form solutions except in very specific cases. He stated that, "In general, no closed form relations are available for the probabilities $P_{i,0}$, and, except for numerical results, no analytic comparison to the elegant results of the classical $M/M/1$ queue can be made." [35]

3.2.2 The Single Server, Single Spare System

This section will show a technique to solve analytically for the stationary line length distribution for a single server queue subject to breakdown, where one spare server is provided. In addition to the assumptions stated in Section 2.3, we will also assume that operable spares do not fail while held in the spare server pool. This is a reasonable restriction considering the application; however, it can be lifted without significantly affecting the analysis.

The solution procedure to be used follows. A state space for a continuous-time Markov process will be defined. The Markovian nature of the state space allows us to describe the flows in the system with balance equations. These equations cannot be solved directly so probability generating functions are derived. These will assist us in

solving for three of the stationary probabilities. When these probabilities are known, the entire distribution can be obtained using the balance equations.

We start by defining a state space for this system as

$$\{(i,j) \mid i = \text{number of operable servers} = 0,1,2, \\ j = \text{number of units in the system} = 0,1,\dots\} .$$

Since the state transition times are all independent and exponentially distributed, the process described is Markovian. Define $P_{i,j}$ as the steady state probability of being in state (i,j) . The transition flows are illustrated in Figure 3.2.

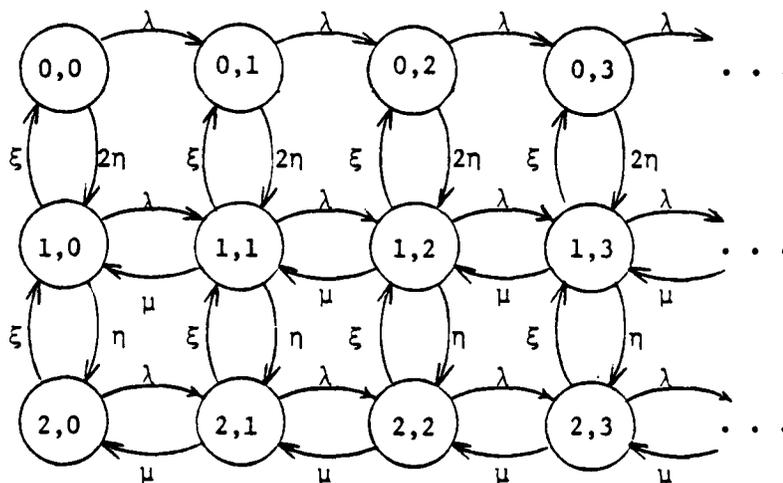


Figure 3.2: State Space Transitions

The balance equations for this system are:

$$(\lambda+2\eta)P_{0j} = \lambda P_{0,j-1} + \xi P_{1,j} \quad , \quad j=0,1,\dots, \quad (3.22)$$

$$(\lambda+\xi+\eta)P_{10} = 2\eta P_{00} + \xi P_{20} + \mu P_{11} \quad , \quad (3.23)$$

$$(\lambda+\xi+\mu+\eta)P_{1j} = 2\eta P_{0j} + \xi P_{2j} + \mu P_{1,j+1} + \lambda P_{1,j-1} \quad , \quad j=1,2,\dots, \quad (3.24)$$

$$(\lambda+\xi)P_{20} = \eta P_{10} + \mu P_{21} \quad , \quad (3.25)$$

$$(\lambda+\xi+\mu)P_{2j} = \eta P_{1j} + \mu P_{2,j+1} + \lambda P_{2,j-1} \quad , \quad j=1,2,\dots \quad (3.26)$$

Define the probability generating functions (pgf)

$$G_i(z) = \sum_{j=0}^{\infty} P_{ij} z^j \quad , \quad i=0,1,2.$$

Note that these generating functions evaluated at $z = 1$ give the stationary probability of the number of operational servers. Applying Theorem 3.1 we have

$$G_i(1) = q_i = \begin{cases} \frac{\xi^2}{\xi^2 + 2\eta\xi + 2\eta^2} & , \quad i=0 \\ \frac{2\eta\xi}{\xi^2 + 2\eta\xi + 2\eta^2} & , \quad i=1 \\ \frac{2\eta^2}{\xi^2 + 2\eta\xi + 2\eta^2} & , \quad i=2. \end{cases} \quad (3.27)$$

Also note that

$$G_0(1) + G_1(1) + G_2(1) = 1.$$

We can solve for each pgf by multiplying equations (3.22) through (3.26) by the appropriate z^j and summing. The result is:

$$[\lambda(1-z)+2\eta]G_0(z) - \xi G_1(z) = 0, \quad (3.29)$$

$$-2\eta z G_0(z) + [\lambda z(1-z) + \mu(z-1) + \xi z + \eta z] G_1(z) - \xi z G_2(z) = \mu(z-1) P_{10}, \quad (3.30)$$

$$-\eta z G_1(z) + [\lambda z(1-z) + \mu(z-1) + \xi z] G_2(z) = \mu(z-1) P_{20}. \quad (3.31)$$

After some elementary row operations we have an equivalent matrix form

$A(z) \cdot \underline{g}(z) = \mu(z-1) \underline{b}$, where:

$$A(z) = \begin{bmatrix} \lambda(1-z)+2\eta & -\xi & 0 \\ \lambda z(1-z) & (\lambda z - \mu)(1-z) & (\lambda z - \mu)(1-z) \\ 0 & -\eta z & \lambda z(1-z) + \mu(z-1) + \xi z \end{bmatrix},$$

$$\underline{g}(z) = \begin{bmatrix} G_0(z) \\ G_1(z) \\ G_2(z) \end{bmatrix} \quad \text{and} \quad \underline{b} = \begin{bmatrix} 0 \\ P_{10} + P_{20} \\ P_{20} \end{bmatrix}$$

If we had values for P_{00} , P_{10} and P_{20} we could solve for all $P_{i,j}$ via the balance equations. Equation (3.22) provides one equation in P_{00} and P_{10} . The above pgf's will be used to provide two more independent equations.

Using Cramer's method for solving simultaneous equations we have:

$$G_i(z) = \frac{|A_i(z)|}{|A(z)|} \mu(z-1), \quad i=0,1,2, \quad (3.32)$$

where $|A_i(z)|$ is the determinate of $A(z)$ with column $(i+1)$ replaced by \underline{b} . Examining row 2 of $A(z)$ it is clear that $|A(z)| = (z-1) |Q(z)|$

where

$$Q(z) = \begin{bmatrix} \lambda(1-z)+2\eta & -\xi & 0 \\ -\lambda z & -\lambda z+\mu & -\lambda z+\mu \\ 0 & -\eta z & \lambda z(1-z)-\mu(1-z)+\xi \end{bmatrix}.$$

So we have

$$G_i(z) = \frac{|A_i(z)|}{|Q(z)|} \mu, \quad i=0,1,2. \quad (3.33)$$

By its definition, $G_i(z)$ must be continuous and bounded on the interval $[0,1]$ and we know $G_i(1) = q_i$, where q_i is given in (3.27). Thus we have

$$\frac{G_i(1) \cdot |Q(1)|}{\mu} = |A_i(1)|, \quad i = 0,1,2. \quad (3.34)$$

It is easy to show that these three equations are dependent; therefore, we only get one useful equation in P_{10} and P_{20} . For $i = 0$ we get

$$\frac{-2\eta(\lambda-\mu)(\xi+\eta)+\lambda\xi^2}{\mu(\xi^2+2\xi\eta+2\eta^2)} = P_{10} + P_{20}. \quad (3.35)$$

If we had a root for $|Q(z)|$ on the interval $(0,1)$, then by continuity of $G_0(z)$ on $(0,1)$ we would have another independent equation for $P_{1,0}$ and $P_{2,0}$. Evaluating $|Q(z)|$ at $z = 0$ and $z = 1$ we have:

$$|Q(0)| = -(\lambda + 2\eta) \mu^2 < 0 \quad (3.36)$$

and

$$|Q(1)| = -2\eta(\lambda - \mu)(\xi + \eta) - \lambda\xi^2 \quad (3.37)$$

Forcing $|Q(1)| > 0$ yields the relationship:

$$\frac{\lambda}{\mu} < \frac{2\xi\eta + 2\eta^2}{\xi^2 + 2\xi\eta + 2\eta^2} = q_1 + q_2. \quad (3.38)$$

This is the familiar constraint requiring the expected service demanded per unit time to be less than the expected service available. Given that this condition is met, the fourth degree polynomial $|Q(z)|$ has a root, z_1 , on the interval $(0,1)$. Then we have $|A_0(z_1)| = 0$, or equivalently

$$[(\lambda z_1 - \mu)(1 - z_1) + \xi z_1]P_{10} + \xi z_1 P_{20} = 0 \quad (3.39)$$

Equations (3.22), (3.35), and (3.39) provide three independent equations in P_{00} , P_{10} , and P_{20} . Using these, the balance equations and the traffic intensity condition (3.38) we have:

Theorem 3.4

Suppose we have a system with:

- (i) independent, exponentially distributed inter-arrival times, rate λ ,
- (ii) independent, exponentially distributed service times with mean $1/\mu$,
- (iii) one server and one spare, with independent, exponentially distributed channel inter-failure times, rate ξ , and
- (iv) independent, exponentially distributed repair times with mean $1/\eta$.

Define

$$K(z) = (\lambda z - \mu)(1 - z) + \xi z \quad (3.40)$$

$$D(z) = [\lambda(z-1) - 2\eta][\lambda z - \mu][K(z) + \eta z] - \lambda \xi z K(z) \quad (3.41)$$

and
$$q_0 = \frac{\xi^2}{\xi^2 + 2\eta\xi + 2\eta^2} \quad (3.42)$$

If $\lambda/\mu < 1 - q_0$, then

- a) $D(z)$ has a root, z_1 , on the interval $(0,1)$, and
- b) the stationary probability distribution

$$P_{ij} = \Pr \{ \begin{array}{l} i \text{ servers are operative,} \\ j \text{ units are in the system} \end{array} \}$$

exists and is defined by:

$$P_{20} = \frac{-q_0 D(1) K(z_1)}{\xi^2 \mu \cdot [\xi z_1 - K(z_1)]} \quad , \quad (3.43)$$

$$P_{10} = \frac{q_0 z_1 D(1)}{\xi \mu [\xi z_1 - K(z_1)]} \quad , \quad (3.44)$$

$$P_{00} = \xi P_{10} / (\lambda + 2\eta) \quad , \quad (3.45)$$

and the balance equations (3.22) through (3.26).

3.3 Diffusion Approximations For Queueing Systems

It should be evident that analytic results for large systems are algebraically cumbersome and do not provide any insight into the nature of the desired distribution. A computationally efficient and relatively simple approximation technique would greatly assist us in solving the optimization problem proposed earlier. For the situation that we are modeling, diffusion approximations provide simple and accurate representations for the queue size distribution.

The application of diffusion approximations in the study of queueing systems was introduced in 1961 by Kingman [15]. Subsequently Iglehart [11], Kingman [16], and Newell [24] provided substantial results in 1965. These approximations have also been applied to problems found in such diverse

fields as statistics, engineering, physics, genetics and neurophysiology. An interesting historical review of the use of the diffusion equation is in reference 14.

When employed in describing queueing systems with congestion or heavy traffic, diffusion approximations provide very accurate results. A system is considered congested when the traffic intensity, ρ , is never much less than unity (say, $\rho > .70$), where the traffic intensity generally measures the ratio of the system's input rate to its output rate. The system we are modeling should fit into this category. The fixed cost of each channel is extremely high, and thus, the imputed cost of server idle time is high. Therefore, it is reasonable to expect the number of service channels will be kept to a minimum, forcing the traffic intensity to be high in many real situations.

The previous work done in approximating the line length distribution for multi-server queues can be divided into three categories: 1) $\rho < 1$, 2) $\rho > 1$, and 3) $\rho = 1$. The third category is highly unlikely in practice, so this case will not be discussed.

In the case $\rho < 1$, Iglehart [11] developed approximations for the M/M/C queue and a machine-repairman problem. Although, he provides weak convergence results in some extreme cases, his approximations are generally not very accurate because he did not restrict the queue lengths to be non-negative. Halachmi and Franta [10] incorporated this restriction into their analysis and demonstrated good approximations for the GI/M/C queue. Their work will be discussed in greater detail later in this section. Fischer [6] introduced an approximation method for the distribution of the virtual waiting time in an M/M/1 queue subject to breakdowns. His analysis, however, cannot be extended to multiple

server systems nor does it help in approximating line length distributions.

When $\rho > 1$, the number of units in the queue will become unbounded, almost surely, as time goes to infinity. Although a stationary probability distribution does not exist, we can explore the transient distribution. The non-stationary problem is not discussed until Chapter 5. For continuity of development, these diffusion approximation methods will be covered in this section. Iglehart and Whitt [12] developed a diffusion approximation to the transient distribution of the line length for a GI/G/C queue. Their results are accurate for the case when $\rho > 1$. Although they prove some weak convergence results, Newell [27] provides an improvement to the transient distribution for all categories of traffic intensity. Both of these approaches will be covered in more detail later.

Diffusion approximations are developed by essentially replacing the queueing system's discrete state space with a continuous state space. Conditions are imposed on the continuous state space so that the newly defined process captures the characteristics of the original process. When modeling a queueing process in this manner, one gets a partial differential equation with boundary conditions that play a rather natural role. This differential equation is called the diffusion equation and its solution will yield a probability density function. Since we are trying to find a distribution describing a discrete process, the density function will have to be integrated over specific intervals to yield the desired approximation.

In this section we first derive the basic diffusion equation and the required boundary conditions. We will then review the previous work

done in obtaining diffusion approximations to the line length distributions for multi-server queues. Then an approximation will be derived for the stationary line length distribution for a multi-server queue subject to server breakdown. Numerical examples will be given to compare this approximation to analytic and simulation results.

3.3.1 Derivation Of The Diffusion Equation

This section will discuss some of the underlying assumptions necessary to develop a diffusion approximation, introduce some new notions and notation, and then display the derivation of the diffusion equation and the boundary conditions imposed on it. The derivation follows one found in Kleinrock. (See [17], pp. 69-71.)

The first assumption, which was previously mentioned, is that we will only be examining queues with heavy traffic. Thus it is reasonable to expect that the number of units in the system, $N(t)$, is relatively large compared to unity. On a coarse scale of measurement, $N(t)$ changes very little in a short period of time. Although $N(t)$ is discrete, it is mathematically convenient to view it as a continuous random variable, $X(t)$, and thereby allow "infinitesimal" queue changes. To be consistent with the nature of the queue being modeled, we will define $X(t)$ as a continuous-time, continuous-state Markov process with conditional transition probability

$$F(w,\tau;y,t) = \Pr[X(t) \leq y \mid X(\tau) = w] \text{ for } \tau < t. \quad (3.46)$$

So $F(w,\tau;y,t)$ is the probability that the process $X(t)$ is no greater than state y at time t given that it was in state w at time τ .

We will assume that the conditional probability density function, $f(w,\tau;y,t)$, exists, is continuous and twice differentiable. Then we

have

$$f(w, \tau; y, t) = \frac{\partial F(w, \tau; y, t)}{\partial y} \quad (3.47)$$

This density function satisfies the Chapman-Kolmogorov equation

$$f(w, \tau; y, t) = \int_{-\infty}^{\infty} f(x, u; y, t) f(w, \tau; x, u) dx \quad \text{for } \tau < u < t. \quad (3.48)$$

Define the conditional mean, $M(x, \tau; t)$, to be the expected value of the process X at time t , given it was at x at time τ . Define the conditional variance, $V(x, \tau; t)$, in the same manner. Thus we have:

$$M(x, \tau; t) = E[X(t) \mid X(\tau) = x] \quad \text{for } \tau < t, \quad (3.49)$$

$$V(x, \tau; t) = E \{ [X(t) - M(x, \tau; t)]^2 \mid X(\tau) = x \} \quad \text{for } \tau < t. \quad (3.50)$$

Notice that these moments are dependent on the state of the process as well as the time. This is an important distinction which is valuable in approximating multi-server queue distributions (as opposed to single server queues). We shall assume that these moments have continuous derivatives which are defined as the infinitesimal mean, $m(x, t)$, and the infinitesimal variance, $\sigma^2(x, t)$. Specifically we have:

$$m(x, t) = \left. \frac{\partial M(x, t; \tau)}{\partial \tau} \right|_{\tau=t} \quad (3.51)$$

$$\sigma^2(x, t) = \left. \frac{\partial V(x, t; \tau)}{\partial \tau} \right|_{\tau=t} \quad (3.52)$$

Incorporating definition (3.49) and the fact that $M(x, t; t) = x$, we have

$$\begin{aligned} m(x, t) &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} [M(x, t; t+\Delta t) - M(x, t; t)] \\ &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \left[\int_{-\infty}^{\infty} y f(x, t; y, t+\Delta t) dy - x \right] \\ &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{-\infty}^{\infty} (y-x) f(x, t; y, t+\Delta t) dy \quad (3.53) \end{aligned}$$

Similarly,

$$\sigma^2(x,t) = \lim_{\Delta t \rightarrow 0} \int_{-\infty}^{\infty} (y-x)^2 \frac{f(x,t;y,t+\Delta t)}{\Delta t} dy, \quad (3.54)$$

and, in general, we define

$$E_n(x,t) = \lim_{\Delta t \rightarrow 0} \int_{-\infty}^{\infty} (y-x)^n \frac{f(x,t;y,t+\Delta t)}{\Delta t} dy, \quad n=1,2,\dots \quad (3.55)$$

where $E_n(x,t)$ is the infinitesimal n th moment. These relations will become useful later.

Now we will derive the forward diffusion equation by employing an expedient analytical technique. An arbitrary integral, I , will be defined. Then an alternate representation using Taylor series will be derived. By taking the difference of these two integrals and by employing a theorem of integration we will obtain the diffusion equation.

Consider an arbitrary function $Q(y)$ which is infinitely differentiable and sufficiently bounded so that the integral

$$I = \int_{-\infty}^{\infty} Q(y) \frac{\partial f(w,\tau;y,t)}{\partial t} dy \quad (3.56)$$

is well defined. Using the definition of a partial derivative and the Chapman-Kolmogorov equation (3.48) we get:

$$\begin{aligned} I &= \int_{-\infty}^{\infty} Q(y) \lim_{\Delta t \rightarrow 0} \left[\frac{f(w,\tau;y,t+\Delta t) - f(w,\tau;y,t)}{\Delta t} \right] dy \\ &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \left\{ \int_{-\infty}^{\infty} Q(y) \left[\int_{-\infty}^{\infty} f(w,\tau;x,t) f(x,t;y,t+\Delta t) dx \right] dy \right. \\ &\quad \left. - \int_{-\infty}^{\infty} Q(y) f(w,\tau;y,t) dy \right\} \quad (3.57) \end{aligned}$$

Let $I = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} (I_1 - I_2)$, where I_1 and I_2 are the two integrals above.

Examining I_1 alone, we substitute the Taylor series expansion for Q about x and then interchange order of integration to obtain

$$\begin{aligned}
 I_1 &= \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{1}{n!} (y-x)^n \frac{d^n Q(x)}{dx^n} \left[\int_{-\infty}^{\infty} f(w, \tau; x, t) f(x, t; y, t+\Delta t) dx \right] dy \\
 &= \int_{-\infty}^{\infty} f(w, \tau; x, t) \left[\sum_{n=0}^{\infty} \frac{1}{n!} \frac{d^n Q(x)}{dx^n} \int_{-\infty}^{\infty} (y-x)^n f(x, t; y, t+\Delta t) dy \right] dx.
 \end{aligned} \tag{3.58}$$

Careful examination of the integrand reveals that the inner integral is the definition for the n th moment.

Taking the limit of I_1 we get

$$\begin{aligned}
 \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} I_1 &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{-\infty}^{\infty} Q(x) f(w, \tau; x, t) dx + \int_{-\infty}^{\infty} f(w, \tau; x, t) \left[\sum_{n=1}^{\infty} \frac{1}{n!} E_n(x, t) \frac{d^n Q(x)}{dx^n} \right] dx \\
 &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} I_2 + \sum_{n=1}^{\infty} \frac{1}{n!} \int_{-\infty}^{\infty} f(w, \tau; x, t) E_n(x, t) \frac{d^n Q(x)}{dx^n} dx.
 \end{aligned} \tag{3.59}$$

Substituting back into (3.57) we get

$$I = \sum_{n=1}^{\infty} \frac{1}{n!} \int_{-\infty}^{\infty} f(w, \tau; x, t) E_n(x, t) \frac{d^n Q(x)}{dx^n} dx. \tag{3.60}$$

The n th term ($n = 1, 2, \dots$) in this equation can be integrated by parts n times to remove derivatives of Q . For example, define

$$I_n = \int_{-\infty}^{\infty} f(w, \tau; x, t) E_n(x, t) \frac{d^n Q(x)}{dx^n} dx, \quad n = 1, 2, \dots \tag{3.61}$$

Let

$$u = f(w, \tau; x, t) E_n(x, t) \text{ and } dv = \frac{d^n Q(x)}{dx^n} dx.$$

Then

$$\begin{aligned} I_n &= uv \Big|_{x=-\infty}^{\infty} - \int_{-\infty}^{\infty} v du \\ &= f(w, \tau; x, t) E_n(x, t) \frac{d^{(n-1)} Q(x)}{dx^{(n-1)}} \Big|_{x=-\infty}^{\infty} \\ &\quad - \int_{-\infty}^{\infty} \left[\frac{d^{(n-1)} Q(x)}{dx^{(n-1)}} \right] \left[\frac{\partial}{\partial x} [f(w, \tau; x, t) E_n(x, t)] \right] dx \\ &= - \int_{-\infty}^{\infty} \frac{d^{(n-1)} Q(x)}{dx^{(n-1)}} \left[\frac{\partial}{\partial x} [f(w, \tau; x, t) E_n(x, t)] \right] dx, \quad n = 1, 2, \dots, \end{aligned} \quad (3.62)$$

since by previous assumption $Q(x)$ and its derivatives must vanish at $\pm \infty$. Thus by an inductive argument we have

$$I = \int_{-\infty}^{\infty} Q(x) \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial x^n} [E_n(x, t) f(w, \tau; x, t)] dx. \quad (3.63)$$

Subtracting this equation from the original definition of I , (3.56),

yields

$$0 = \int_{-\infty}^{\infty} Q(x) \left\{ \frac{\partial f(w, \tau; x, t)}{\partial t} - \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial x^n} [E_n(x, t) f(w, \tau; x, t)] \right\} dx. \quad (3.64)$$

By assumption, $Q(x)$ was an arbitrary function. Thus the second factor in the integrand must be identically zero, giving

$$\frac{\partial f(w, \tau; x, t)}{\partial t} = \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial x^n} [E_n(x, t) f(w, \tau; x, t)]. \quad (3.65)$$

It is reasonable to expect the conditional density, f , to be "tightly

concentrated" around the value x . Thus we shall assume that the third and higher infinitesimal moments are negligible (i.e., $E_n(x,t) = 0$, $n = 3, 4, \dots$). We finally get the second order partial differential equation

$$\frac{\partial f}{\partial t} = - \frac{\partial}{\partial x} [m(x,t)f] + \frac{1}{2} \frac{\partial^2}{\partial x^2} [\sigma^2(x,t)f]. \quad (3.66)$$

This equation is known as the forward diffusion equation, the one-dimensional Fokker-Plank equation and the forward Kolmogorov equation. When the infinitesimal mean and variance are not time dependent, the process $X(t)$, which this equation describes, is defined as a stationary Ornstein-Uhlenbeck process. When the infinitesimal mean and variance are constants, the process defined is a Brownian motion or Weiner process with drift. Thus, in some cases, the normal probability density or Gaussian function is a solution to the diffusion equation.

Boundary conditions must be imposed upon the density, f , in order to assure a unique and meaningful solution to the diffusion equation. The most natural conditions are to require f to be a probability density which is non-zero in the positive quadrant. By this we mean

$$f(w, \tau; x, t) = 0, \text{ for } x < 0 \text{ and when } t = \tau, x \neq w, \quad (3.67)$$

and

$$\int_0^{\infty} f(w, \tau; x, t) dx = 1, \text{ for } t > \tau. \quad (3.68)$$

As the process $X(t)$ wanders through its domain, we expect it to spend very little time around its lower limit, zero. Most of the probability mass of $X(t)$ should be in the tail of the distribution and so it is natural to impose a boundary condition on the diffusion equation to enhance this concept. We can take advantage of the first two con-

ditions to derive another boundary condition, often termed the "reflecting barrier" condition. The first step is to integrate the diffusion equation,

$$\int_y^{\infty} \frac{\partial f}{\partial t} dx = - \int_y^{\infty} \frac{\partial}{\partial x} [m(x,t)f] dx + \frac{1}{2} \int_y^{\infty} \frac{\partial^2}{\partial x^2} [\sigma^2(x,t)f] dx$$

$$\frac{\partial}{\partial t} \int_y^{\infty} f dx = -m(x,t)f \Big|_{x=y}^{\infty} + \frac{1}{2} \frac{\partial}{\partial x} [\sigma^2(x,t)f] \Big|_{x=y}^{\infty} \quad (3.69)$$

The nature of the system we are modeling suggests that $X(t)$ will not become infinite in finite time. Thus condition (3.68) and the continuity of f yield

$$\lim_{x \rightarrow \infty} f(w, \tau; x, t) = 0, \quad (3.70)$$

$$\lim_{x \rightarrow \infty} \frac{\partial}{\partial x} f(w, \tau; x, t) = 0, \text{ and} \quad (3.71)$$

$$\frac{\partial}{\partial t} \int_0^{\infty} f dx = \frac{\partial}{\partial t} [1] = 0. \quad (3.72)$$

By taking the limit of (3.69) we get the boundary condition

$$\lim_{x \rightarrow 0} \left\{ -m(x,t)f + \frac{1}{2} \frac{\partial}{\partial x} [\sigma^2(x,t)f] \right\} = 0. \quad (3.73)$$

The diffusion equation and the three boundary conditions derived above, given in this general form, have never been solved analytically. If the infinitesimal mean and variance are expressed as functions of only one variable or as constants, then a solution can be found. The remainder of this section will display these solutions.

3.3.2 Previous Approximations For Queueing Systems

The key to finding an accurate diffusion approximation for a queueing system is in finding good representations for the infinitesimal

mean, $m(x,t)$, and variance, $\sigma^2(x,t)$. In order to find these, it is convenient to think of the random variable $N(t)$ as the difference of two random variables, $N(t) = A(t) - D(t)$, where $A(t)$ represents the arrival process and $D(t)$ represents the departure process. The distributions underlying $A(t)$ and $D(t)$ then determine whether $m(x,t)$ and $\sigma^2(x,t)$ are constants or functions of one variable. We will explore three diffusion approximations which differ in infinitesimal moments and traffic intensity.

3.3.2.1 Stationary Distributions For The GI/M/C Queue

The most accurate approximation of the stationary line length distribution for the GI/M/C queue was developed by Halachmi and Franta [10]. For stationary results they assume the traffic intensity is less than unity ($\rho < 1$). They were able to capture the nature of this system by making the infinitesimal moments depend upon the state of the system. The methodology used to arrive at these moments will be displayed first and then the solution to the diffusion equation will be provided. Finally, some comparative results will be discussed.

We start by assuming that the number of units in the system is continuous-valued random variable, $X(t)$. Letting $X(t) = A(t) - D(t)$, we can define

$$\begin{aligned}
 M(x,t) &= \lim_{\Delta t \rightarrow 0} \frac{E[X(t+\Delta t) - X(t) | X(t) = x]}{\Delta t} \\
 &= \lim_{\Delta t \rightarrow 0} \frac{E[A(t+\Delta t) - A(t) | X(t) = x]}{\Delta t} - \\
 &\quad \lim_{\Delta t \rightarrow 0} \frac{E[D(t+\Delta t) - D(t) | X(t) = x]}{\Delta t}
 \end{aligned} \tag{3.74}$$

The arrival process is independent of the state of the system so we can drop the conditional statement in its expectation. Since we are interested in obtaining steady state results, we need to find the limit of (3.74) as $t \rightarrow \infty$. Since $A(t)$ is a renewal (or counting) process, we can use Blackwell's Theorem [4] if we assume that the interarrival times are non-lattice (or non-arithmetic) random variables. This theorem yields

$$\lim_{t \rightarrow \infty} E[A(t+\Delta t) - A(t)] = \lambda \Delta t, \quad (3.75)$$

where λ is the interarrival rate.

The departure process does depend upon the state of the system. Since the interdeparture times for each active server are independent and exponentially distributed, we can take advantage of the memoryless property of this distribution. Let μ be the service rate of an active channel. To accommodate the continuity assumption for the system's state space, we assume that the servers act as independent infinitesimal units. This allows the definition

$$\Pr[D(t+\Delta t) - D(t) > 0 | X(t) = x] = \begin{cases} x\mu\Delta(t) + O(\Delta t), & 0 \leq x < C \\ c\mu\Delta(t) + O(\Delta t), & x > C, \end{cases} \quad (3.76)$$

which together with Blackwell's Theorem gives

$$\lim_{t \rightarrow \infty} E[D(t+\Delta t) - D(t) | X(t) = x] = \begin{cases} x\mu\Delta t, & 0 \leq x < C \\ C\mu\Delta t, & x \geq C \end{cases} \\ = \min(x, C)\mu\Delta t. \quad (3.77)$$

Thus, using (3.74) the infinitesimal mean is

$$m(x) = \lambda - \min(x, C)\mu. \quad (3.78)$$

Similarly, we can define the infinitesimal variance

$$\begin{aligned}\sigma^2(x, t) &= \lim_{\Delta t \rightarrow 0} \frac{\text{Var}[X(t+\Delta t) - X(t) | X(t) = x]}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{\text{Var}[A(t+\Delta t) - A(t)]}{\Delta t} + \lim_{\Delta t \rightarrow 0} \frac{\text{Var}[D(t+\Delta t) - D(t) | X(t) = x]}{\Delta t},\end{aligned}\tag{3.79}$$

where we again assume that the arrival process is independent of the state of the system. Since we do not have the equivalent of Blackwell's Theorem for the variance of a renewal process we need to employ a transformation to fit the conditions of a known renewal theorem. Define

$\Delta A(t) = A(t + \Delta t) - A(t)$, $n = [t/\Delta t]$, and $\Delta A_i = A(i \cdot \Delta t) - A((i-1) \cdot \Delta t)$, where the brackets, $[\cdot]$, denote the greatest integer function. Then by our assumption that arrivals are independent, we have

$$v = \text{Var} \left[\frac{1}{n} \sum_{i=1}^n \Delta A_i \right] = \text{Var} [\Delta A(t)]/n.\tag{3.80}$$

But assuming $A(0) = 0$ we also get

$$\begin{aligned}v &= \text{Var} \left[\frac{1}{n} \sum_{i=1}^n \Delta A_i \right] = \text{Var} \left\{ [A(n \cdot \Delta t) - A(0)]/n \right\} \\ &= \frac{1}{n^2} \text{Var}[A(t)]\end{aligned}\tag{3.81}$$

Combining (3.80) and (3.81) yields

$$\text{Var}[\Delta A(t)] = nV = \frac{1}{n} \text{Var} [A(t)] = \text{Var} [A(t)]/[t/\Delta t]\tag{3.82}$$

Taking limits we get the equivalent statements

$$\begin{aligned}\lim_{t \rightarrow \infty} \text{Var} [\Delta A(t)] &= \lim_{t \rightarrow \infty} \left\{ \text{Var}[A(t)]/[t/\Delta t] \right\} \\ &= \lim_{t \rightarrow \infty} \left\{ \text{Var}[A(t)] \Delta t/t \right\}.\end{aligned}\tag{3.83}$$

From renewal theory (see [28], pp. 180) we have

$$\lim_{t \rightarrow \infty} \frac{\text{Var} [A(t)]}{t} = \lambda^3 \sigma_a^2, \quad (3.84)$$

where σ_a^2 is the variance of the interarrival times. So we finally get

$$\lim_{t \rightarrow \infty} \text{Var} [A(t+\Delta t) - A(t)] = \lambda^3 \sigma_a^2 \Delta t. \quad (3.85)$$

In a similar manner, we can find the variance of the departure process, keeping in mind the dependence on the state of the system. Since the service times for each active server are exponentially distributed, we know the variance σ_d^2 of each time is the square of the expected service time (i.e., $\sigma_d^2 = 1/\mu^2$). Therefore, we have

$$\lim_{t \rightarrow \infty} \text{Var} [D(t+\Delta t) - D(t) | X(t) = x] = \min(x, C) \mu \Delta t. \quad (3.86)$$

Thus, using (3.79), the infinitesimal variance is

$$\sigma^2(x) = \lambda^3 \sigma_a^2 + \min(x, C) \mu. \quad (3.87)$$

The diffusion equation (3.66) derived in the previous section relates a change in state to a change in time. Since we seek a stationary distribution, we must discard the conditioning on the initial state and take the limit as $t \rightarrow \infty$. Thus

$$\lim_{t \rightarrow \infty} f(w, \tau; x, t) = f(x), \quad \text{and} \quad \lim_{t \rightarrow \infty} \frac{\partial f}{\partial t} = 0$$

are necessary for the existence of a stationary distribution. We then get the diffusion equation

$$\frac{1}{2} \frac{d^2}{dx^2} [\sigma^2(x) f(x)] - \frac{d}{dx} [m(x) f(x)] = 0, \quad (3.88)$$

and the associated boundary conditions

$$\frac{1}{2} \frac{d}{dx} [\sigma^2(x) f(x)] \Big|_{x=0} - m(0) f(0) = 0, \quad (3.89)$$

and

$$\int_0^1 f(x) dx = 1. \quad (3.90)$$

The solution arrived at for this differential equation by Halachmi and Franta [10] is

$$f(x) = \begin{cases} H_1 [\sigma^2(x)]^{u-1} \exp(-2x) & , \quad 0 < x = C \\ H_2 \exp\left[\frac{2m(C)x}{\sigma^2(C)}\right] & , \quad x > C, \end{cases} \quad (3.91)$$

where

$$u = \frac{2\lambda}{\mu} [(\lambda\sigma_a)^2 + 1], \quad (3.92)$$

$$\int_0^{\infty} f(x) dx = 1, \quad (3.93)$$

and

$$H_1 [\sigma^2(C)]^{u-1} \exp(-2C) = H_2 \exp\left[\frac{2m(C) \cdot C}{\sigma^2(C)}\right]. \quad (3.94)$$

Conditions (3.93) and (3.94) solve precisely for the constants H_1 and H_2 . Condition (3.94) requires f to be continuous. The solution, f , is a probability density function. To get the approximation for the distribution of the number of units, N , in the system, we define

$$\tilde{p}_n = \Pr [N = n] = \int_{n-.5}^{n+.5} f(x) dx \quad , \quad n = 1, 2, \dots \quad (3.95)$$

Certain adjustments must be made since the approximation for p_0 is not well defined. These are discussed in another section.

Some comparisons can be made between the diffusion approximation and the analytical solution for the GI/M/C queue, but unfortunately only for $n \geq C$. A known queueing result for the GI/M/C queue [8] states

that when $\rho = \lambda/C\mu < 1$, the stationary probability distribution, p_n , exists and there is an r , $0 < r < 1$, such that

$$p_n = Ar^n, \quad n \geq C. \quad (3.96)$$

When we evaluate the approximating density (3.91), we get

$$\begin{aligned} \tilde{p}_n &= \int_{n-.5}^{n+.5} f(x) dx \\ &= \frac{H_2 \sigma^2(C)}{2m(C)} \left\{ \exp \left[\frac{2m(C)}{\sigma^2(C)} (n+.5) \right] - \exp \left[\frac{2m(C)}{\sigma^2(C)} (n-.5) \right] \right\} \\ &= Ks^n, \quad n \geq C, \end{aligned} \quad (3.97)$$

where

$$s = \exp \left[\frac{2m(C)}{\sigma^2(C)} \right], \quad (3.98)$$

and

$$\begin{aligned} K &= \frac{H_2 \sigma^2(C)}{2m(C)} [s^{n+.5} - s^{n-.5}] \\ &= K' (1 - s). \end{aligned} \quad (3.99)$$

Thus, comparing (3.96) and (3.97), we can see that p_n and \tilde{p}_n agree in form.

Examining the M/M/C queue yields some convergence results. Equation (3.96) becomes

$$p_n = Ap^n, \quad n \geq C, \quad \text{where } A = (1-\rho)/\rho \text{ for } C = 1. \quad (3.100)$$

The infinitesimal mean and variance in (3.98) become

$$m(C) = \lambda - C\mu \text{ and } \sigma^2(C) = \lambda^3 \left(\frac{1}{\lambda^2} \right) + C\mu = \lambda + C\mu,$$

so that the multiplicative factor in (3.97) is

$$s = \exp \left[\frac{2(\lambda - C\mu)}{\lambda + C\mu} \right] = \exp \left[\frac{-2(1-\rho)}{1+\rho} \right]. \quad (3.101)$$

Thus as $\rho \rightarrow 1$, then $s \rightarrow 1$. Moreover, for an M/M/1 queue, Kobayashi [18] points out that if we define $\tilde{p}_0 = 1 - \rho$, then the approximation (3.97) will become

$$\tilde{p}_n = \begin{cases} 1 - \rho & , \quad n = 0 \\ \rho(1 - s)s^{n-1} & , \quad n \geq 1 \end{cases} \quad (3.102)$$

As well as being very close in form to (3.100), this suggests that convergence is rapid as $\rho \rightarrow 1$.

Basically these convergence results are due to the fact that $\sigma_a^2 = 1/\lambda^2$ which simplifies $\sigma^2(C)$. This suggests that the diffusion approximation developed here would also be accurate for systems which have a coefficient of variation for the arrival process close to unity (i.e., $\sigma_a^2/\lambda^2 \approx 1$).

3.3.2.2 Transient Distributions For The GI/G/C Queue

Iglehart and Whitt [12] explored a diffusion approximation for the transient line length distribution for the GI/G/C queue. They considered the cases $\rho = 1$ and $\rho > 1$. The latter case, only, will be discussed. Their derivations do not involve the solution of the diffusion equation; they are lengthy and involve much notation, so only the major results will be displayed. The analysis involves modifying the queueing system, finding an approximation for the modified process, and proving the same results are good for the unmodified process. Since the approximation converges it can be expressed as a theorem.

Theorem 3.5. (Iglehart and Whitt)

Suppose we are given a GI/G/C queue with $N(0) = 0$.

Let λ represent the arrival rate of units,

μ represent the service rate of a busy server,

σ_a^2 represent the variance of interarrival times, and

σ_d^2 represent the variance of the service times.

Define $\gamma^2 = \lambda^3 \sigma_a^2 + C \mu^3 \sigma_s^2$. If $\rho = \frac{\lambda}{C\mu} > 1$,

then

$$\lim_{t \rightarrow \infty} \Pr \left\{ \frac{N(t) - (\lambda - \mu)t}{\gamma t^{1/2}} \leq x \right\} = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^x \exp(-y^2/2) dy. \quad (3.103)$$

This approximation is a direct solution of the diffusion equation without regard for the boundary conditions or the dependence of the infinitesimal moments on the size of the queue. If $N(t)$ were allowed to be continuous, then it would represent a Brownian motion process with drift which would give positive probability to negative line lengths. Since the process "drifts" away from zero, the approximation becomes more accurate as time progresses. If an initial condition, say $N(0) = k$, is introduced, where k is large, then the approximation would also be improved.

Newell [27] points out that when some of the boundary conditions are ignored, the solution found for the diffusion equation may not be unique. For this case, the diffusion equation is

$$\frac{\partial f}{\partial t} = -m \frac{\partial f}{\partial x} + \frac{\sigma^2}{2} \frac{\partial^2 f}{\partial x^2} \quad (3.104)$$

and the boundary conditions are

$$f(w, 0; x, t) = 0, \text{ for } x < 0 \text{ and } t = 0, x \neq w, \quad (3.105)$$

$$\int_0^{\infty} f(w, 0; x, t) dx = 1, \text{ for } t \geq 0, \quad (3.106)$$

and

$$-mf(w,0;0,t) + \frac{\sigma^2}{2} \left. \frac{\partial f(w,0,x,t)}{\partial x} \right|_{x=0} = 0. \quad (3.107)$$

Using the same notation defined in Theorem 3.5, Newell provides the following solution

$$f(w,0;x,t) = \frac{1}{\sqrt{2\pi\gamma^2 t}} \left\{ \exp \left[\frac{-[x-w-(\lambda-\mu)t]^2}{2\gamma^2 t} \right] + \exp \left[\frac{+2x(\lambda-\mu)}{\gamma^2} \right] \right. \\ \left. \left[\exp \left[\frac{-[x+w+(\lambda-\mu)t]^2}{2\gamma^2 t} \right] - \frac{2(\lambda-\mu)}{\gamma^2} \int_x^\infty \exp \left[\frac{-[y+w+(\lambda-\mu)t]^2}{2\gamma^2 t} \right] dy \right] \right\}. \quad (3.108)$$

This solution satisfies all the boundary conditions and is valid for all values of ρ . Notice Iglehart and Whitt's approximation, (3.103), corresponds to the first term in this solution. They actually provide a solution to the diffusion equation (3.104) but do not meet the boundary conditions. The second term in Newell's solution is a correcting factor so that the reflecting barrier condition (3.107) is met.

3.3.3 Approximation For GI/M/C Queue Subject To Server Breakdown

The method we use to find an approximation for the line length distribution in a GI/M/C queue subject to server breakdown is the same as the one discussed in Section 3.3.2.1. A representation is found for the infinitesimal moments, the diffusion equation is solved and the derived density is integrated to get the approximation for the stationary distribution of the number of units in the system. In this section, we shall also restrict the traffic intensity, $\rho = \lambda/\mu\bar{C}$, to be less than one. After the approximation is developed, a region in which to expect the best results is created. Finally, some numerical examples are given

so that we may compare the approximation to analytic and simulation results.

3.3.3.1 Derivation Of Infinitesimal Moments

We start, as before, by assuming that the number of units in the system is a continuous random variable, $X(t)$. We then express $X(t)$ as the difference between the arrival process, and the departure process, that is, $X(t) = A(t) - D(t)$. Given that $C(t)$ is the number of operational service channels, we can define the infinitesimal mean

$$\begin{aligned} m(x,k,t) &= \lim_{h \rightarrow 0} \frac{E[A(t+h)-D(t+h)-A(t)+D(t) | C(t)=k, X(t)=x]}{h} \\ &= \lim_{h \rightarrow 0} \frac{E[A(t+h)-A(t)]}{h} - \lim_{h \rightarrow 0} \frac{E[D(t+h)-D(t) | C(t)=k, X(t)=x]}{h} \end{aligned} \quad (3.109)$$

Notice that in this case, the departure process is dependent upon the number of operational service channels as well as the state of the system. Using the same procedure as described in Section 3.3.2.1, we have

$$\lim_{t \rightarrow \infty} E[A(t+h) - A(t)] = \lambda h,$$

and

$$\lim_{t \rightarrow \infty} E[D(t+h) - D(t) | C(t) = k, X(t) = x] = \min(X,k) \mu h \quad (3.110)$$

Thus,

$$m(x,k) = \lambda - \min(x,k) \mu. \quad (3.111)$$

Notice $m(x,k)$ depends on both x and k . We can find the infinitesimal mean $m(x)$ using $m(x,k)$ as follows:

$$\begin{aligned} m(x) &= \sum_{k=0}^C m(x,k) \Pr \{k \text{ channels operational} | x \text{ units in the system}\} \\ &= \sum_{k=0}^C [\lambda - \min(x,k) \mu] \cdot \Pr \{k|x\} \end{aligned}$$

$$= \left\{ \begin{array}{l} \lambda - (x\mu) \cdot \sum_{k=1}^C \text{Pr}\{k|x\} \quad , \quad 0 \leq x \leq 1 \\ \lambda - (x\mu) \sum_{k=2}^C \text{Pr}\{k|x\} - \mu \text{Pr}\{1|x\} \quad , \quad 1 < x \leq 2 \\ \lambda - (x\mu) \sum_{k=3}^C \text{Pr}\{k|x\} - \mu \sum_{k=1}^2 k \text{Pr}\{k|x\} \quad , \quad 2 < x \leq 3 \\ \vdots \\ \lambda - \mu \sum_{k=1}^C k \text{Pr}\{k|x\} \quad , \quad x > C. \end{array} \right.$$

(3.112)

The infinitesimal variance, can be written as follows:

$$\begin{aligned} \sigma^2(x, k, t) &= \lim_{h \rightarrow 0} \frac{\text{Var} [A(t+h) - D(t+h) - A(t) + D(t) | C(t)=k, X(t)=x]}{h} \\ &= \lim_{h \rightarrow 0} \frac{\text{Var}[A(t+h) - A(t)]}{h} + \lim_{h \rightarrow 0} \frac{\text{Var}[D(t+h) - D(t) | C(t)=k, X(t)=x]}{h} \end{aligned}$$

(3.113)

Again we know that

$$\lim_{t \rightarrow \infty} \text{Var}[A(t+h) - A(t)] = \lambda^3 \sigma_a^2 h,$$

(3.114)

and

$$\lim_{t \rightarrow \infty} \text{Var}[D(t+h) - D(t) | C(t)=k, X(t)=x] = \min(x, k) \mu h.$$

(3.115)

Thus

$$\sigma^2(x, k) = \lambda^3 \sigma_a^2 + \min(x, k) \mu.$$

(3.116)

Taking expectations, we see that

$$\sigma^2(x) = \sum_{k=0}^C \sigma^2(x, k) \cdot \text{Pr} \{k \text{ channels operational} \mid x \text{ units in the system}\}$$

$$= \begin{cases} \lambda^3 \sigma_a^2 + x\mu \sum_{k=1}^C \Pr\{k|x\} & , \quad 0 \leq x \leq 1 \\ \lambda^3 \sigma_a^2 + x\mu \sum_{k=2}^C \Pr\{k|x\} + \mu \Pr\{1|x\} & , \quad 1 < x \leq 2 \\ \lambda^3 \sigma_a^2 + x\mu \sum_{k=3}^C \Pr\{k|x\} + \mu \sum_{k=1}^2 k \cdot \Pr\{k|x\} & , \quad 2 < x \leq 3 \\ \vdots & \\ \lambda^3 \sigma_a^2 + \mu \sum_{k=1}^C k \Pr\{k|x\} & , \quad x > C. \end{cases}$$

(3.117)

3.3.3.2 Definition of an Approximation Region

Finding the conditional distribution $\Pr\{k|x\}$ is as difficult as solving for the distribution we wish to approximate. Using Bayes' Theorem to solve for $\Pr\{k|x\}$, where the continuous variable $X(t)$ has been replaced by the original discrete variable $N(t)$, we get

$$\Pr\{k|n\} = \frac{\Pr\{n|k\} \Pr\{k\}}{\sum_{k=0}^C \Pr\{n|k\} \Pr\{k\}} = \frac{\Pr\{k,n\}}{\Pr\{n\}} \quad (3.118)$$

If we knew any of the probabilities in the above equation, then we would not need to use approximations since solving for $\Pr\{n\}$ is our objective.

Under some conditions, however, it may be reasonable to use $\Pr\{k\}$ in place of $\Pr\{k|n\}$. Looking back at equations (3.112) and (3.117), one can see that the probabilities of interest are only $\Pr\{k|n\}$

when $n > C$. For example, in the two channel case, we wish to have

$$\Pr\{k|n\} \approx \Pr\{k\}, \text{ for } k = 1, 2 \text{ and } n = 3, 4, \dots \quad (3.119)$$

The stationary probability distribution for the number of operational channels is given by Theorem 3.1, equation (3.2).

A comparison between the conditional distribution $\Pr\{k|n\}$ and $\Pr\{k\}$ was done for one and two server systems using the results of Theorems 3.2 and 3.3. It appears that the relative error,

$\frac{\Pr\{k\} - \Pr\{k|n\}}{\Pr\{k|n\}}$, is a function of λ/ξ , μ/η and $\lambda/\bar{C}\mu$, where

$\bar{C} = \sum_{k=0}^C k\Pr\{k\}$ and the last term is the traffic intensity, ρ . As $\rho \rightarrow 1$ the relative error decreases. As λ/ξ increases beyond a certain point, the approximation improves. As μ/η increases, the approximation gets steadily worse. A proposed region for approximations which have less than ten percent relative error is as follows:

$$\rho \geq .75, \quad (3.120)$$

$$\lambda/\xi \geq 1.00, \quad (3.121)$$

$$\mu/\eta \leq 1.00. \quad (3.122)$$

The comparative data are displayed in Figures 3.3, 3.4 and 3.5. Figure 3.3 shows the results in a system having one server and no spares. Figure 3.4 shows the results for two servers and no spares. Figure 3.5 compares results for the single server, single spare system. Negative errors show up because one density does not necessarily bound the other for all values of the parameters. The natural logarithm of λ/ξ was used for the abscissa to show the insensitivity of the relative error to large changes in λ/ξ . The numerical results also exposed considerable round-off error problems when trying to apply Theorems 3.2 and 3.3.

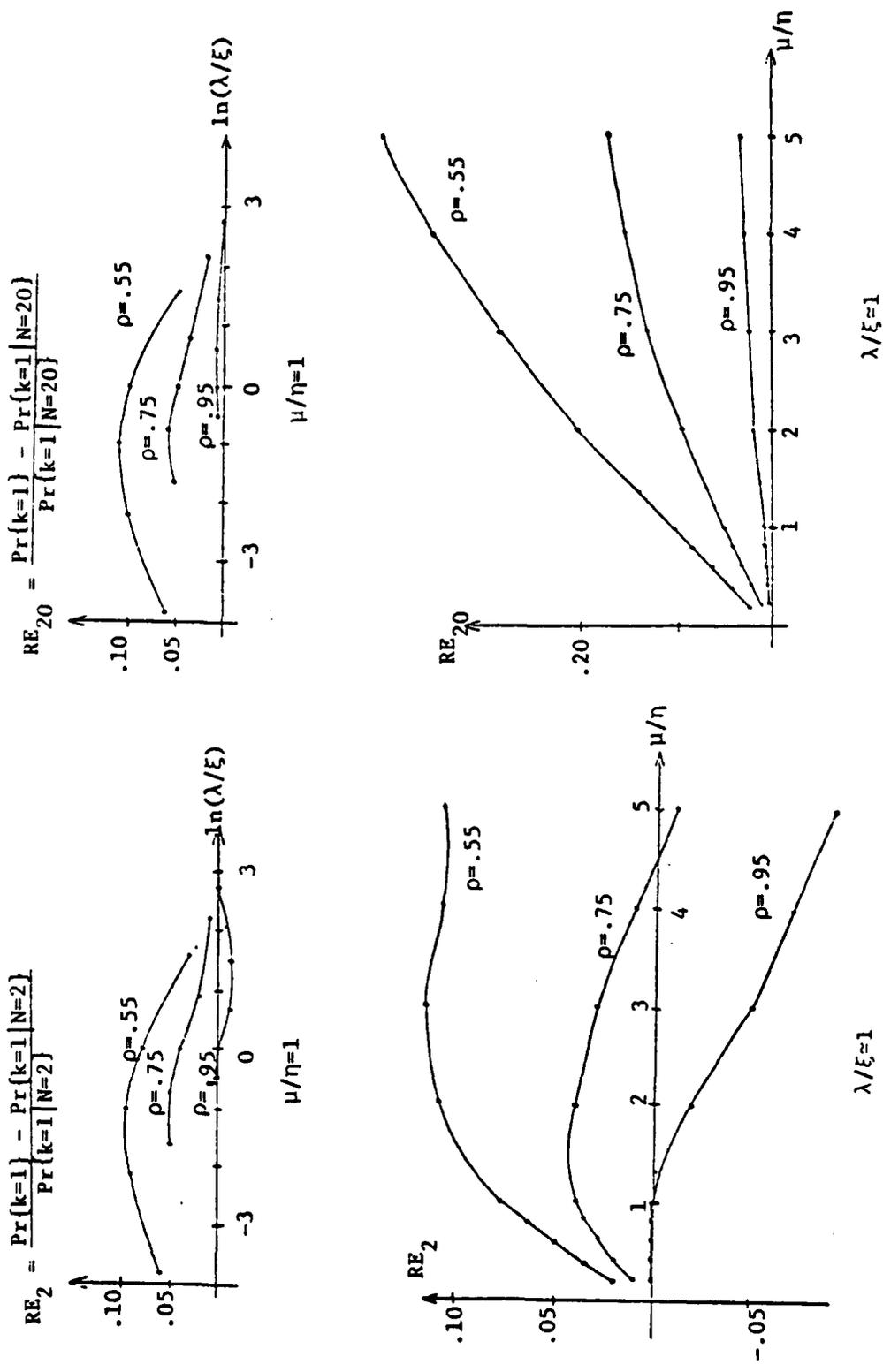


Figure 3.3: Relative Errors for Single Server, No Spare System

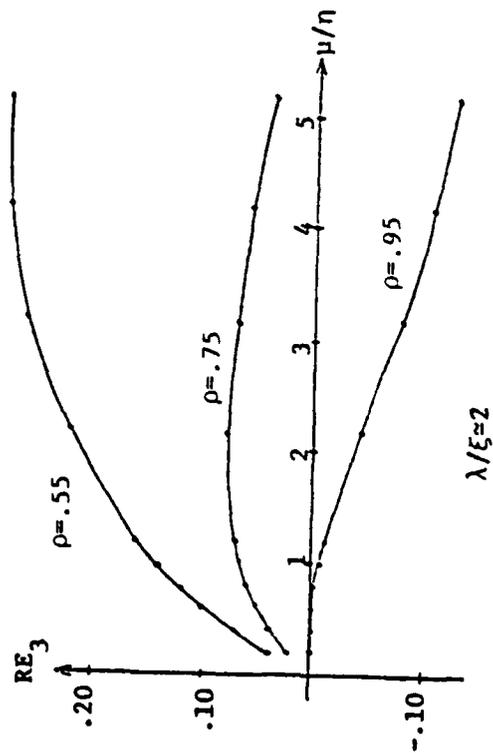
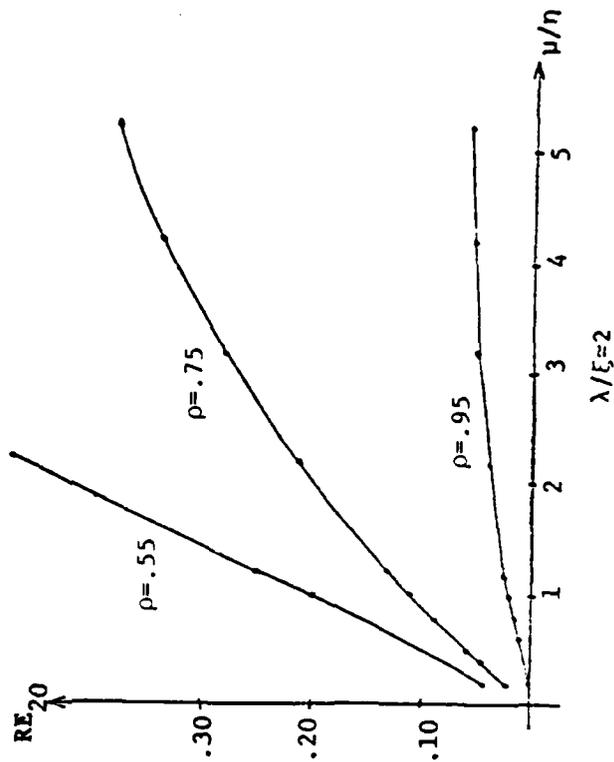
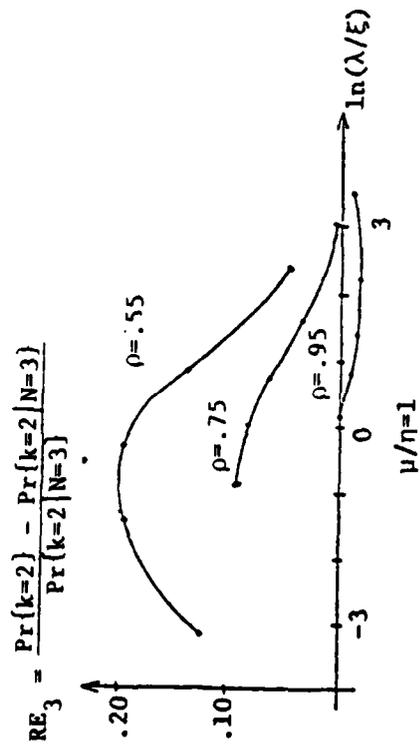
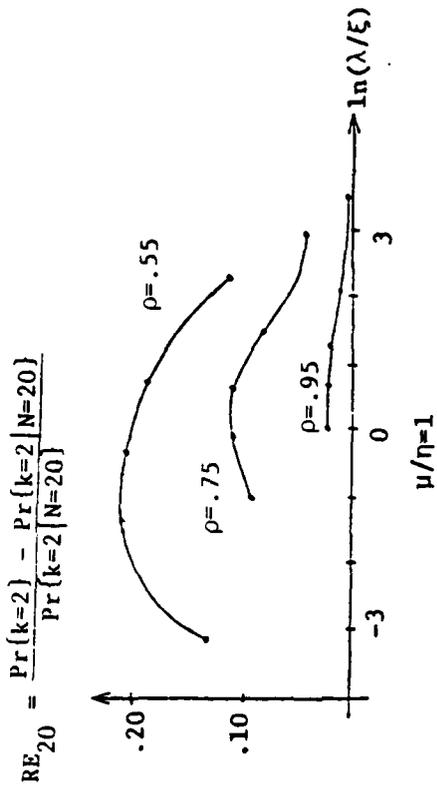


Figure 3.4: Relative Errors for Two Server, No Spare System

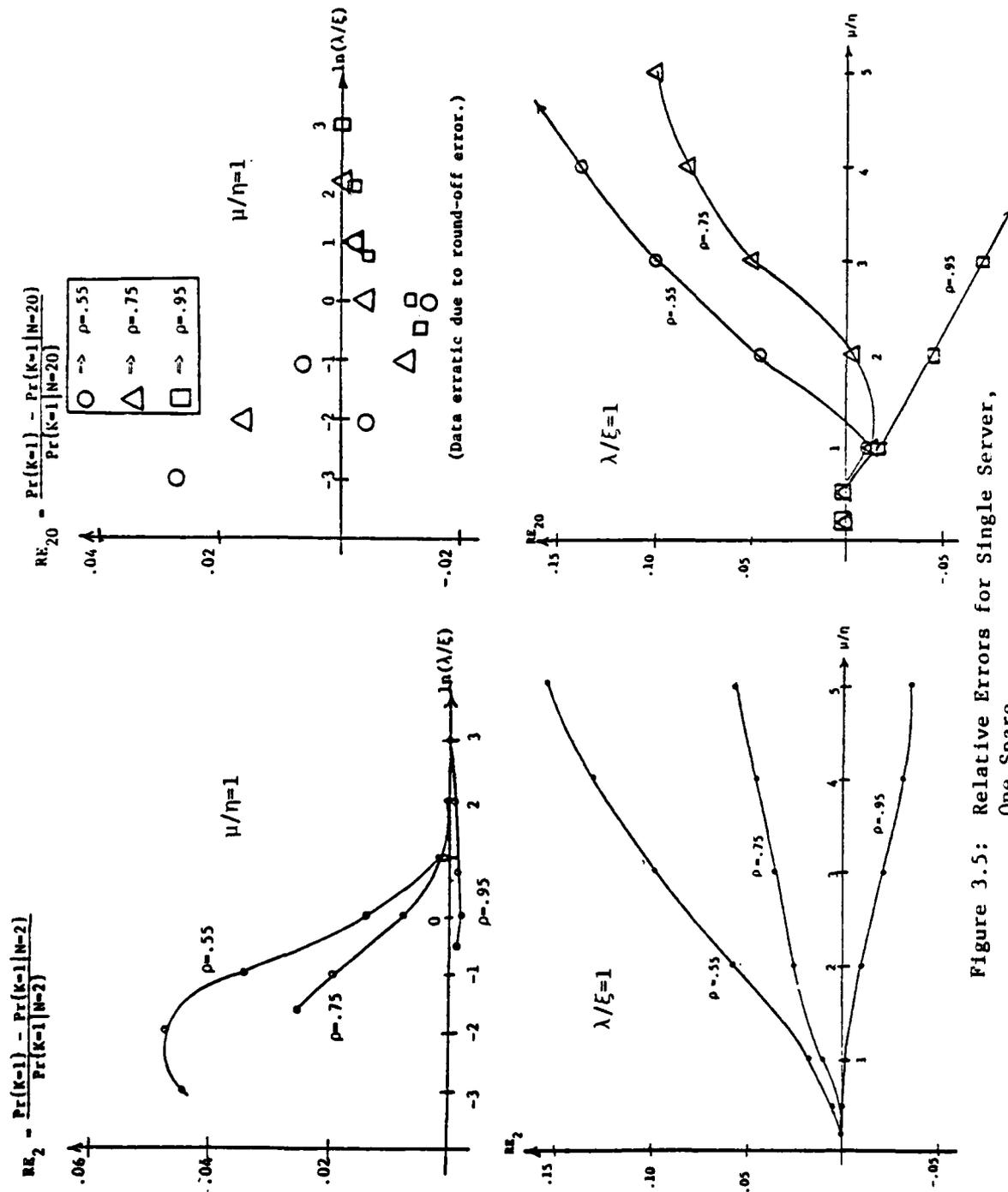


Figure 3.5: Relative Errors for Single Server, One Spare

An interesting effect was also observed when the relative sizes of the server breakdown rate and repair rate were increased while maintaining the same ratio. Suppose we fix a set of base rates, ξ and η , and define $\rho_S = K\xi / K\eta$, where $K\xi$ and $K\eta$ are the actual breakdown and repair rates. Then as K increases, $\Pr \{k\}$ remains the same (see equation (3.2)); but, the relative error decreases. These results are displayed for the single server, single spare case in Figure 3.6. It could be said that server breakdowns induce a non-stationary service rate and that the "more stationary" the service process is, the more independent the line length is from the number of operating channels. This is similar to a conjecture by Ross [30] for a single server system with non-stationary arrivals. Unfortunately, it is difficult to quantify this phenomenon and define a reasonable bound.

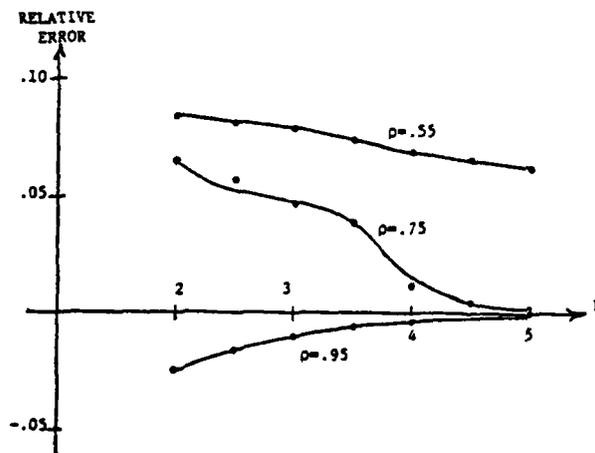


Figure 3.6: Effects of Relative Size of $K\xi$ and $K\eta$

3.3.3.3 Solution Of The Diffusion Equation

Within the region defined by relations (3.120), (3.121) and (3.122), we shall define

$$m(x) = \begin{cases} \lambda - \mu \sum_{k=1}^{n-1} k \Pr\{k\} - \mu x \sum_{k=n}^C \Pr\{k\}, & n-1 < x \leq n, n=1,2,\dots,C \\ \lambda - \mu \sum_{k=1}^C k \Pr\{k\} & , x > C \end{cases}$$

$$= \begin{cases} \lambda_n - x\mu_n & , n-1 < x \leq n, n=1,2,\dots,C \\ \lambda_{C+1} & , x > C, \end{cases} \quad (3.123)$$

where

$$\lambda_n = \lambda - \mu \sum_{k=1}^{n-1} k \Pr\{k\} \quad , n=1,\dots, C+1, \quad (3.124)$$

and

$$\mu_n = \mu \sum_{k=n}^C \Pr\{k\} \quad , n=1,\dots,C. \quad (3.125)$$

Similarly,

$$\sigma^2(x) = \begin{cases} v_n + x\omega_n & , n-1 < x \leq n, n=1,\dots,C, \\ v_{C+1} & , x > C, \end{cases} \quad (3.126)$$

where

$$v_n = \lambda^3 \sigma_a^2 + \mu \sum_{k=1}^{n-1} k^2 \Pr\{k\} \quad , n=1,\dots,C+1, \quad (3.127)$$

and

$$\omega_n = \mu \sum_{k=n}^C k \Pr\{k\} \quad , n=1,2,\dots,C. \quad (3.128)$$

These moments can now be used to solve the diffusion equation subject to the boundary conditions, which are repeated from Section 3.3.2.1,

$$\frac{1}{2} \frac{d^2}{dx^2} [\sigma^2(x)f(x)] - \frac{d}{dx} [m(x)f(x)] = 0, \quad (3.88)$$

$$\frac{1}{2} \frac{d}{dx} [\sigma^2(x)f(x)] \Big|_{x=0} - m(0)f(0) = 0, \quad (3.89)$$

$$\int_0^1 f(x) dx = 1. \quad (3.90)$$

It is easy to establish that $m(x)$ and $\sigma^2(x)$ are continuous for $x > 0$. These functions, however, are not differentiable at the lattice points $\{1, 2, \dots, C\}$. Thus one must be careful to evaluate the diffusion equation only on the interior of each interval, $(n-1, n)$, where $n = 1, 2, \dots, C$. Integration of the diffusion equation yields

$$\frac{1}{2} \frac{d}{dx} [\sigma^2(x)f(x)] - m(x)f(x) = H_n, \quad (3.129)$$

where H_n is the constant of integration peculiar to the evaluation on the interval $(n-1, n)$, $n = 1, 2, \dots, C$, and H_{C+1} is for the interval (C, ∞) .

From the reflecting boundary condition, (3.89), we get $H_1 = 0$. In order to have a proper density function, we need $\lim_{x \rightarrow \infty} f(x) = 0$ and $\lim_{x \rightarrow \infty} \frac{d f(x)}{dx} = 0$. Thus $H_{C+1} = 0$.

Let $m_n(x)$, $\sigma_n^2(x)$, and $f_n(x)$ be the respective evaluations of these functions on the intervals $(n-1, n)$ for $n = 1, \dots, C$, and (C, ∞) , for $n = C+1$. Then by continuity of $m(x)$ and $\sigma^2(x)$ we have (for $n = 1, \dots, C$)

$$m_n(n) = m_{n+1}(n), \quad (3.130)$$

$$\text{and } \sigma_n^2(n) = \sigma_{n+1}^2(n). \quad (3.131)$$

To establish the composition of f , we shall require f to be continuous.

Thus

$$f_n(n) = f_{n+1}(n), \quad n = 1, \dots, C. \quad (3.132)$$

With the three continuity relations above it is natural to set $H_n = 0$ for $n = 2, \dots, C$.

Now the differential equation (3.129) is a homogeneous, first order equation with variable coefficients. Using well known techniques, the solution, in general form, is found to be

$$f_n(x) = \frac{H_n}{\sigma_n^2(x)} \exp [K_n(x)], \quad n-1 < x \leq n \text{ for } n=1, \dots, C \text{ and } x > C \text{ for } n=C+1, \quad (3.133)$$

where
$$K_n(x) = \int \frac{2m(x)}{\sigma^2(x)} dx \quad \text{and } H_n \text{ is a new}$$

constant of integration. Examining $K_n(x)$ more closely for $n = 1, \dots, C$, we have

$$\begin{aligned} K_n(x) &= \int \frac{2(\lambda_n - x\mu_n)}{v_n + x\omega_n} dx \\ &= \frac{-2x\mu_n}{\omega_n} + 2 \left[\frac{\lambda_n}{\omega_n} + \frac{\mu_n v_n}{\omega_n^2} \right] \ln(v_n + x\omega_n). \end{aligned} \quad (3.134)$$

For $n = C + 1$,

$$K_{C+1} = \int \frac{2\lambda_{C+1}}{v_{C+1}} dx = \frac{2\lambda_{C+1}x}{v_{C+1}}. \quad (3.135)$$

One can also show that $K(x)$ is continuous for $x \geq 0$. Thus the final solution form for $f(x)$ is

$$f(x) = \begin{cases} H_n [v_n + x\omega_n]^{u-1} \exp[-2x\mu_n/\omega_n], & n-1 < x \leq n, \quad n=1, \dots, C, \\ H_{C+1} \exp [2\lambda_{C+1} x/v_{C+1}], & x > C, \end{cases} \quad (3.136)$$

where

$$u_n = \frac{2}{\omega_n} \left[\lambda_n + \frac{\mu_n \nu_n}{\omega_n} \right], \quad n=1, \dots, C. \quad (3.137)$$

It now remains to solve for H_n and evaluate \tilde{p}_n . First we shall point out an ambiguity in evaluating \tilde{p}_0 . We will use the method

$\tilde{p}_n = \int_{n-.5}^{n+.5} f(x) dx$ to compute the approximating density. This evaluation, however, is not well suited for p_0 . Methods for approximating p_0 will be discussed in the next section.

The constants H_n can be evaluated by using the continuity constraint and the last boundary condition, which requires f to be a proper density function. If we define $g_n(x)$ to be the variable portion of $f_n(x)$ in equation (3.136), then $f_n(x) = H_n g_n(x)$ and continuity requires

$$H_n g_n(n) = H_{n+1} g_{n+1}(n), \quad n = 1, \dots, C. \quad (3.138)$$

We can solve for H_n in terms of H_{C+1} using (3.138) recursively.

$$\begin{aligned} H_C &= \frac{H_{C+1} g_{C+1}(C)}{g_C(C)} \\ H_{C-1} &= \frac{H_{C+1} g_{C+1}(C) g_C(C-1)}{g_C(C) g_{C-1}(C-1)} \\ &\vdots \\ H_n &= \frac{H_{C+1} g_{C+1}(C) \dots g_{n+1}(n)}{g_C(C) \dots g_n(n)} = H_{C+1} R_n, \quad n = 1, \dots, C. \end{aligned} \quad (3.139)$$

Define

$$F_n = \int_{n-.5}^n g_n(x) dx \quad \text{and} \quad G_{n+1} = \int_n^{n+.5} g_{n+1}(x) dx, \quad n=1, \dots, C.$$

Then

$$\begin{aligned} \tilde{p}_n &= \int_{n-.5}^{n+.5} f(x) dx = \int_{n-.5}^n f_n(x) dx + \int_n^{n+.5} f_{n+1}(x) dx \\ &= H_n F_n + H_{n+1} G_{n+1}, \quad n = 1, \dots, C. \end{aligned} \quad (3.140)$$

Assuming we have a good approximation for p_0 , the boundary condition (3.90) yields

$$\begin{aligned} 1 - \tilde{p}_0 &= \int_{.5}^{\infty} f(x) dx \\ &= \sum_{n=1}^C \tilde{p}_n + \int_{C+.5}^{\infty} H_{C+1} g_{C+1}(x) dx \\ &= \sum_{n=1}^C [H_n F_n + H_{n+1} G_{n+1}] + H_{C+1} \int_{C+.5}^{\infty} \exp\left[\frac{2\lambda_{C+1} x}{\nu_{C+1}}\right] dx. \end{aligned}$$

Substituting in (3.139) gives

$$1 - \tilde{p}_0 = H_{C+1} \left\{ \sum_{n=1}^C R_n F_n + \sum_{n=2}^C R_n G_n + G_{C+1} - \frac{\nu_{C+1}}{2\lambda_{C+1}} \exp\left[\frac{2\lambda_{C+1}(C+.5)}{\nu_{C+1}}\right] \right\} \quad (3.141)$$

Finally we have

$$H_{C+1} = \frac{1 - \tilde{p}_0}{R_1 F_1 + G_{C+1} + \sum_{n=2}^C R_n [F_n + G_n] - \frac{\nu_{C+1}}{2\lambda_{C+1}} \exp\left[\frac{2\lambda_{C+1}(C+.5)}{\nu_{C+1}}\right]} \quad (3.142)$$

We can then use equation (3.139) to solve for the remaining H_n , $n=1, \dots, C$, and equation (3.140) to solve for \tilde{p}_n , $n=1, 2, \dots$.

3.3.3.4 Approximation For P_0

As pointed out in the last section, we cannot use diffusion approximations to estimate p_0 accurately. Defining $\tilde{p}_0 = \int_0^{.5} f(x) dx$ underestimates the value and empirically we find that $\tilde{p}_0 = \int_{.5}^{n+1} f(x) dx$ is also not a good estimator. Redefining all estimates $\tilde{p}_n = \int_n^{n+1} f(x) dx$, $n = 0, 1, \dots$, is also ineffective. This section will provide some analytic formulas which are good in specific cases and then discuss several alternative estimates. In all cases, the estimates have been empirically found to be at least as large as p_0 . The cases considered are briefly described below.

Case 1. M/M/1 subject to breakdown, no spares:

$$P_0 = \frac{[\mu\eta - \lambda(\eta + \xi)] (\eta + \lambda + \xi)}{\mu(\xi + \eta) (\lambda + \eta)}, \quad (3.143)$$

from Theorem 3.2, equations (3.9) and (3.10).

Case 2. M/M/2 subject to breakdowns, no spares:

$P_0 = P_{00} + P_{10} + P_{20}$, where these probabilities are defined in Theorem 3.3, equations (3.17), (3.18), and (3.19).

Case 3. M/M/1 subject to breakdowns, one spare:

$P_0 = P_{00} + P_{10} + P_{20}$, where these probabilities are defined in Theorem 3.4, equations (3.43), (3.44), and (3.45).

Case 4. M/M/C with no breakdowns:

$$P_0 = \left[\sum_{n=0}^{C-1} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n + \frac{1}{C!} \left(\frac{\lambda}{\mu}\right)^C \left(\frac{C\mu}{C\mu - \lambda}\right) \right]^{-1}. \quad (3.144)$$

This analytic result was suggested by Halachmi and Franta [10] as an approximation for GI/M/C queues when p_0 is very small. An adjustment to μ , to reflect slower service due to failures, will be provided later.

Case 5. M/G/ ∞ :

$$p_0 = e^{-\lambda E[S]} \quad (3.145)$$

where $E[S]$ is the expected service time. This equation is a result of Palm's Theorem [8] and is useful when the number of servers is greater than two. Here we assume that when a channel fails, the unit being repaired does not change channels. Evaluation of $E[S]$ requires some analysis which will be provided later.

Case 6. M/G/1:

$$p_0 = 1 - \lambda E[S], \quad (3.146)$$

where $E[S]$ is the expected service time. This is the analytical solution for the M/G/1 queue, proven in reference 6. No analytic results exist for the M/G/C queue. For a multi-channel queue, we will assume that all units are served by a single server which works as a "super server" at a rate $\bar{C}\mu$, where \bar{C} is the expected number of operational channels.

Cases 1, 2, and 3 are analytical results and can be directly applied to solve for \tilde{p}_0 . Cases 1, 2, 4, 5, and 6 are analytic results for specific systems which can be adapted to provide reasonable approximations. Finding a way to adjust the service rate (μ) is the key to altering the equations (3.143) through (3.146) to suit our needs. This will be done by assuming that all channels are occupied and finding the expected service rate when servers are subject to breakdown.

The total time a typical unit occupies a channel can be expressed as

$$T = S + D_1 + D_2 + \dots + D_{N(S)}, \quad (3.147)$$

where S is the service time, D_i is the delay caused by a server failure and $N(S)$ is the number of channel failures encountered in a time S .

Given all channels are occupied and no other channels fail while the inoperative channel is being replaced, then we can find the conditional expectation

$$E[D_i | K = k] = \begin{cases} \sum_{i=k}^{C-1} \frac{1}{(C+L-i)\eta} & , \text{ when } k \text{ other channels were operating just prior to the failure,} \\ & k=0, 1, \dots, C-1 \\ 0 & , \text{ otherwise.} \end{cases} \quad (3.148)$$

Then using the results of Corollary 3.1 and assuming all delays, D_i , are independent

$$E[D] = \sum_{k=0}^{C-1} q_{k+1}' E[D | K = k], \quad (3.149)$$

where q_k' is the conditional probability that k servers are operating just prior to a failure.

We now need an expression for $N(S)$, the expected number of failures in a time S , which has an exponential distribution with rate μ . Since failures occur as a Poisson process having rate ξ ,

$$\Pr \{N(S) = k\} = \int_0^{\infty} \mu e^{-\mu t} \frac{e^{-\xi t} (\xi t)^k}{k!} dt. \quad (3.150)$$

Then

$$E[N(S)] = \sum_{k=0}^{\infty} k \Pr\{N(S) = k\}$$

$$\begin{aligned}
&= \int_0^{\infty} \mu e^{-\mu t} \left[\sum_{k=0}^{\infty} k \frac{e^{-\xi t} (\xi t)^k}{k!} \right] dt \\
&= \xi \int_0^{\infty} t \mu e^{-\mu t} dt = \xi/\mu.
\end{aligned} \tag{3.151}$$

Reference 28 provides the result

$$E[D_1 + D_2 + \dots + D_{N(S)}] = E[N(S)] \cdot E[D]. \tag{3.152}$$

So, assuming independence for all random variables,

$$\begin{aligned}
E[T] &= E[S] + E[N(S)] \cdot E[D] \\
&= \frac{1}{\mu} + \frac{\xi}{\mu} E[D] = \frac{1}{\mu} \left[1 + \xi E[D] \right].
\end{aligned} \tag{3.153}$$

Therefore we can define the adjusted service rate as

$$\mu' = \frac{\mu}{1 + \xi E[D]}. \tag{3.154}$$

We will apply μ' in cases 1 and 2 to approximate p_0 for systems with one or two channels and spare servers. The modified service rate will be employed in Case 4 for all systems with over two channels. For Case 5 we can use $1/\mu'$ as an estimate for $E[S]$. Case 5 will be used for systems with over two channels and Case 6 will be used for all systems with spare servers.

Some empirical results are compared to values for p_0 estimated by simulations in Table 1. In all cases, p_0 was overestimated. Thus a reasonable approximation would be to select the smallest value.

3.3.3.5 Comparative Analysis

The diffusion approximation derived in the preceding sections was tested against analytic and simulation results. The actual stationary

C	L	ρ^*	SIMULATION	1	2	Case 4	5	6
1	1	.75	.231**	.244				.250
1	1	.95	.023**	.043				.050
2	2	.75	.159		.181	.204		.250
2	2	.95	.026		.056	.083		.050
3	0	.75	.108			.183	.150	.250
3	0	.95	.034			.126	.098	.050
3	3	.75	.083			.153	.121	.250
3	3	.95	.023			.076	.069	.050
5	0	.75	.051			.126	.064	.250
5	0	.95	.008			.095	.031	.005
5	3	.75	.030			.098	.034	.250
5	3	.95	.005			.062	.014	.050

$$*\rho = \frac{\lambda}{c\mu}$$

**Analytic Result

Comparison of Estimates for p_0

TABLE 1

probability distributions were derived from the results of Theorems 3.2, 3.3 and 3.4. The simulated distributions were developed using a simulation routine involving at least 20,000 state changes. The results are graphically displayed in the figures that follow. Each figure is labeled using the following nomenclature:

C = the number of channels,

L = the number of spare servers provided,

ρ = the traffic intensity = $\lambda/\bar{C}\mu$.

Each graph includes information on the ratios, λ/ξ and μ/η .

We know from Section 3.3.3.2 that the accuracy of the approximation $\Pr \{k | n\} \approx \Pr \{k\}$ improves as ξ and η get large; therefore, we should expect the size of ξ and η to affect our diffusion approximation. As the relative size of the rate of repair for servers (η) decreases, the ratio μ/η gets large. Examining the data, we see, the approximation suffers as the ratio μ/η increases above the proposed bound of one. On the other hand, the approximation is relatively insensitive to the ratio λ/ξ . Thus, this diffusion approximation is most sensitive to changes in the server repair rate (η) and improves as the rate increases relative to all other parameters.

Some simulation results show a slight perturbation due to auto correlation effects. Regardless, it appears the exponential tail of the diffusion approximation matches the true shape of the distribution in all cases. When the approximation is inaccurate, the tail of the approximate distribution is slightly low.

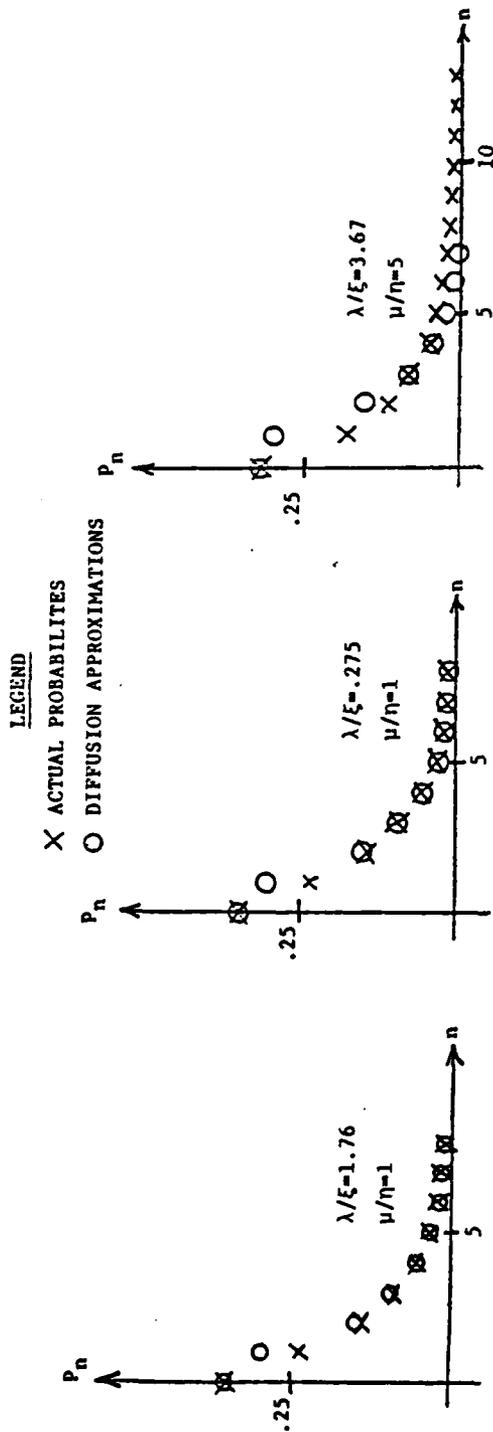


Figure 3.7: $C=1, L=0, \rho=0.55$

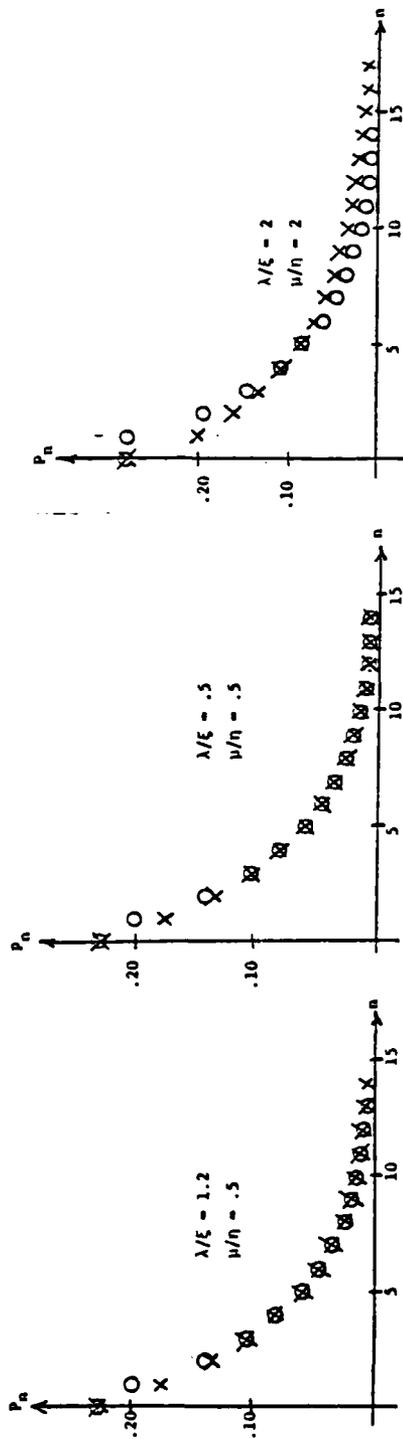
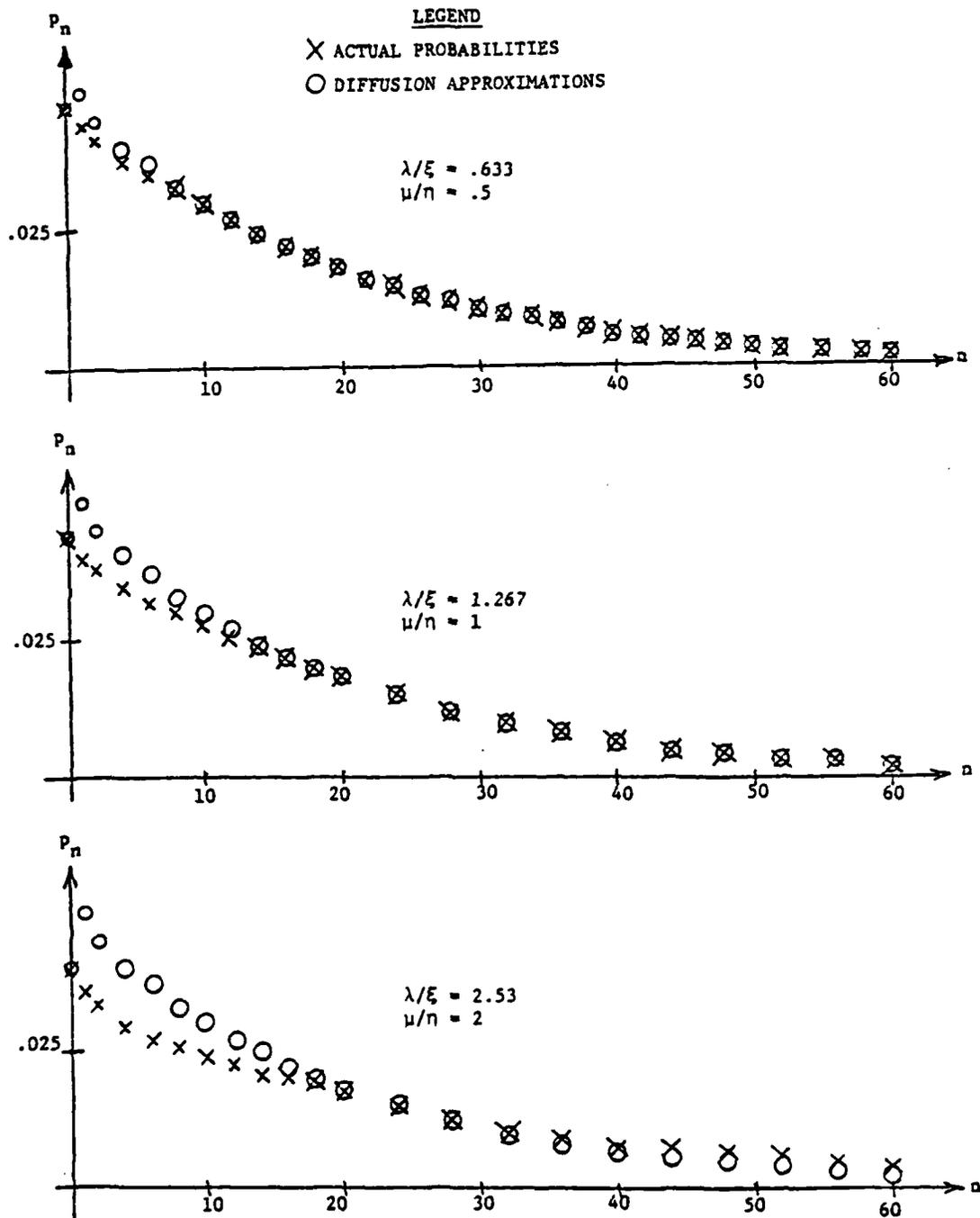


Figure 3.8: $C=1, L=0, \rho=0.75$

Figure 3.9: $C=1, L=0, \rho=.95$

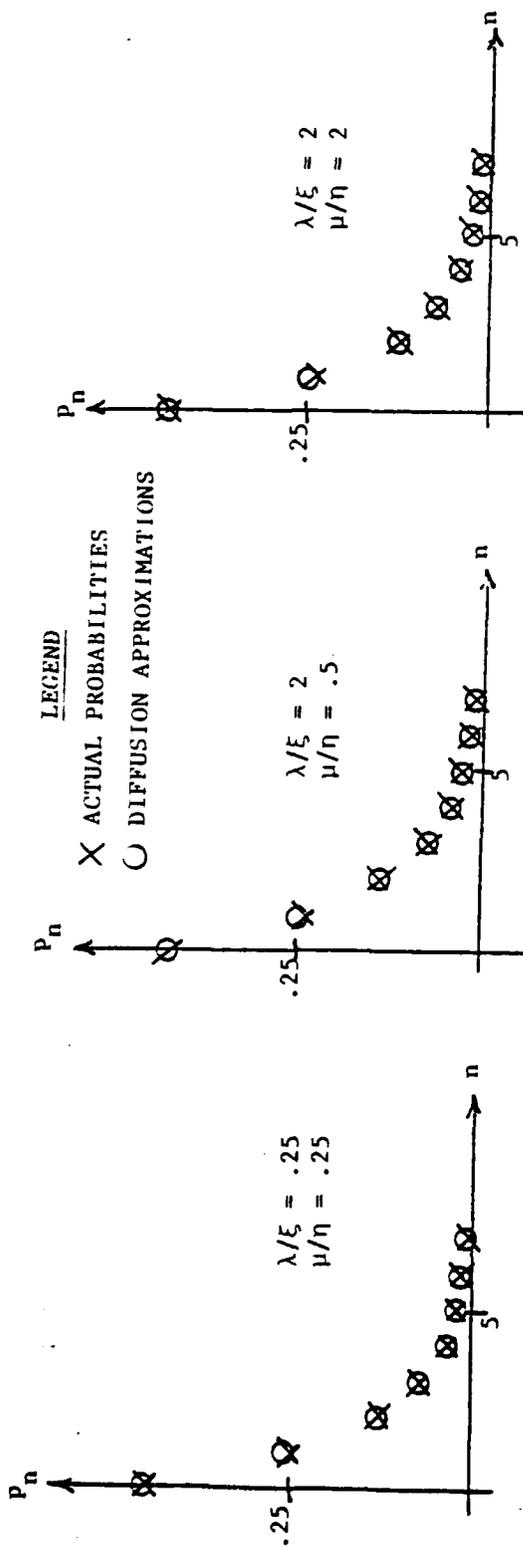


Figure 3.10: $C=1$, $L=1$, $\rho=.55$

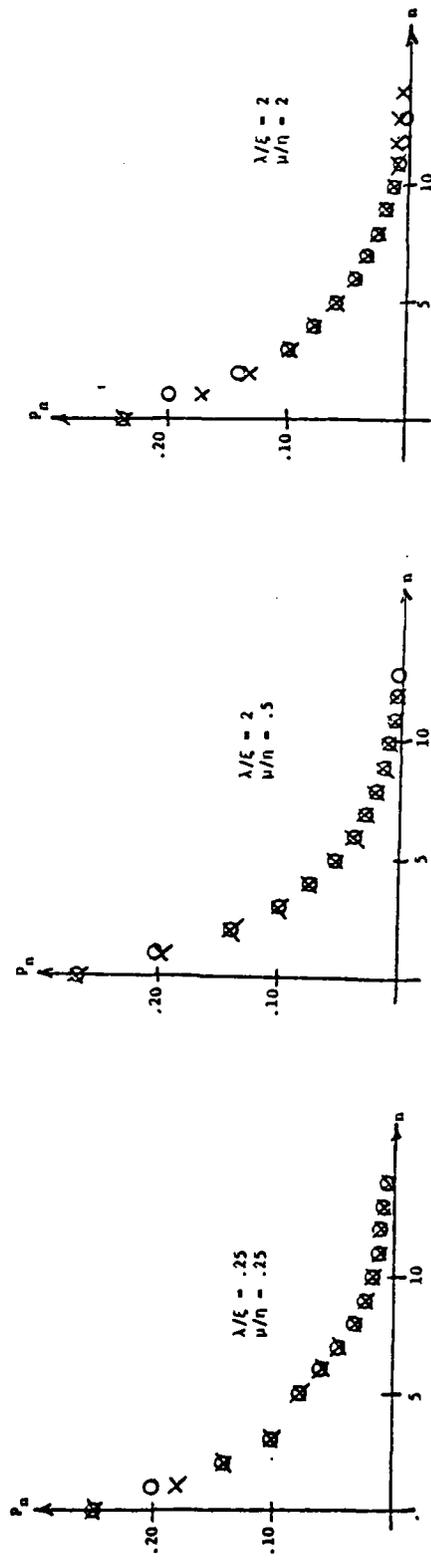


Figure 3.11: $C=1$, $L=1$, $\rho=.75$

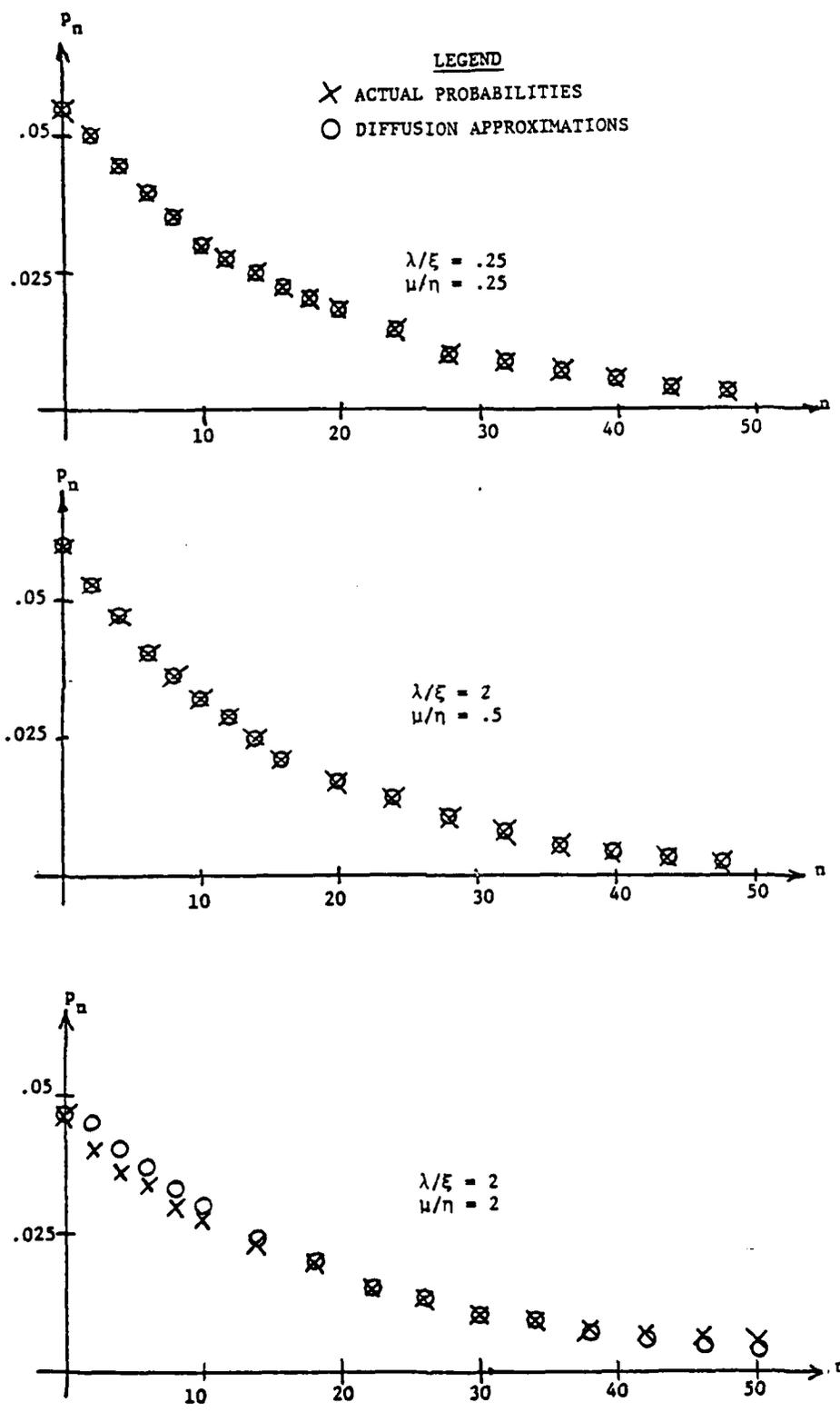


Figure 3.12: C=1, L=1, $\rho=.95$

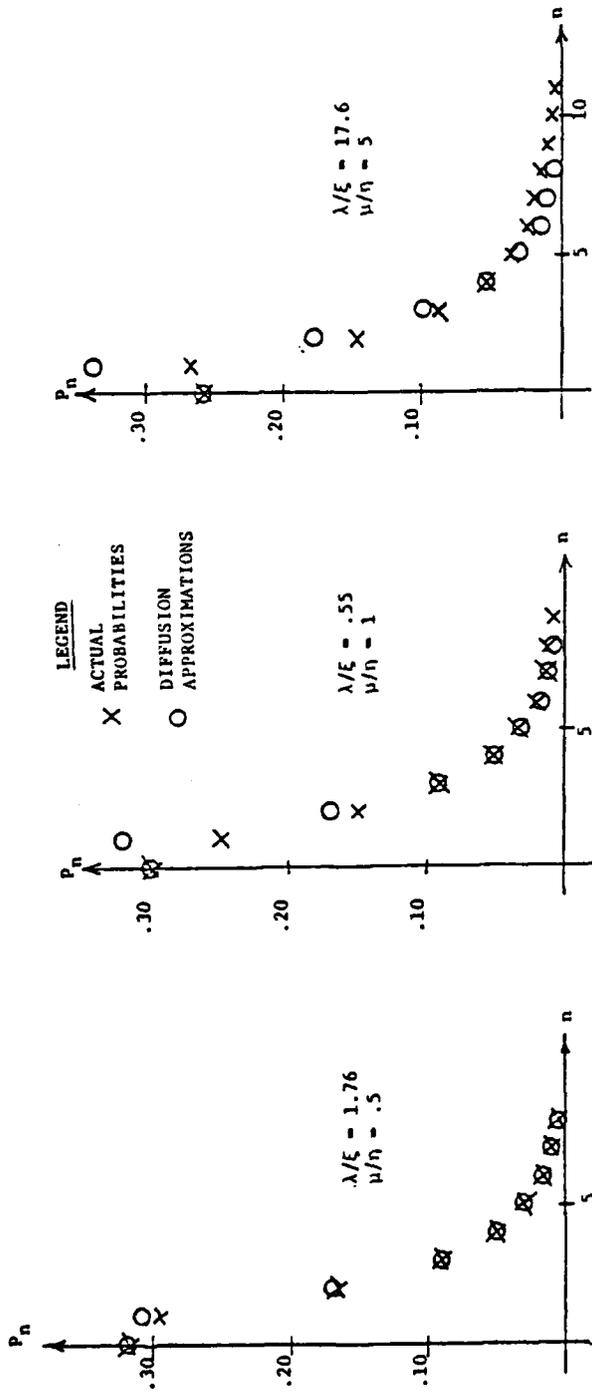


Figure 3.13: C=2, L=0, $\rho=.55$

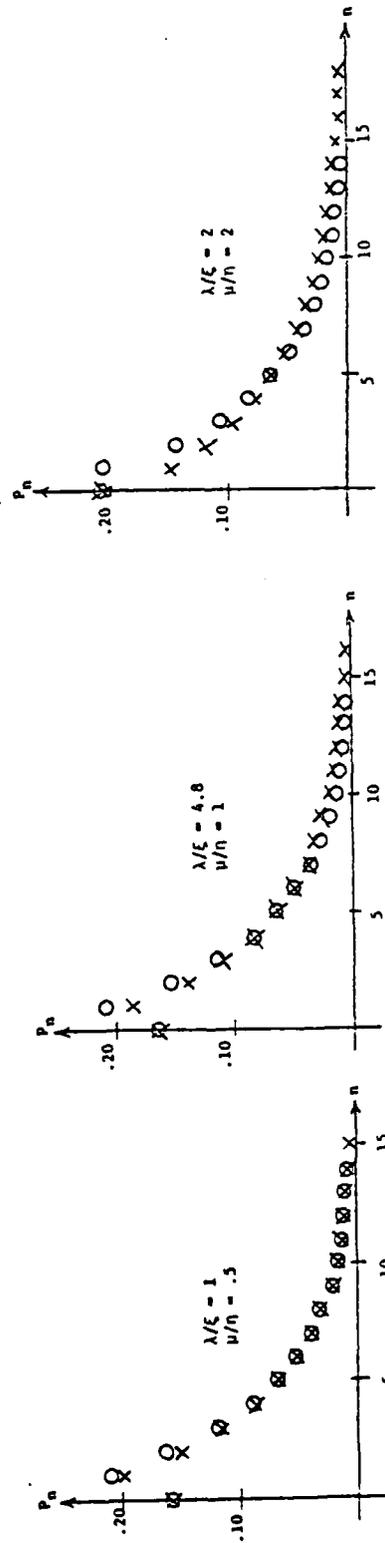


Figure 3.14: C=2, L=0, $\rho=.75$

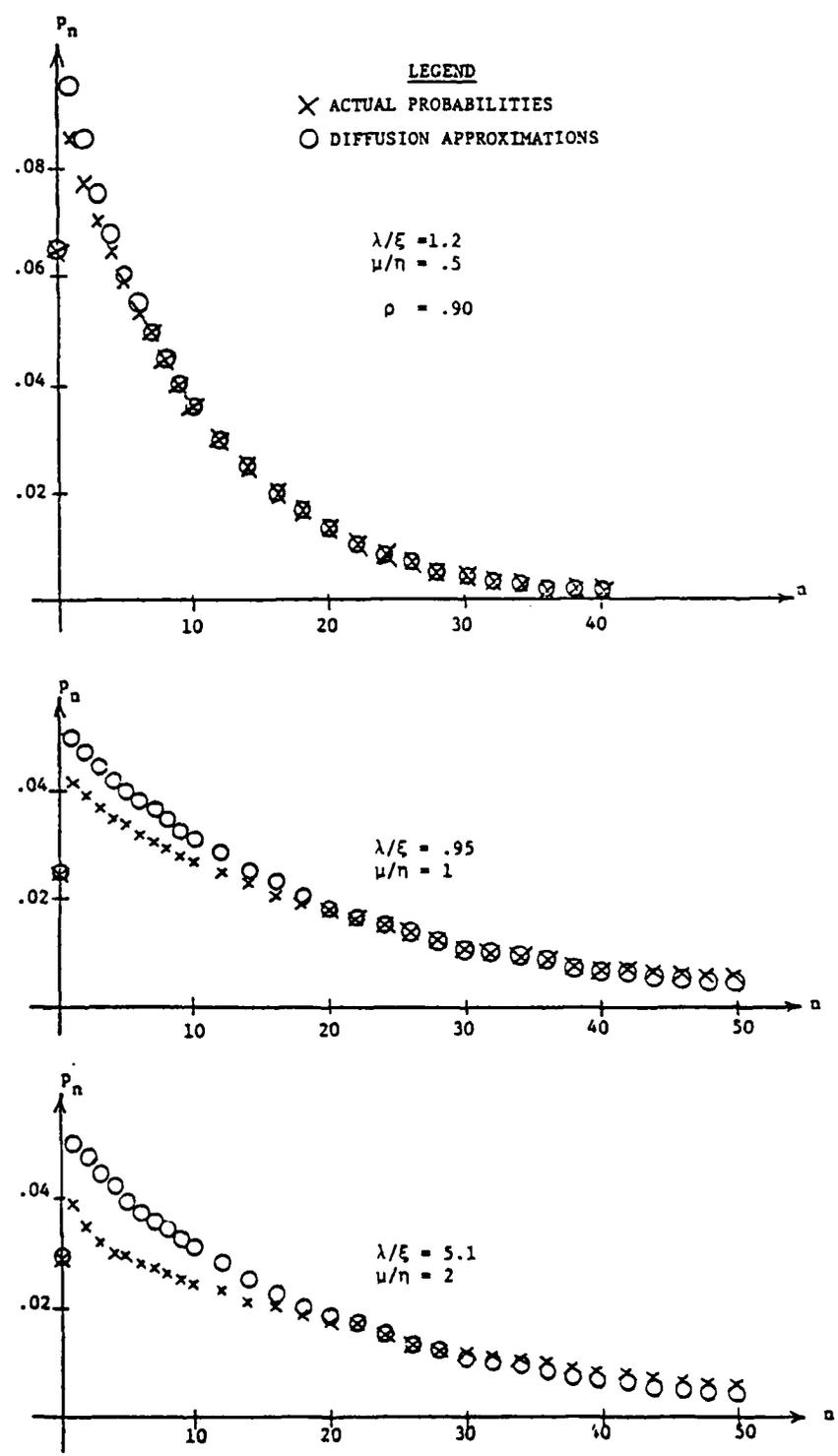


Figure 3.15: $C=2, L=0, \rho=.95$

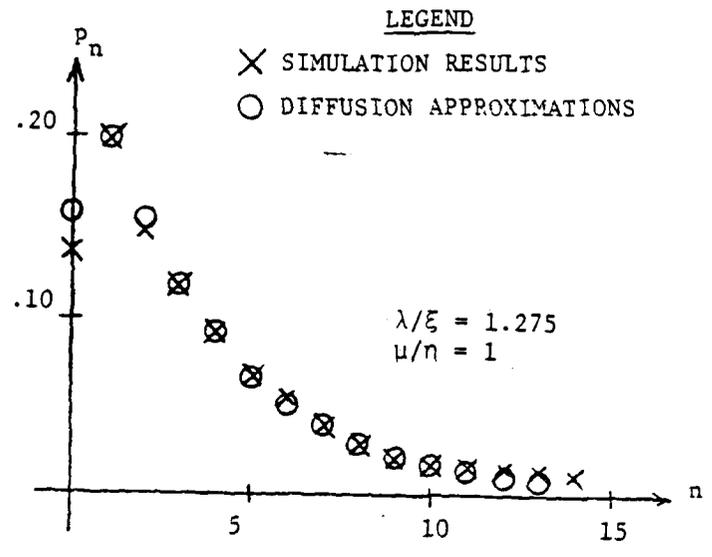


Figure 3.16: C=2, L=2, $\rho=0.75$

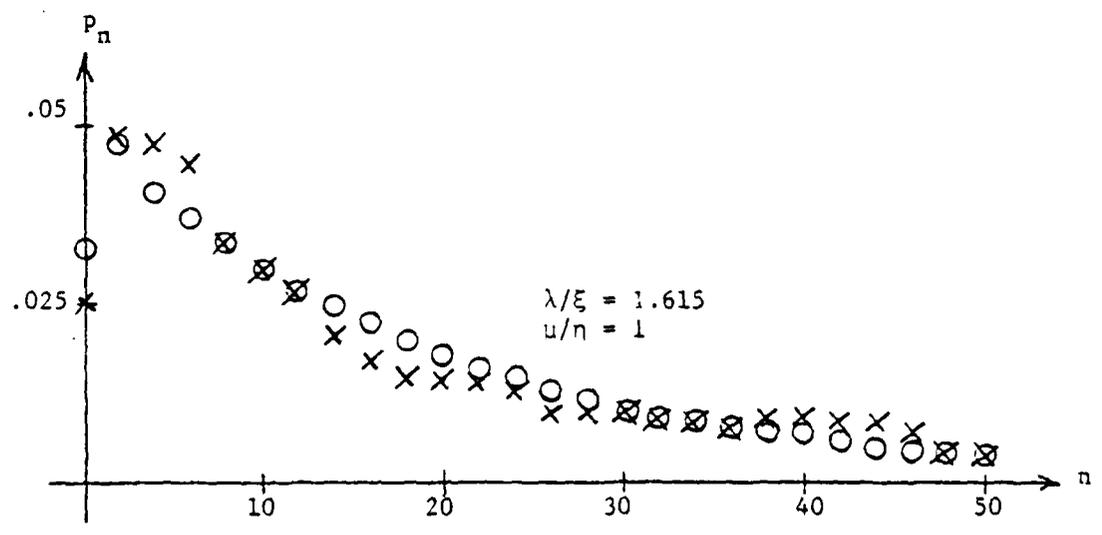
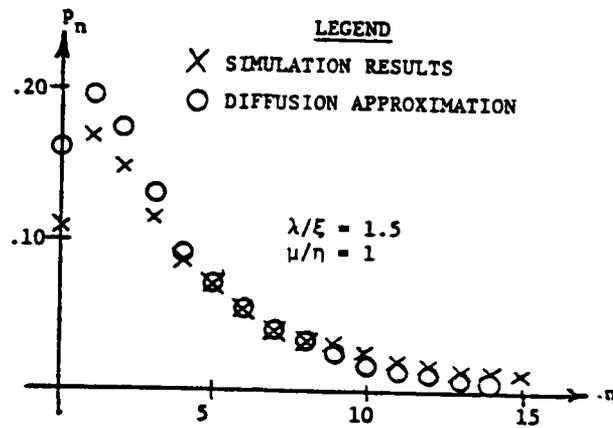
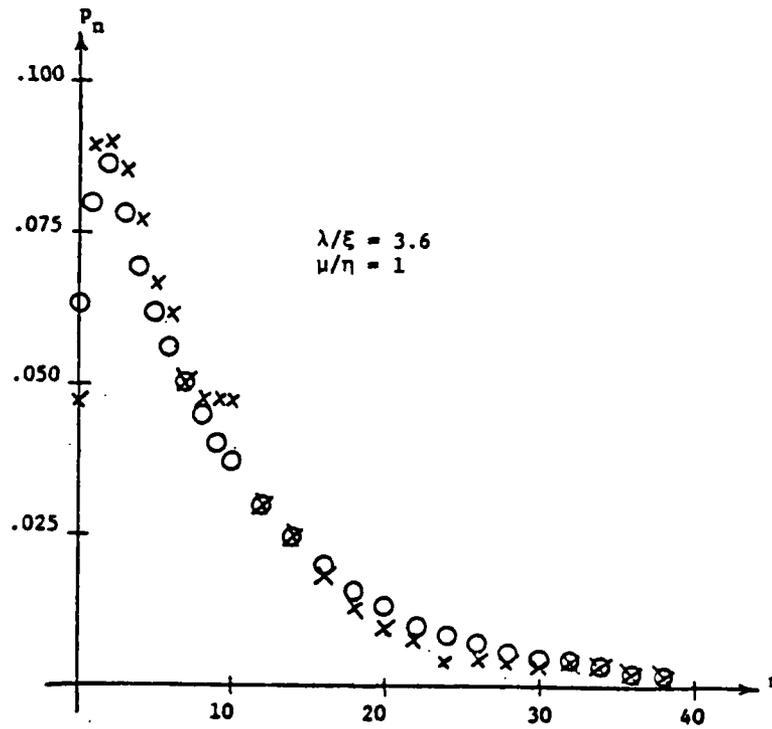
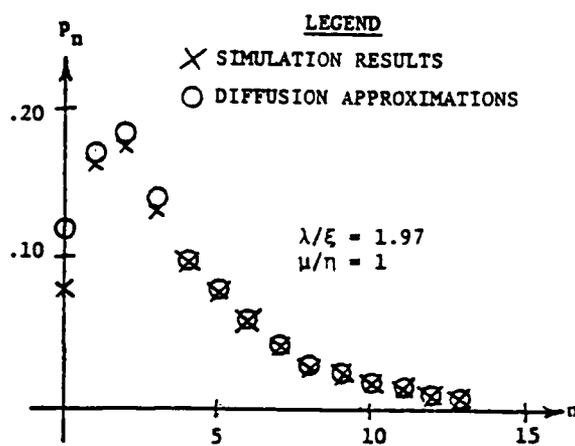
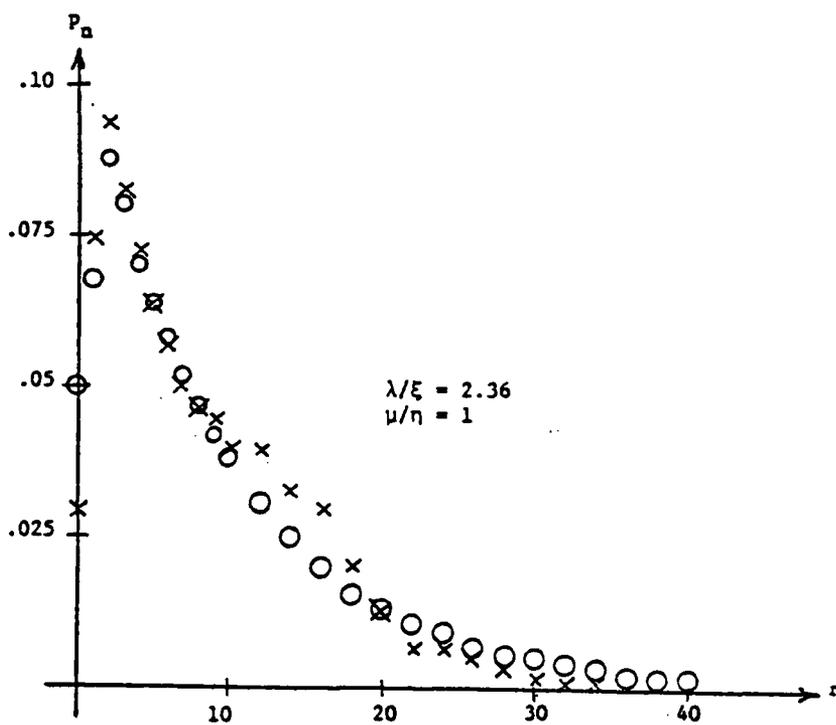
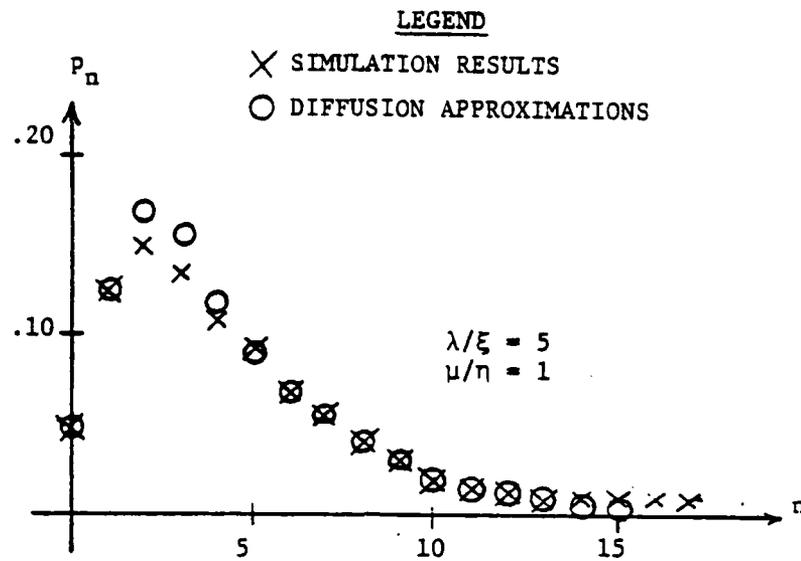
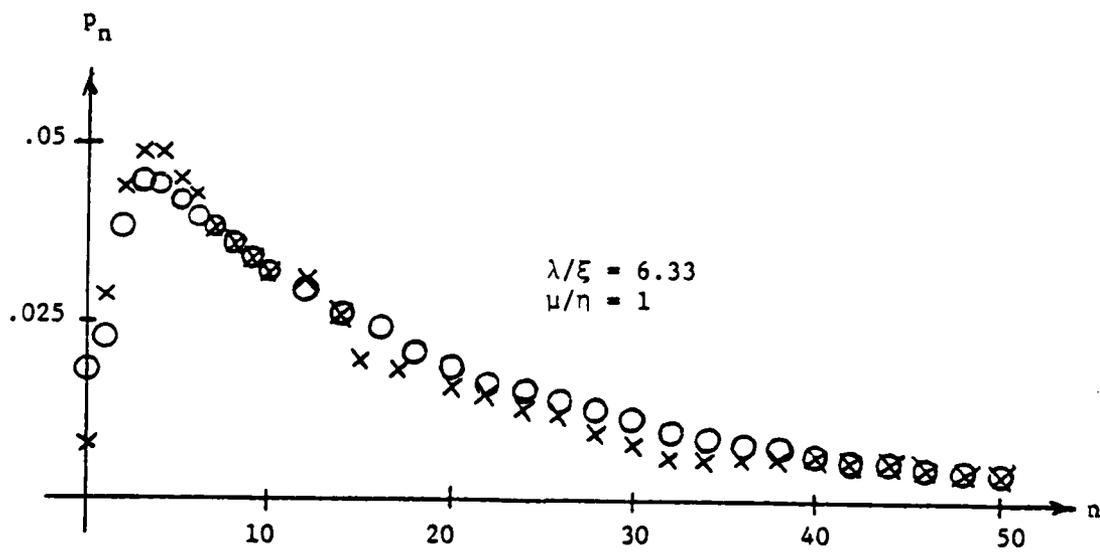


Figure 3.17: C=2, L=2, $\rho=0.95$

Figure 3.18: $C=3, L=0, \rho=0.75$ Figure 3.19: $C=3, L=0, \rho=0.90$

Figure 3.20: $C=3, L=3, \rho=0.75$ Figure 3.21: $C=3, L=3, \rho=0.90$

Figure 3.22: $C=5, L=0, \rho=0.75$ Figure 3.23: $C=5, L=0, \rho=0.95$

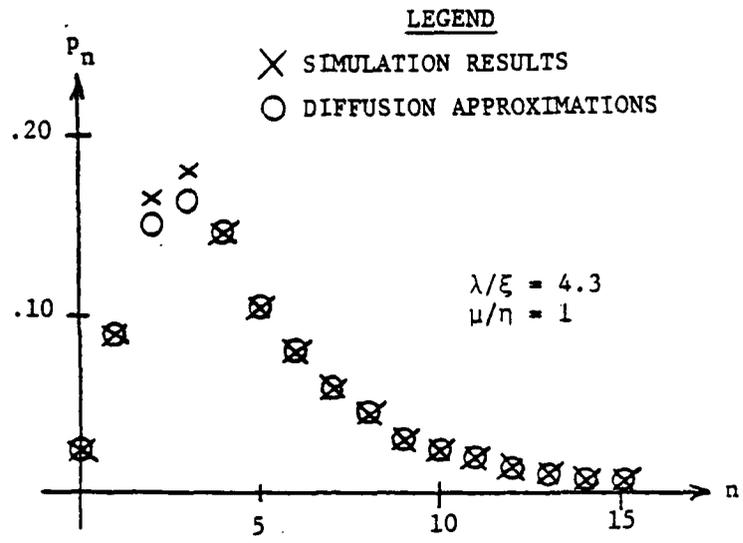


Figure 3.24: C=5, L=3, $\rho=0.75$

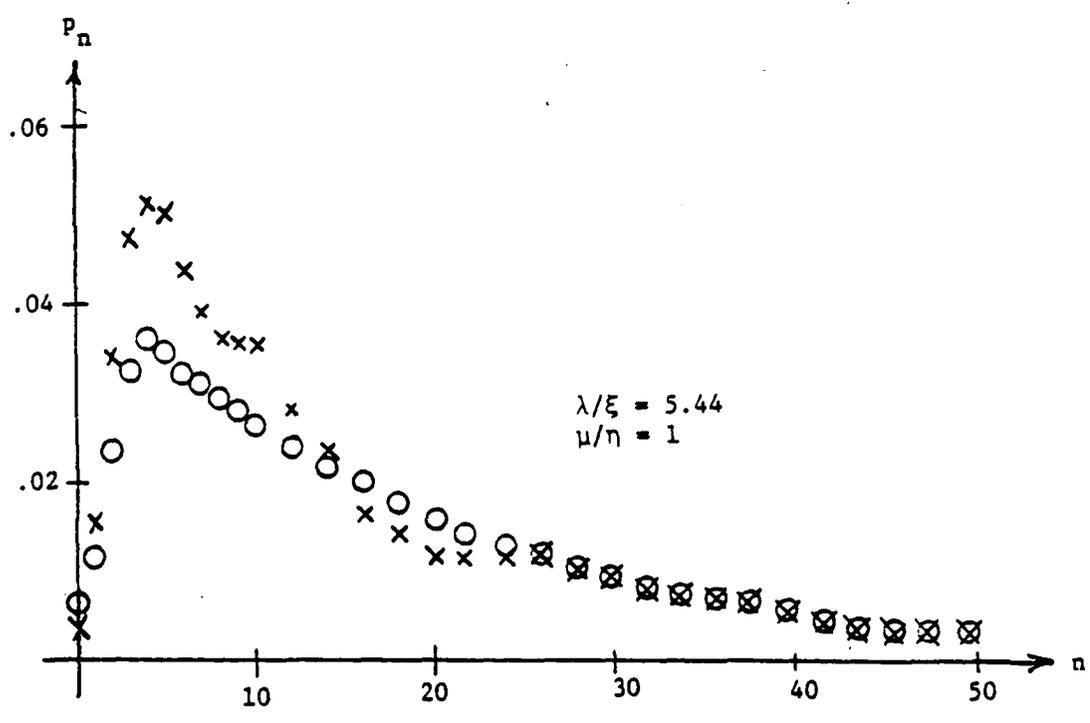


Figure 3.25: C=5, L=3, $\rho=0.95$

CHAPTER IV
AN OPTIMIZATION METHOD

This chapter will derive a method to find the optimal allocation of resources for the inventory system we have modeled. We will first examine the backorder function given the diffusion approximation developed in the previous chapter for the stationary distribution. Since the distribution is for the total number of units in the system, it will be modified to describe a multi-item system. Then using the exponential tail of this distribution, we develop sufficient conditions for convexity of the backorder function. Finally, a simple marginal analysis technique is given to find the solution to the optimization problem, P.

The optimization problem, P, we wish to solve is:

$$\text{minimize} \quad \sum_{i=1}^m E_i \sum_{n>s_i} (n - s_i) p(n|\lambda_i, \mu, \xi, \eta, C, L), \quad (4.1)$$

subject to

$$C \cdot C_C + L \cdot C_s + \sum_{i=1}^m C_i s_i \leq B, \quad (4.2)$$

and

$$\lambda < \mu \bar{C}. \quad (4.3)$$

Here C , L , and s_i , $i = 1, \dots, m$, are the decision variables. Because of the high setup costs for service channels, it is reasonable to assume C is small. We shall show later that, for fixed values of C , P is relatively easy to solve; thus, we shall enumerate solutions for small values of C . For notational simplicity in the following developments, the variable C and parameters μ , ξ , and η will be suppressed.

4.1 Approximating The Backorder Function

After examining the objective function, we can see that P is not separable. In fact, it is difficult to explicitly express $p(n | \lambda_i, L)$, the distribution of the number of units of type i in the system, as a function of L . An expression for the expected backorders of item i , $B_i(s_i, L)$, will be created that will assist in the analysis.

Define

$$B_i(s_i, L) = \sum_{n > s_i} (n - s_i) p(n | \lambda_i, L). \quad (4.4)$$

We now develop an expression for an approximation to $p(n | \lambda_i, L)$.

Since the influx of each type unit is an independent Poisson process, the entire input process is Poisson. The conditional distribution for the number of units of type i in the system is binomial, i.e.,

Pr { k units of type i in system | n total units in system}

$$= \begin{cases} 0 & , n < k \\ \binom{n}{k} \left(\frac{\lambda_i}{\lambda}\right)^k \left(\frac{\lambda - \lambda_i}{\lambda}\right)^{n-k} & , n \geq k. \end{cases} \quad (4.5)$$

Using the law of total probability gives

$$p(k | \lambda_i, L) = \sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{\lambda_i}{\lambda}\right)^k \left(\frac{\lambda - \lambda_i}{\lambda}\right)^{n-k} p_n, \quad (4.6)$$

where p_n is the stationary probability of n total units in the system.

For $n > c$, the diffusion approximation, \tilde{p}_n , for p_n given in the previous chapter is

$$\tilde{p}_n = \int_{n-.5}^{n+.5} f(x) dx = H \int_{n-.5}^{n+.5} \exp(Kx) dx$$

$$\begin{aligned}
&= \frac{H}{K} \{ \exp [K(n + .5)] - \exp [K(n - .5)] \} \\
&= \frac{H}{K} [\exp (.5K) - \exp (-.5K)] \exp (Kn) \\
&= H' r^n, \quad n > C, \tag{4.7}
\end{aligned}$$

where

$$K = \frac{2(\lambda - \mu \bar{C})}{\lambda + \mu \bar{C}}, \tag{4.8}$$

$$H' = \frac{H}{K} [\exp (.5K) - \exp (-.5K)], \tag{4.9}$$

$$\text{and } r = \exp (K). \tag{4.10}$$

Recall that the average number of operating channels, \bar{C} , is a function of the number of spares provided, L , so that $K = K(L)$ and $r = r(L)$.

Unfortunately, the expression for the few values of \tilde{p}_n where $n \leq C$ is not as concise.

Applying equation (4.7) to equation (4.5) for $k > C$, we get

$$\begin{aligned}
p(k|\lambda_i, L) &\approx \sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{\lambda_i}{\lambda} \right)^k \left(\frac{\lambda - \lambda_i}{\lambda} \right)^{n-k} \tilde{p}_n \\
&= \sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{\lambda_i}{\lambda} \right)^k \left(\frac{\lambda - \lambda_i}{\lambda} \right)^{n-k} H' r^n \\
&= H' \left(\frac{\lambda_i r}{\lambda} \right)^k \sum_{n=k}^{\infty} \binom{n}{k} \left[\frac{(\lambda - \lambda_i) r}{\lambda} \right]^{n-k} \tag{4.11}
\end{aligned}$$

Under the assumption that $\lambda < \mu \bar{C}$, we have $K < 0$ and thus $r < 1$. Then the factor being raised to a power in the summand is less than unity and we can apply the binomial expansion to get

$$p(k|\lambda_i, L) \approx \frac{H' \left(\frac{\lambda_i r}{\lambda} \right)^k}{\left[1 - \frac{(\lambda - \lambda_i) r}{\lambda} \right]^{k+1}}$$

$$\begin{aligned}
&= \frac{H \lambda}{\lambda_i r} \left[\frac{\lambda_i r}{\lambda(1-r) + \lambda_i r} \right]^{k+1} \\
&= A_i b_i^{k-1}, \tag{4.12}
\end{aligned}$$

$$\text{where } b_i = \frac{\lambda_i r}{\lambda(1-r) + \lambda_i r}, \tag{4.13}$$

$$\text{and } A_i = \frac{H \lambda b_i^2}{\lambda_i r}. \tag{4.14}$$

Equation (4.12) shows the similarity to the geometric distribution.

Generalizing for all $k \geq 0$ gives

$$\begin{aligned}
p(k|\lambda_i, L) &\approx \sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{\lambda_i}{\lambda}\right)^k \left(\frac{\lambda - \lambda_i}{\lambda}\right)^{n-k} \tilde{p}_n \\
&= \sum_{n=\max(C+1, k)}^{\infty} \binom{n}{k} \left(\frac{\lambda_i}{\lambda}\right)^k \left(\frac{\lambda - \lambda_i}{\lambda}\right)^{n-k} H r^n \\
&\quad + \sum_{n=k}^C \binom{n}{k} \left(\frac{\lambda_i}{\lambda}\right)^k \left(\frac{\lambda - \lambda_i}{\lambda}\right)^{n-k} \tilde{p}_n \\
&= A_i b_i^{k-1} + d_{i,k}, \quad k \geq 0, \tag{4.15}
\end{aligned}$$

$$\text{where } d_{i,k} = \begin{cases} \sum_{n=k}^C \binom{n}{k} \left(\frac{\lambda_i}{\lambda}\right)^k \left(\frac{\lambda - \lambda_i}{\lambda}\right)^{n-k} \tilde{p}_n, & 0 \leq k \leq C \\ 0, & k > C, \end{cases} \tag{4.16}$$

and A_i and b_i are defined in (4.13) and (4.14).

We can now solve for a general approximation of $B_i(s_i, L)$, the expected backorders for item i . Since $b_i < 1$, we have

$$B_i(s_i, L) = \sum_{n=s_i}^{\infty} (n-s_i) p(n|\lambda_i, L)$$

$$\begin{aligned}
&= \sum_{n=s_1}^C (n-s_1) p(n|\lambda_1, L) + \sum_{n=C+1}^{\infty} (n-s_1) p(n|\lambda_1, L) \\
&\approx \sum_{n=s_1}^C (n-s_1) (A_1 b_1^{n-1} + d_{1,n}) + \sum_{n=C+1}^{\infty} (n-s_1) A_1 b_1^{n-1} \\
&= \sum_{n=s_1}^{\infty} (n-s_1) A_1 b_1^{n-1} + \sum_{n=s_1}^C (n-s_1) d_{1,n} \\
&= \frac{A_1 b_1^{s_1}}{(1-b_1)^2} + \sum_{n=s_1}^C (n-s_1) d_{1,n}.
\end{aligned} \tag{4.17}$$

We expect $B_1(s_1, L)$ to be a monotonically decreasing function of s_1 . In fact,

$$\begin{aligned}
\Delta_s B_1(s, L) &= B_1(s+1, L) - B_1(s, L) \\
&\approx \frac{A_1}{(1-b_1)^2} \left[b_1^{s+1} - b_1^s \right] + \sum_{n=s+1}^C n d_{1,n} - \sum_{n=s}^C n d_{1,n} \\
&= \frac{A_1 b_1^s}{(1-b_1)^2} (b_1 - 1) - s d_{1,s} < 0
\end{aligned}$$

since $b_1 < 1$ and $d_{1,s} \geq 0$. It is not clear that $d_{1,s}$ is decreasing in s for $0 \leq s \leq C$; however, since $d_{1,s} = 0$ for $s > C$, the function $\Delta_s B(s, L)$ is increasing in s for all $s \geq C$. Therefore, $B_1(s_1, L)$ is convex in s_1 alone, for $s_1 \geq C$.

The parameter b_1 , as a function of L , has some properties which will be useful later. These are easiest to display after first ex-

ploring \bar{C} , K and r as functions of L . The expected number of operative channels, \bar{C} , is a bounded, strictly increasing function of L . Furthermore, $\lim_{L \rightarrow \infty} \bar{C}(L) = C$. For feasibility we assume there is an L such that $\lambda < \mu \bar{C}(L)$. As a result, K is a bounded, monotonically decreasing function of L . Specifically,

$$K_{\infty} = \lim_{L \rightarrow \infty} K(L) = \lim_{L \rightarrow \infty} \frac{2[\lambda - \mu \bar{C}(L)]}{\lambda + \mu \bar{C}(L)} = \frac{2(\lambda - \mu C)}{\lambda + \mu C} < 0,$$

since $\lambda < \mu C$. To show monotonicity, we note that

$$\begin{aligned} \Delta K(L) &= K(L+1) - K(L) \\ &= \frac{2[\lambda - \mu \bar{C}(L+1)]}{\lambda + \mu \bar{C}(L+1)} - \frac{2[\lambda - \mu \bar{C}(L)]}{\lambda + \mu \bar{C}(L)} \\ &= \frac{-4\lambda\mu [\bar{C}(L+1) - \bar{C}(L)]}{[\lambda + \mu \bar{C}(L+1)][\lambda + \mu \bar{C}(L)]} < 0, \end{aligned}$$

since $\bar{C}(L+1) > \bar{C}(L)$. Since $\bar{C}(L)$ increases towards its upper bound C , we expect $\Delta \bar{C}(L)$ to be a decreasing function of L . Then $\Delta K(L)$ would be increasing (becoming less negative), and therefore $K(L)$ is a convex function of L .

The parameter $r = \exp(K)$ is a convex function of a convex variable; thus, r is convex in L . It is bounded below by $r_{\infty} = \exp(K_{\infty})$ which is in the interval $(0,1)$. If we assume, for $b_i = \frac{\lambda_i r}{\lambda(1-r) + \lambda_i r}$, that L is a continuous variable we have

$$\begin{aligned} \frac{dB_i}{dL} &= \frac{\lambda_i r' [\lambda(1-r) + \lambda_i r] - \lambda_i r (-\lambda r' + \lambda_i r')}{[\lambda(1-r) + \lambda_i r]^2} \\ &= \frac{\lambda_i \lambda r'}{[\lambda(1-r) + \lambda_i r]^2} < 0, \end{aligned}$$

since $r' = \frac{dr}{dL} < 0$. Thus b_i is strictly decreasing in L . Additionally,

$$b_{i\infty} = \lim_{L \rightarrow \infty} b_i = \frac{\lambda_i r_{\infty}}{\lambda(1-r_{\infty}) + \lambda_i r_{\infty}}$$

lies in the interval $(0,1)$ and is the lower bound for b_i .

The above analysis is displayed graphically in Figures 4.1 and 4.2. Figure 4.1 displays parameters for a system which is saturated when no spares are provided. In Figure 4.2, the system is never saturated. The variable L is considered a continuous variable for purposes of illustration.

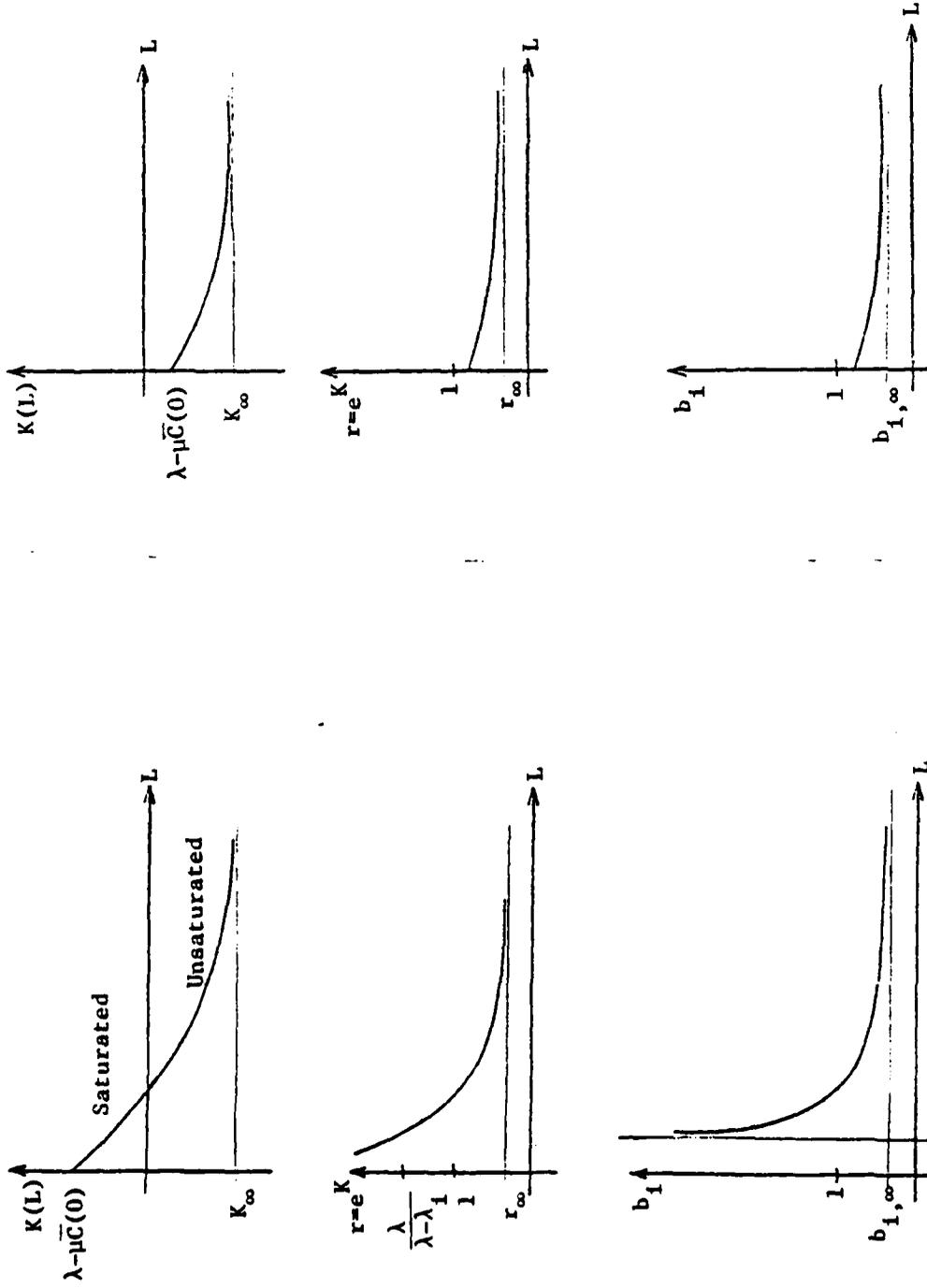


Figure 4.1: Parameters K , r , and b_1
 $(\lambda > \mu \bar{C}(0))$

Figure 4.2: Parameters k , r , and b_1
 $(\lambda < \mu \bar{C}(0))$

4.2 An Algorithm For Determining Unit Stocklevels And Service System Design

The steps for a generalized algorithm will be presented with an explanation and justification of each given afterwards. The algorithm is given in the context of solving the problem graphically:

0. Set $C_{\min} = \left[\frac{\lambda}{\mu} \right]$, where $[\cdot]$ is the greatest integer function.

1. Determine lower and upper bounds for L , L_{\min} and L_{\max} , by

$$L_{\min} = \inf \{L \geq 0 \mid \frac{\lambda}{\mu} \leq \bar{C}(L)\} \text{ and } L_{\max} = \inf \{L \geq L_{\min} \mid \bar{C}(L) \geq .95C\}$$

2. Solve for the optimal allocation of spares, $S^* = (s_1^*, \dots, s_m^*)$, for a given budget in item spares and a mid-range L , using marginal analysis.

3. Compute $B(S^*, L)$ for each value of L with S^* found in step 2.

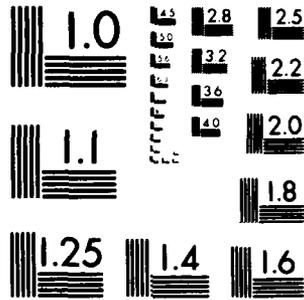
4. Select a new investment level for item spares and go to step 2.

5. Increment C by one and go to step 1.

6. Plot fixed total investment curves and select the optimal solution.

The selection of C_{\min} and L_{\min} in the first two steps provide the lowest feasible investment in service facilities. In a very congested system, it is possible that $L_{\min} = L_{\max}$ since L_{\min} would have to be large enough to provide a very high level of service reliability.

Steps 2 and 3 fix the investment in unit spares and compute the expected backorders over a range of server spares. Since $B_i(s_i, L)$, $i = 1, \dots, m$, are convex for $s_i \geq C$, then the total backorder function, $B(S, L) = \sum_{i=1}^n B_i(s_i, L)$, is convex for $s_i \geq C$. In practice, inventory systems are often budgeted so that all items are stocked with at least



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A.

the mean number of units expected in the system (commonly called the "expected pipeline inventory"). Generally, the mean of the distribution $p(n | \lambda_i, L)$ is greater than C ; so, we should be outside the region $s_i \leq C$, $i = 1, \dots, m$, provided the investment budget is sufficiently large. Under these conditions, the optimal values s_i^* , $i = 1, \dots, m$, can be found by marginal analysis due to the decreasing incremental returns to scale. We shall start with all s_i set to some minimum value and then successively increment the level of item i^* where

$$\frac{\Delta B_{i^*}(s_{i^*}^*, L)}{C_{i^*}} = \max_{1 \leq i \leq n} \left\{ \frac{\Delta B_i(s_i, L)}{C_i} \right\}.$$

That is, we increase the spare level of the item which provides the greatest improvement in the objective per dollar invested. This allocation can be followed until the entire budget is used since the objective function $B(S, L)$ is strictly decreasing in s_i , $i = 1, \dots, m$. Notice this procedure would not be altered if essentialities, E_i , were assigned to each item and our objective function becomes

$$B(S, L) = \sum_{i=1}^n E_i B_i(s_i, L).$$

We assume that for a given level of investment in unit spares, the optimal values s_i^* , $i = 1, \dots, m$, do not change for different investments in server spares. Figures 4.1 and especially 4.2 show that for small changes in L , b_i does not change significantly. Moreover, we can expect the same relative changes in b_i for all $i = 1, \dots, m$. We should, therefore, not expect great changes in the backorder functions for various values of L . The only area in which the backorder function changes significantly is in the range $0 \leq s_i \leq C$. This is most easily explained by the presence of the nuisance parameter $\sum_{n=s_i}^C (n - s_i) d_{i,n}$, in

equation (4.17) for the expected backorders when $s_1 \leq C$. As previously mentioned, we should expect a budget large enough to prevent this from becoming a factor.

An illustration of the performance curves obtained for a fixed value of C is given in Figure 4.3. The solid curves represent fixed investments in spare servers. As increased investments are made in spare items, the objective function decreases. The dotted line represents the trade-off curve for a fixed net investment in item spares and spare servers.

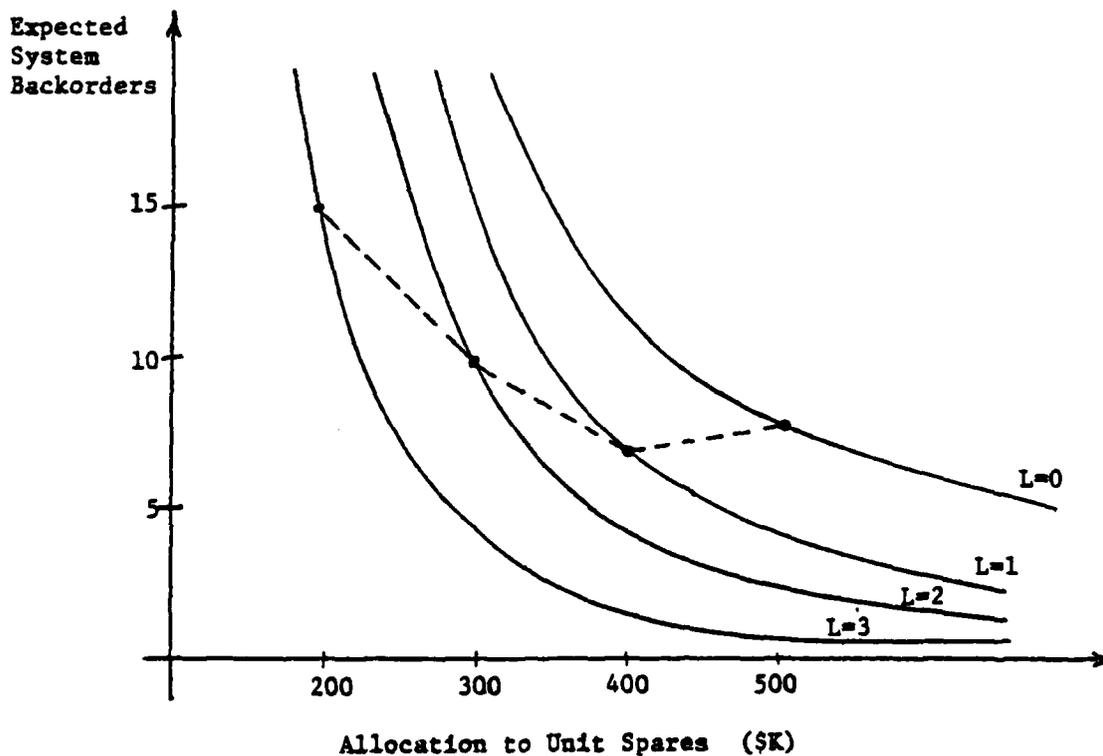


Figure 4.3: Performance and Tradeoff Curves (C Fixed)

Note that we must include the cost of establishing C service channels to this to get the gross investment. The lowest point on the trade off curve represents the optimal mix of investments for this budget. In this illustration, it is best to purchase one spare server (at \$100,000) and allocate \$400,000 to unit spares when the investment that can be made is \$500,000. With all the data displayed in this fashion, a manager can plot several trade-off curves and observe how expected backorders vary with changes in total investment.

CHAPTER V
NON-STATIONARY ANALYSIS

5.1 Introduction and Motivation

Inventory systems must often operate in dynamic environments which preclude the use of stationary planning models. For example, demand rates may cycle much like traffic during the day on a city street; surges in demand can occur when a military environment goes from peacetime to wartime; or, expected demands could steadily increase as a new aircraft system is being purchased. Sometimes inventory problems require a model for short horizon planning only, such as, equipping an aircraft carrier for a four month cruise. In all cases, the objective is to describe the system's ability to provide support through time. This chapter is devoted to modeling the changes through time of a recoverable item inventory system, with servers subject to failure, where the unit demands may be non-stationary.

In general, we are considering a system whose input and output processes are both non-stationary. As will be shown, it is extremely difficult to describe a multi-server system's transient behavior for the case when both processes are stationary. We can remove one element of non-stationarity by considering the service facility as operating in different "service states" through time. Each state would represent a period when the output process is stationary. The length of this period can be a random variable.

We will define the service state as the number of servers requiring repair. Thus, the state space is $\{0, 1, \dots, L, L + 1, \dots, L + C\}$. For each service state, we want the transient distribution of the number of

units in the system. For example, let G represent the state set when all service channels are operating, $G = \{0, 1, \dots, L\}$, and B represent the state when one service channel is inoperative, $B = \{L + 1\}$. The length of time the system is in each state will be represented by T_G and T_B .

Figure 5.1 shows how the system behaves through time.

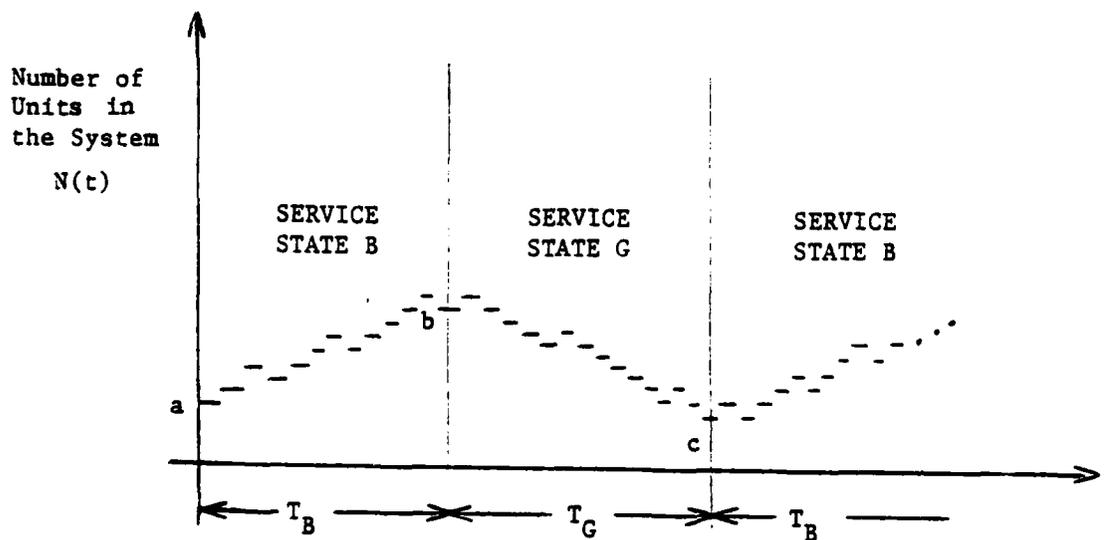


Figure 5.1: System Performance for Different Service States

Notice, while the system is in state B , the service facility is not as efficient and the number of units in the system drifts up. Let $N(t)$ be the number of units in the system at time t . Define the time dependent transition probability,

$$P_B(a, b, t) = \Pr\{N(0) = a, N(t) = b \mid \text{service state } B\}.$$

The time period we are concerned with, T_B , is a random variable. Then the transition probability for this period is

$$P_B(a, b) = \Pr\{N(0) = a, N(T_B) = b\}$$

$$P_B(a,b) = \int_0^{\infty} P_B(a,b,t) S_B(t) dt, \quad (5.1)$$

where $S_B(t)$ is the probability density function for the variable T_B .

Using the same notation, we can express the transition probability of going from b units at the beginning of service state G to c units at the end of this service state as

$$P_G(b,c) = \Pr\{N(T_B) = b, N(T_B + T_G) = c\}.$$

This analysis can be carried further to describe the transition from a to c through service states B and G :

$$P_{BG}(a,c) = \sum_{b=0}^{\infty} P_B(a,b) P_G(b,c). \quad (5.2)$$

It is apparent that in order to model the system in this manner, we must know two different probability functions: (1) the time dependent transition distribution for the number of units in the system given a particular service state, and (2) the density for the length of time each service state exists. The latter density is frequently referred to as the "passage time" density. We will explore each probability function in the context of the inventory system we are modeling.

5.2 Time Dependent Distribution Analysis

The recoverable item inventory system will be modeled as a queueing system, as discussed in Chapter 3. In this case, we are not concerned with server failures since we need the line length distribution when the system is in a particular service state (i.e., a fixed number of operative channels). Thus, we want to describe the time dependent behavior of an $M(t)/M/k$ queue, where $M(t)$ is the shorthand notation for a non-stationary Poisson input process and $k = 0, 1, \dots, C$. This section will

provide the closed form solutions available for some special cases and then review some approximation procedures that are in the literature.

5.2.1 Closed Form Solutions

Consider a system for the period when exactly k service channels are operational, $k = 0, 1, \dots, C$. Without loss of generality, we will assume the period starts at $t = 0$. To describe the system's time dependent behavior during this period, we need the transition probability

$$P_k(a, b, t) = \Pr\{N(t) = b \mid N(0) = a \text{ and exactly } k \text{ service channels operational}\}.$$

For the case $k = 0$, we have a counting process with Poisson arrivals.

Thus given a time dependent arrival rate of $\lambda(t)$, then

$$P_0(a, b, t) = \Pr\{b - a \text{ arrivals in time } t\}$$

$$= \frac{e^{-m(t)} [m(t)]^{b-a}}{(b-a)!}, \quad b \geq a, \quad (5.3)$$

where

$$m(t) = \int_0^t \lambda(t) dt.$$

Closed form expressions for $P_k(a, b, t)$ for $k \geq 1$ are very complex. Saaty [31] provides a general procedure to find these distributions for queues with stationary Poisson arrivals. The procedure involves establishing the balance equations, obtaining a partial differential equation for a probability generating function, and using integral transforms to arrive at a transform representation of the distribution. To get the distribution from its transform requires a lengthy inversion process.

For the case $k = 1$, we have

$$P_1(a, b, t) = e^{-(\lambda + \mu)t} \left[\left(\frac{\mu}{\lambda} \right)^{(a-b)/2} I_{b-a}(2\sqrt{\lambda\mu} t) \right]$$

$$\begin{aligned}
& + \left(\frac{\mu}{\lambda}\right)^{(a-b+1)/2} I_{a+b+1}(2\sqrt{\lambda\mu} t) \\
& + \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^b \sum_{n=a+b+2}^{\infty} \left(\frac{\mu}{\lambda}\right)^{n/2} I_n(2\sqrt{\lambda\mu} t), \quad b \geq 1, \quad (5.4)
\end{aligned}$$

where $I_n(y) = \sum_{k=0}^{\infty} \frac{(y/2)^{n+2k}}{k! (n+k)!}$ is the modified Bessel function of the first kind. Obtaining solutions for $k \geq 2$ is extremely arduous. The resulting solutions are not easy to apply and are too complex to be useful for simple comparisons. Thus, we will consider approximation methods to arrive at these distributions.

5.2.2 Approximation Methods

The application and accuracy of approximations proposed in the literature depend upon the stationarity of the arrival process and the traffic intensity. Unfortunately, there is a paucity of computational comparisons among methods. We will review the results for multi-server queues with (1) stationary Poisson arrivals (M/M/k) and (2) non-stationary Poisson arrivals (M(t)/M/k).

Kotiah [19] uses an approximate transform inversion method to arrive at an approximate distribution for the line length. He uses the integral transform given by Saaty [31] to arrive at some numerical results for the case M/M/1. His numerical examples are accurate for $.5 \leq \rho \leq 4$ and, in some cases, provide bounds for the actual distribution. He describes a method for generating a distribution for the M/M/2 queue, but does not provide any numerical results.

For congestion cases, Newell's diffusion approximation (given in Chapter 3, equation (3.108)) is very accurate, especially as

ρ exceeds unity. Equation (3.108) can be applied to cases where $k \geq 2$; however, when $\rho < 1$, the accuracy suffers since the infinitesimal moments are not functions of the system size. This approximation has computational advantages over Kotiah's method.

For the $M(t)/M/1$ queue, Moore [21] provides the most general results for a single server queue. He uses an interactive procedure on an imbedded Markov chain to find the line length distribution. His model allows for bulk arrivals and Erlang distributed service times as well (so called, $M^X(t)/E_\gamma/1$ queue), and his numerical examples are accurate for $.3 \leq \rho(t) \leq .9$.

A computationally simpler approximation is given by Pokress [29]. He compares the finite server queue ($k > 1$) to the $M(t)/M/\infty$ queue and gets very accurate results when the range of $\rho(t)$ is less than .8. The probability distribution for the number of customers in an infinite server system, which is empty at time zero, is Poisson with mean

$$m(t) = \int_0^t e^{-\mu(t-x)} \lambda(x) dx,$$

where $\lambda(t)$ is the time dependent arrival rate.

Finally, Newell [25,26] proposes a diffusion approximation for the $M(t)/M/k$ queue which is similar in form to equation (3.108).

Define $F_k(x,t) = \text{Pr}\{\text{number of units in the system at time } t \text{ is } \leq x\}$

$$= \int_0^x P_k(0,y,t) dy.$$

Then

$$F_k(x,t) \approx \begin{cases} \phi\left(\frac{x-m(t)}{\sigma(t)}\right) - \phi\left(\frac{-m(t)}{\sigma(t)}\right) \exp\left[\frac{-2(k\mu-\lambda(t))x}{k\mu+\lambda(t)}\right], & x > 0, \\ 0, & x \leq 0, \end{cases}$$

where

$$\phi\left(\frac{x-m(t)}{\sigma(t)}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-m(t)}{\sigma(t)}} e^{-y^2/2} dy,$$

$$m(t) = A - \int_0^t [k\mu - \lambda(y)] dy,$$

and

$$\sigma^2(t) = B + \int_0^t [k\mu + \lambda(y)] dy. \quad (5.5)$$

Newell [26] claims that for t "large enough", the constants, A and B , are negligible. He suggests that for $k > 1$, the accuracy improves when $\rho(t) < 1$ and as t gets large. The approximations of Newell and Pokress have computational advantages over Moore's method.

5.3 Passage Time Distributions

The distribution of transition times from one service state to another is developed in this section. The flows in the system are displayed in Figure 5.2.

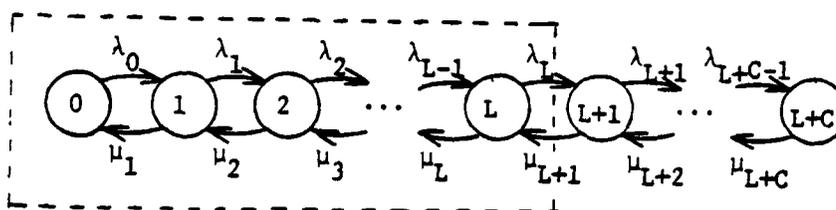


Figure 5.2: Transition Flows for Number of Inoperative Servers

Since the state transitions are only to adjacent states, we have a birth-death process. The states $\{0, 1, \dots, L\}$ are outlined because all channels are in operating order when the system is in these service states. The work of Keilson, et. al., [7,13,14] assume that the birth

and death processes are Poisson. This section will discuss their work and provide extensions for Erlang and deterministic death (server repair) processes.

5.3.1 Previous Results

If we assume the birth and death processes are Poisson, then the birth and death times are exponentially distributed. Define $S_k(t)$ as the probability density for the time in service state k . Then we have

$$S_k(t) = (\lambda_k + \mu_k) e^{-(\lambda_k + \mu_k)t}, \quad k = 0, 1, \dots, L + C, \quad (5.6)$$

where $\mu_0 = \lambda_{L+C} = 0$. In addition, the probability of a transition from service state i to state j is

$$P_{ij} = \begin{cases} \frac{\lambda_i}{\lambda_i + \mu_i}, & j = i + 1 \\ \frac{\mu_i}{\lambda_i + \mu_i}, & j = i - 1 \\ 0, & \text{otherwise.} \end{cases} \quad (5.7)$$

To best describe the service system, we need the passage time densities for transitions in the number of operational channels. For the service state set $\{0, 1, \dots, L\}$ we have C operational channels and for service states $k > L$ we have $C + L - k$ operational channels. Keilson, et. al., describe the passage time from a "good" state to a "bad" state. In our case, the GOOD state is defined as the service state $\{0, 1, \dots, L\}$ and the BAD state is defined as the remaining state set $\{L + 1, \dots, L + C\}$. This is illustrated in Figure 5.3. The server system can wander through the GOOD state for some time before going to service

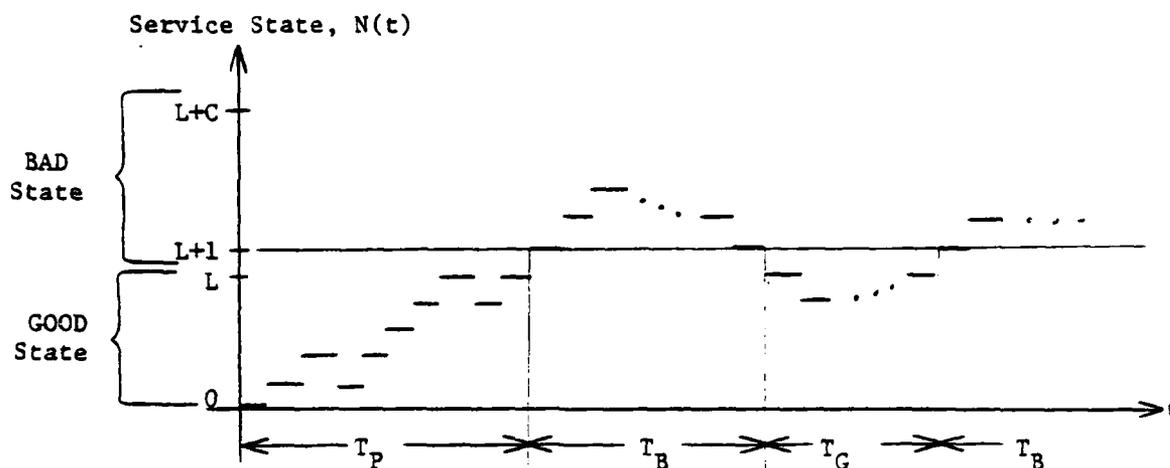


Figure 5.3: Transitions from GOOD to BAD States

state $L + 1$ and thus enter the BAD state. For this reason, the distribution of the passage time from the GOOD to the BAD state is not as simple in form as the distributions previously described.

Define the time from the perfect state, T_P , to be the passage time from service state 0 to $L + 1$ and the post recovery failure time, T_G , to be the passage time from first entering service state L to first entering state $L + 1$. The time T_P can be useful when observing a system which starts in perfect operating condition (e.g., an aircraft carrier starting a cruise). Let

$S_P(t)$ = the probability density function for T_P ,

$S_G(t)$ = the probability density function for T_G ,

$\bar{F}_P(t)$ = the survival function of T_P
 $= \int_t^{\infty} S_P(y) dy,$

$\bar{F}_G(t)$ = the survival function of T_G
 $= \int_t^{\infty} S_G(y) dy, \text{ and}$

$S_k^+(t)$ = the probability density function of the first passage time from service state k to $k + 1$.

Then we have

$$S_p(t) = S_0^+ * S_1^+ * \dots * S_L^+(t), \quad (5.8)$$

and

$$S_G(t) = S_L^+(t), \quad (5.9)$$

where "*" denotes convolution. Using equations (5.6) and (5.7), we have

$$S_0^+(t) = \lambda_0 e^{-\lambda_0 t}, \quad (5.10)$$

and

$$S_k^+(t) = \lambda_k e^{-(\lambda_k + \mu_k)t} + \mu_k e^{-(\lambda_k + \mu_k)t} * S_{k-1}^+(t) * S_k^+(t), \quad k \geq 1. \quad (5.11)$$

Graves and Keilson [7] take advantage of the convolution properties of Laplace transforms to arrive at the desired densities. Define

$$\sigma_k^+(s) = \text{Laplace transform of } S_k^+(t) = \int_0^{\infty} e^{-st} S_k^+(t) dt.$$

Then we get the relations

$$\begin{aligned} \sigma_0^+ &= \frac{\lambda_0}{s + \lambda_0}, \\ \sigma_k^+(s) &= \frac{\lambda_k}{s + \lambda_k + \mu_k - \mu_k \sigma_{k-1}^+(s)}, \quad k=1,2,\dots, \\ &= \frac{\lambda_k P_k(s)}{P_{k+1}(s)} \end{aligned} \quad (5.12)$$

$$\text{and} \quad \sigma_p(s) = \prod_{k=0}^L \sigma_k^+(s), \quad (5.13)$$

where $P_0(s) = 1$

$$P_1(s) = s + \lambda_0$$

$$P_{k+1}(s) = (s + \lambda_k + \mu_k) P_k(s) - \mu_k \lambda_{k-1} P_{k-1}(s), \quad k=2,3,\dots \quad (5.14)$$

Examining the polynomials $P_k(s)$ we get the following properties:

Theorem 5.1. (Graves and Keilson)

- (i) $P_k(s)$ has k simple roots, q_1, \dots, q_k ,
 (ii) $P_{k+1}(s)$ has $k+1$ simple roots, r_1, \dots, r_{k+1} , such that

$$-\infty < r_{k+1} < q_k < r_k < \dots < q_1 < r_1 < 0.$$

Since $P_k(s)$ is a polynomial of degree k , Theorem 5.1 provides the complete factorization of $P_k(s)$. Property (ii) of the theorem proves that the roots of each polynomial provide upper and lower bounds for all but one of the roots of the succeeding polynomial. Thus, the roots for each polynomial can be determined using a recursive algorithm. From equation (5.13) we have

$$\begin{aligned} \sigma_p(s) &= \prod_{k=0}^L \sigma_k^+(s) = \prod_{k=0}^L \frac{\lambda_k P_k(s)}{P_{k+1}(s)} = \frac{\lambda_0 \lambda_1 \dots \lambda_L}{P_{L+1}(s)} \\ &= \frac{\lambda_0 \lambda_1 \dots \lambda_L}{L+1} \prod_{k=1}^{L+1} (s - r_k), \end{aligned} \quad (5.15)$$

where r_i , $i = 1, \dots, L+1$, are the distinct negative roots of $P_{L+1}(s)$.

Using partial fractions and the fact that $\frac{1}{s-r_k}$ is the transform of $e^{r_k t}$, we get the following result:

Theorem 5.2.

The probability density function of the passage time from the perfect state, T_p , is

$$S_p(t) = \sum_{k=1}^{L+1} \beta_k e^{r_k t}, \quad t \geq 0, \quad (5.16)$$

where $r_k < 0$ and $\beta_k = \frac{\lambda_0 \lambda_1 \dots \lambda_L}{L+1} \frac{\prod_{i=1, i \neq k}^{L+1} (r_k - r_i)}{\prod_{i=1, i \neq k}^{L+1} (r_k - r_i)}$, $k=1, \dots, L+1$

Notice $\sum_{k=1}^{L+1} \beta_k = 1$; however $\beta_k < 0$ is possible for some k . In a similar fashion we get:

Theorem 5.3.

The probability density function of the post recovery failure time, T_G , is

$$S_G(t) = \sum_{k=1}^{L+1} \alpha_k e^{r_k t}, \quad (5.17)$$

where q_k , $k = 1, \dots, L$, are the distinct negative roots of $P_L(s)$,

r_k , $k = 1, \dots, L+1$ are the distinct negative roots of $P_{L+1}(s)$,

and

$$\alpha_k = \frac{\lambda_{L+1} \prod_{i=1}^L (r_k - q_i)}{\prod_{i=1, i \neq k}^{L+1} (r_k - r_i)}$$

In this case, $\sum_{k=1}^{L+1} \alpha_k = 1$ and $\alpha_k \geq 0$ due to property (ii) of Theorem 5.1.

Notice both distributions are mixed exponential distributions. This leads to some interesting properties of these distributions. Below are some useful properties given by Keilson [13].

Definition

A function f is log concave (convex) if $\ln(f)$ is a concave (convex) function.

Theorem 5.4.

If $S_1(t)$ and $S_2(t)$ are log convex probability densities defined on some connected interval T , with means μ_i and variances σ_i^2 , then

(i) $\lambda_1 S_1(t) + \lambda_2 S_2(t)$ is log convex on T where $\lambda_1 + \lambda_2 = 1$ and $\lambda_1, \lambda_2 \geq 0$,

(ii) $\bar{F}_1(t) = \int_t^\infty S_1(y) dy$ is log convex,

(iii) $\frac{\mu_i^2}{\sigma_i^2} \leq 1$, and

(iv) there is a distribution function G such that

$$S_1(t) = \int_0^\infty ye^{-yt} dG(y).$$

Theorem 5.5

If $S_1(t)$ and $S_2(t)$ are log concave probability densities defined on some connected interval T , with means μ_i and variances σ_i^2 , then

(i) $S_1(t) * S_2(t)$ is log concave on T ,

(ii) $\bar{F}_1(t) = \int_t^\infty S_1(y) dy$ is log concave,

(iii) $\frac{\mu_i^2}{\sigma_i^2} \geq 1$, and

(iv) $\lim_{t \rightarrow \infty} \frac{S_1(t)}{e^{-\mu t}} = 0$ for some $\mu > 0$.

These theorems have obvious extensions to include discrete distributions. The exponential distribution is the only function which is log concave and log convex. Most probability densities which are unimodal are log concave, such as: the normal, binomial, Poisson, geometric, negative binomial, beta and Erlang distributions. The gamma density

$\frac{x^r e^{-x}}{\Gamma(r+1)}$ with $-1 < r < 0$, however, is log convex.

Theorem 5.4 proves that $S_G(t)$ and $\bar{F}_G(t)$ are log convex and Theorem 5.5 proves that $S_p(t)$ and $\bar{F}_p(t)$ are log concave. Moreover, from property (iv) of both theorems, we can expect the tails of $S_p(t)$ and $S_G(t)$ to be bounded by negative exponential curves. Since both $S_G(t)$ and $S_p(t)$ are mixed exponentials, we expect $\bar{F}_G(t)$ and $\bar{F}_p(t)$ to have exponential tails. The numerical examples of Graves and Keilson [7] verify this. A pair of typical survival functions are illustrated in Figure 5.4.

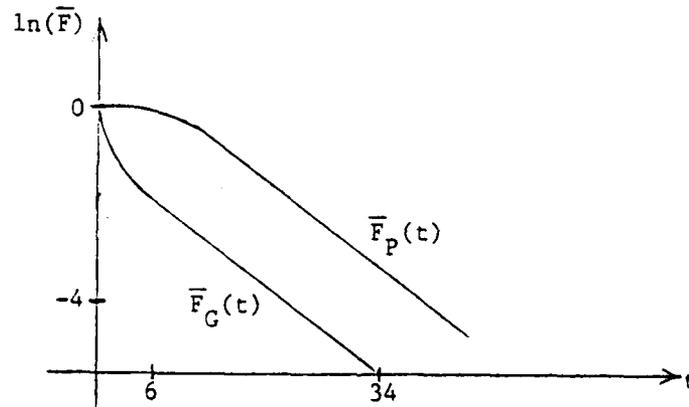


Figure 5.4: Log Survival Function Comparison [7]

5.3.2 Extension to Erlang Distributed Repair Times

We now consider a system with a Poisson failure process and Erlang distributed repair times. A failed server will require k "phases" of repair, each exponentially distributed in length. For this case, the service state space must be modified so that the arrival and departure processes can be accurately expressed as functions of the state of the system. We will use the method of phases and will assume that the server repair facility has only one repairman whose repair rate is proportional to the number of units in the system. It is sufficient to

expand the service state space to the number of phases in the system, $\{0, 1, \dots, k, k+1, \dots, 2k, \dots, k(L+C)\}$. Now the GOOD state is the set $G = \{0, 1, \dots, kL\}$ and the BAD state is the set $B = \{kL+1, \dots, k(L+C)\}$. The BAD state can only be entered upon the arrival of a failed server so we will examine the state of the system at each arrival. After describing the state changes between arrivals, we will construct a transition matrix. Using the structure of this matrix, we can derive the passage time distribution.

The service state changes at each arrival epoch, X_j , are illustrated in Figure 5.5. Let $N(t)$ be the service state at time t . Notice that $N(X_j) \geq k$, for all X_j , since the state is examined just after the arrival of k phases. We know the interarrival times, $X_j - X_{j-1}$, are independent and exponentially distributed, with rate $C\bar{\xi}$, while the system is in the GOOD state. The number of phase completions in an interarrival period is dependent upon the initial service state. The

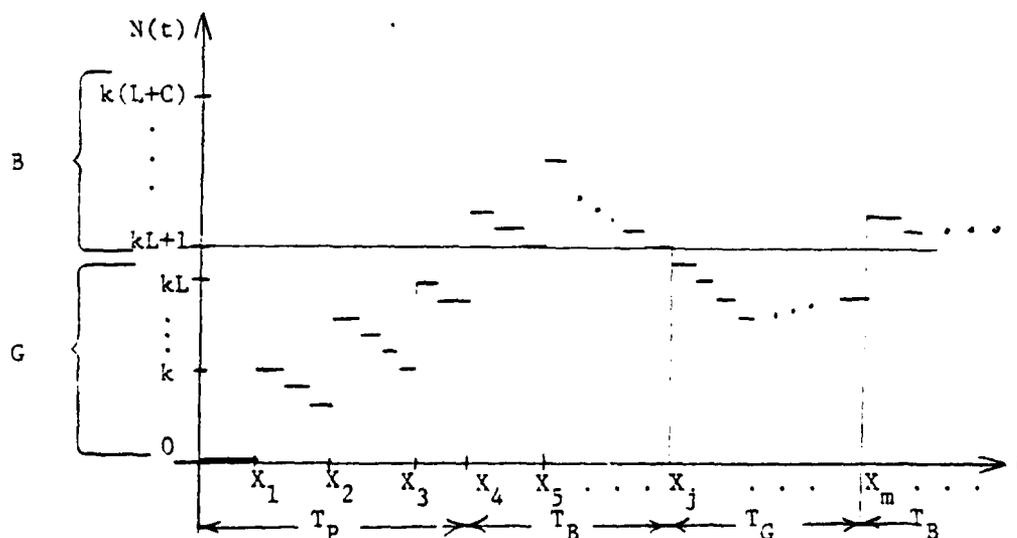


Figure 5.5: Service State Changes Between Arrival Epochs

interdeparture time for a system in service state i is exponentially distributed with rate $[i/k] \cdot \eta$, where $[i/k]$ represents the number of servers being repaired. Thus, the next service state transition is governed by the following probabilities:

$$d_n = \Pr\{\text{a phase completion is the next event} \mid \text{in service state } n\}$$

$$= \begin{cases} \frac{[n/k]\eta}{C\xi + [n/k]\eta}, & 0 \leq n \leq kL \\ \frac{[n/k]\eta}{(C+L - [n/k]\xi + [n/k]\eta)}, & kL+1 \leq n \leq k(L+C), \end{cases} \quad (5.18)$$

and

$$\begin{aligned} U_n &= \Pr\{\text{a server failure is the next event} \mid \text{in service state } n\} \\ &= 1 - d_n. \end{aligned} \quad (5.19)$$

Each service state transition is independent of the previous one because of the memoryless property of the exponential distribution. At an arrival epoch, we add k phases to the current service state. Thus we have the following transition probabilities:

$$P_{ij} = \Pr\{\text{in service state } j \text{ at an arrival epoch} \mid \text{in service state } i \text{ at the previous epoch}\}$$

$$= \Pr\{i + k - j \text{ phase completions in an interarrival period, } i + k - j \geq 0 \mid \text{start in service state } i\}$$

$$= d_i d_{i-1} \cdots d_{j-k+1} U_{j-k}$$

$$= \prod_{n=j-k+1}^i d_n U_{j-k}, \quad i \geq j - k. \quad (5.20)$$

The transition probability matrix is:

$$P = \left[\begin{array}{cccc|cccc} 0 & \dots & 0 & P_{0k} & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & P_{1k} & P_{1,k+1} & 0 & \dots & 0 & 0 & \dots & 0 \\ \cdot & & \cdot & \cdot & & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & & \cdot & & \cdot & \cdot & & \cdot \\ 0 & \dots & 0 & P_{kL,k} & \cdot & \dots & \cdot & P_{kL,kL} & P_{kL,kL+1} & \dots & P_{kL,k(L+C)} \\ \hline 0 & \dots & 0 & P_{kL+1,k} & \cdot & \dots & \cdot & P_{kL+1,kL} & P_{kL+1,kL+1} & \dots & P_{kL+1,k(L+C)} \\ \cdot & & \cdot & \cdot & & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & & \cdot & & \cdot & \cdot & & \cdot \\ 0 & \dots & 0 & P_{k(L+C),k} & \cdot & \dots & \cdot & P_{k(L+C),kL} & P_{k(L+C),kL+1} & \dots & P_{k(L+C),k(L+C)} \end{array} \right]$$

$$= \left[\begin{array}{cc|c} \overbrace{\quad}^{kL+1} & \overbrace{\quad}^{kC} & \\ \hline A & B & \left. \vphantom{\begin{matrix} A \\ B \end{matrix}} \right\} kL+1 \\ C & D & \left. \vphantom{\begin{matrix} C \\ D \end{matrix}} \right\} kC \end{array} \right]$$

Let

a_i^n represent the i^{th} row of A^n ; $n = 0, 1, \dots$,

b_i represent the i^{th} row of B , $i = 0, 1, \dots, kL$, and

$\underline{1}$ represent the kC - dimension column vector $(1, 1, \dots, 1)^T$.

Notice that submatrix A represents transitions from the GOOD state to the GOOD state and submatrix B represents transitions from the GOOD state to the BAD state. If we define N_p to be the number of arrivals required to enter the BAD state from the perfect state (e.g. $N_p = 4$ in Figure 5.5), then

$$\Pr\{N_p = 1\} = \sum_{j \in B} P_{0j} = \sum_{j=kL+1}^{k(L+C)} P_{0j} = b_0 \cdot \underline{1}, \quad (5.21)$$

$$\begin{aligned} \Pr\{N_p = 2\} &= \sum_{i \in G} \sum_{j \in B} P_{0i} P_{ij} = \sum_{i=0}^{kL} P_{0i} \sum_{j=kL+1}^{k(L+C)} P_{ij} \\ &= \sum_{i=0}^{kL} P_{0i} (b_i \cdot \underline{1}) = a_0 \cdot B \cdot \underline{1}, \end{aligned}$$

$$\begin{aligned} \Pr\{N_p = 3\} &= \sum_{i \in G} \sum_{j \in G} \sum_{k \in B} P_{0i} P_{ij} P_{jk} \\ &= \sum_{i \in G} P_{0i} \left(\sum_{j \in G} P_{ij} \sum_{k \in B} P_{jk} \right) = \sum_{i \in G} P_{0i} (a_L \cdot B \cdot \underline{1}) \\ &= a_0^2 B \cdot \underline{1}, \end{aligned}$$

and, in general,

$$\Pr\{N_p = n\} = a_0^{(n-1)} \cdot B \cdot \underline{1}, \quad n = 2, 3, \dots \quad (5.22)$$

Similarly, defining N_G as the number of arrivals required to enter the BAD state just after entering the GOOD state, we get:

$$\Pr\{N_G = 1\} = \sum_{j \in B} P_{kL,j} = b_{kL} \cdot \underline{1}, \quad (5.23)$$

$$\begin{aligned}
\Pr\{N_G = 2\} &= \sum_{i \in G} \sum_{j \in B} P_{kL,i} P_{i,j} \\
&= \sum_{i \in G} P_{kL,i} \sum_{j=kL+1}^{k(L+C)} P_{ij} = \sum_{i=0}^{kL} P_{kl,i} (b_i \cdot \underline{1}) \\
&= a_{kL} \cdot B \cdot \underline{1},
\end{aligned}$$

$$\begin{aligned}
\Pr\{N_G = 3\} &= \sum_{i \in G} \sum_{j \in G} \sum_{k \in B} P_{kL,i} P_{i,j} P_{j,k} \\
&= \sum_{i \in G} P_{kL,i} \left(\sum_{j \in G} P_{i,j} \sum_{k \in B} P_{j,k} \right) \\
&= \sum_{i=0}^{kL} P_{kl,i} (a_i \cdot B \cdot \underline{1}) \\
&= a_{kL}^2 \cdot B \cdot \underline{1},
\end{aligned}$$

and, in general

$$\Pr\{N_G = n\} = a_{kL}^{(n-1)} \cdot B \cdot \underline{1}, \quad n = 2, 3, \dots \quad (5.24)$$

The distribution for T_p and T_G can be derived from the above work. For example, if $N_p = n$, then the time from the perfect state, T_p , is the sum of n identically distributed exponential random variables, i.e., T_p is Erlang distributed with mean $n/C\bar{c}$ and shape parameter n . Thus, we get

$$\bar{F}_p(t) = \Pr\{T_p > t\}$$

$$= 1 - \sum_{n=1}^p Y_n(t) \Pr\{N_p = n\}, \quad t > 0, \quad (5.25)$$

where

$$Y_n(t) = \int_0^t \frac{(n^2/C\xi)^n}{(n-1)!} t^{n-1} \exp(-n^2 t/C\xi) dt. \quad (5.26)$$

Similarly,

$$\bar{Y}_G(t) = 1 - \sum_{n=1}^{\infty} Y_n(t) \Pr\{N_G = n\}, \quad t > 0 \quad (5.27)$$

where $Y_n(t)$ is defined above.

5.3.3 Extension To Deterministic Repair Times

If we assume the repair times are fixed in length, the queuing process becomes non-Markovian; thus, describing the service state of the system through time is an arduous task. The problem can be simplified by considering the state of the system at certain lattice points in time - specifically, at integer multiples of the repair time, R . In this case the GOOD state is the set $G = \{0, 1, \dots, L\}$ and the BAD state is the set $B = \{L + 1, \dots, L + C\}$. The state transitions and passage times are illustrated in Figure 5.6. We will assume that repair is only initiated on the failed servers present at the beginning of the repair period. This could be viewed as a periodic review inventory model where orders are placed for new servers every R days and the lead time is R days. Notice the BAD state can be entered during a repair period and exited at the completion of the period. As a result, we will have to examine the state of the system just prior to the end of a repair period.

We will use a method similar to the one used for Erlang distributed repairs. A transition probability matrix, P , will be created, in order to find a distribution for the number of service periods in a passage time. Then the results will be refined to allow for transitions into

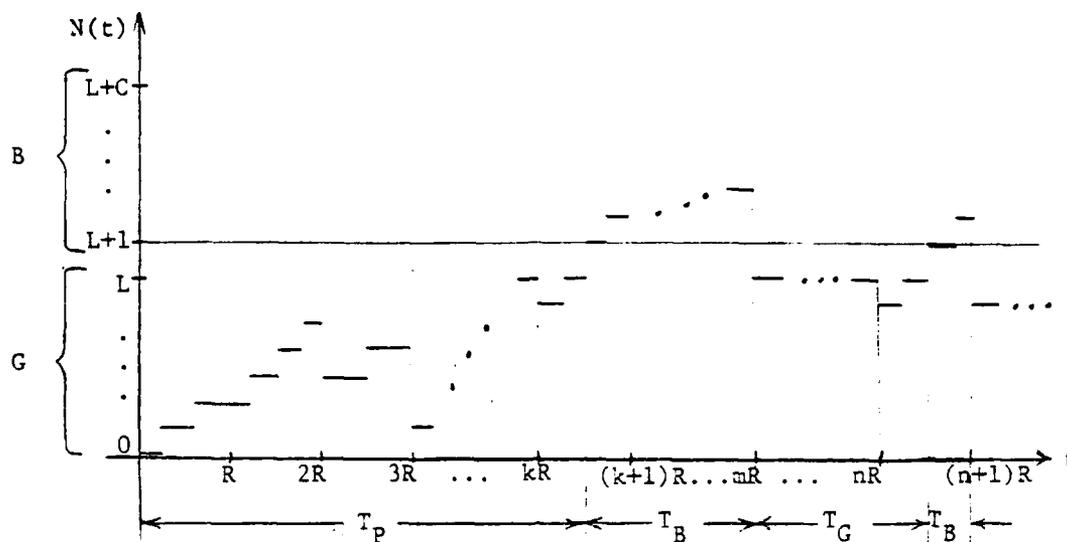


Figure 5.6: Service State Changes Between Service Epochs

the BAD state at non-lattice times. Define

$$P_{ij} = \{ \text{Pr in service state } j \text{ just prior to completion of} \\ \text{repair period} \mid \text{in service state } i \text{ at beginning} \\ \text{of repair period} \} \\ = \text{Pr} \{ (j-i) \text{ servers fail in time } R \mid \text{in service state } i \\ \text{at beginning of service period} \}.$$

As long as the system remains in the GOOD state, the servers will fail as Poisson events with rate $C\xi$. Thus, the number of failures in time R , given $j \leq L$, is Poisson distributed, i.e.,

$$P_{ij} = \frac{e^{-C\xi R} (C\xi R)^{j-i}}{(j-i)!}, \quad 0 \leq i \leq j \leq L. \quad (5.28)$$

Since the server failure rate is constant when j above is in the GOOD state, equation (5.28) is the probability of $(j-i)$ server failures in time R . Thus, we have the following important property:

$$P_{ij} = P_{0, (j-i)} \text{ whenever } j = 0, 1, \dots, L \text{ and } i \leq j. \quad (5.29)$$

The remaining probabilities, however, are not as simple because while the system is in the BAD state the server failure rates are dependent on the service state.

The interarrival times are independent; therefore the number of server failures, $N(t)$, is a renewal process. Define the distributions

$$F_{in}^+(t) = \Pr\{\text{the } n\text{th failure occurs at time } \leq t \mid \\ i \text{ servers are being repaired at time } t = 0\},$$

and

$$F_j^+(t) = \Pr\{\text{the interarrival time between failure } j \text{ and} \\ \text{failure } (j+1) \text{ is } \leq t\}, \quad j = 0, 1, \dots, C+L.$$

Then

$$F_{in}^+(t) = F_1^+ * F_{i+1}^+ * \dots * F_{n-1}^+(t), \quad (5.30)$$

where $*$ denotes convolution.

For the one-step transition distributions, we know

$$F_j^+(t) = \begin{cases} 1 - e^{-C\xi t} & , 0 \leq j \leq L \\ 1 - e^{-(C+L-j)\xi t} & , L+1 \leq j \leq L+C \end{cases} \quad (5.31)$$

Using the Laplace transforms of equations (5.30) and (5.31) we can find the distributions $F_{ij}(t)$, $j = L, L+1, \dots, L+C$.

From renewal theory [28] we now have the remaining probabilities

$$P_{ij} = \Pr\{(j-1) \text{ server failures in time } R \mid \\ i \text{ servers are being repaired at the beginning} \\ \text{of the service period}\} \\ = \begin{cases} F_{ij}(R) - F_{i,j+1}(R), & j = L, L+1, \dots, C+L-1, \\ F_{i,L+C}(R), & j = L+C. \end{cases} \quad (5.32)$$

Now a transition probability matrix, P , can be established:

$$P = \begin{bmatrix} P_{00} & P_{01} & \cdots & P_{0,L} & P_{0,L+1} & \cdots & P_{0,L+C} \\ 0 & P_{11} & & P_{1,L} & P_{1,L+1} & \cdots & P_{1,L+C} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & P_{L,L} & P_{L,L+1} & \cdots & P_{L,L+C} \\ \hline 0 & \cdots & \cdots & 0 & P_{L+1,L+1} & \cdots & P_{L+1,L+C} \\ \vdots & & & \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & P_{L+C,L+C} \end{bmatrix}$$

$$= \begin{bmatrix} \overbrace{\begin{bmatrix} A & B \end{bmatrix}}^{L+1} \\ \hline \underbrace{\begin{bmatrix} 0 & D \end{bmatrix}}_C \end{bmatrix} \begin{matrix} \left. \begin{matrix} L+1 \\ C \end{matrix} \right\} \\ \left. \begin{matrix} L+1 \\ C \end{matrix} \right\} \end{matrix}$$

Let a_i represent the i^{th} row of the submatrix A ,

b_i represent the i^{th} row of the submatrix B , and

$\underline{1}$ represent the C -dimensional column vector $(1,1,\dots,1)^T$.

Let N_p be the number of repair periods prior to entering the BAD state given the initial service state was zero (the "perfect" state). For example $N_p = k$ in Figure 5.6.

Define the matrix

$$W = \begin{bmatrix} P_{00} & P_{01} & \cdot & \cdot & P_{0,L-2} & P_{0,L-1} & P_{0L} \\ P_{00} & P_{01} & \cdot & \cdot & P_{0,L-2} & P_{0,L-1} & 0 \\ P_{00} & P_{01} & \cdot & \cdot & P_{0,L-2} & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ P_{00} & P_{01} & \cdot & \cdot & \cdot & \cdot & 0 \\ P_{00} & 0 & \cdot & \cdot & \cdot & 0 & 0 \end{bmatrix}$$

where w_i is the i^{th} row of W . Then using the property (5.29), we have

$$\Pr\{N_p = 0\} = \sum_{i \in B} P_{0i} = \sum_{i=L+1}^{L+C} P_{0i} = b_0 \cdot \underline{1}. \quad (5.33)$$

$$\begin{aligned} \Pr\{N_p = 1\} &= \sum_{i \in G} \sum_{j \in B} P_{0i} P_{ij} = \sum_{i=0}^L P_{0i} \sum_{j=L+1}^{L+C} P_{ij} \\ &= \sum_{i=0}^L P_{0i} b_i \cdot \underline{1} = a_0 \cdot B \cdot \underline{1}. \end{aligned}$$

$$\begin{aligned} \Pr\{N_p = 2\} &= \sum_{i \in G} \sum_{j \in G} \sum_{k \in B} P_{0i} P_{ij} P_{j-i,k} \\ &= \sum_{i=0}^L P_{0i} \sum_{j=i}^L P_{ij} \sum_{k=L+1}^{L+C} P_{j-i,k} \\ &= \sum_{i=0}^L P_{0i} \sum_{j=0}^{L-i} P_{0,j} \sum_{k=L+1}^{L+C} P_{jk} \\ &= \sum_{i=0}^L P_{0i} \sum_{j=0}^{L-i} P_{0j} b_j \cdot \underline{1} = \sum_{i=0}^L P_{0i} w_i \cdot B \cdot \underline{1} \\ &= a_0 \cdot w \cdot B \cdot \underline{1} \end{aligned}$$

and by an inductive argument, we have

$$\Pr\{N_p = k\} = a_0 \cdot w^{k-1} \cdot B \cdot \underline{1}, \quad k = 1, 2, \dots \quad (5.34)$$

Let N_G be the number of repair periods from the time the system "recovers" from a BAD service state, to just prior to re-entry into the BAD state. For example, in Figure 5.6, $N_G = n-m$. The service state at the beginning of our observation is unknown; it is the number of server failures during the last repair period which contained a BAD service state. This situation is expanded in Figure 5.7.

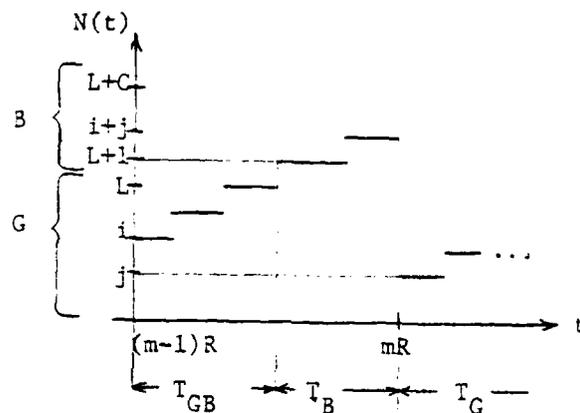


Figure 5.7: GOOD to BAD to GOOD State Transition

Since the server failure rate is dependent, the state j depends upon the previous state i . (Notice, service state i need not be in G .) We are assuming that the number of servers in repair at time $(m-1)R$ does not depend on any initial conditions (i.e., we assume steady state condition). Then $\{Pr N[(m-1)R] = i\} = \pi_i$, where π_i is the stationary probability of being in service state i . We want the probability

$$\begin{aligned}
 v_j &= Pr \{N(mR) = j, j \in G, N[(m-1)R] = i, i + j \in B\} \\
 &= \sum_{\substack{i+j \in B \\ L+C-j \\ i=L-j}} Pr \{j \text{ server failures in time } R \mid N[(m-1)R] = i\} \cdot \pi_i \\
 &= \sum_{i=L-j}^{L+C-j} P_{i, i+j} \pi_i, \quad j = 1, \dots, L.
 \end{aligned} \tag{5.35}$$

Let $\underline{v} = (v_1, \dots, v_L)^T$. Then in a manner similar to that used for N_p , we

$$\begin{aligned} \text{have } \Pr\{N_G = 0\} &= \sum_{i=0}^L \sum_{j=L+1}^{L+C} v_i P_{ij} = \underline{v} \cdot \underline{B} \cdot \underline{1}, \\ \Pr\{N_G = 1\} &= \sum_{i \in G} \sum_{j \in G} \sum_{k \in B} v_i P_{ij} P_{(j-i), k} \\ &= \sum_{i=0}^L v_i \sum_{j=i}^L P_{ij} \sum_{k=L+1}^{L+C} P_{(j-i), k} \\ &= \sum_{i=0}^L v_i \sum_{j=0}^{L-i} P_{0j} \sum_{k=L+1}^{L+C} P_{j, k} \\ &= \underline{v} \cdot \underline{w} \cdot \underline{B} \cdot \underline{1}, \end{aligned}$$

and, in general,

$$\Pr\{N_G = k\} = \underline{v} \cdot \underline{w}^k \cdot \underline{B} \cdot \underline{1}, \quad k = 0, 1, \dots \quad (5.36)$$

We now have distributions on the number of whole repair periods in the times T_p and T_G . It remains to find the distribution for the time within a period until the BAD state is entered. This time is represented in Figure 5.7 as T_{GB} . This passage time is dependent upon the service state, i , at the beginning of the period, where $i \in G$. We have the stationary probability $u_i = \Pr\{N[(m-1)R] = i \mid i \in G\}$

$$= \frac{\pi_i}{\sum_{i=0}^L \pi_i}, \quad i=0, 1, \dots, L. \quad (5.37)$$

The times in each state i , $i+1, \dots, L$ are independent and exponentially distributed with mean $1/C\xi$. Thus the passage time distribution is

$$\Pr\{T_{GB} \leq t \mid N[(m-1)R] = i\} = Y_{L-i+1}(t), \quad i = 0, 1, \dots, L, \quad (5.38)$$

where $Y_k(t)$ is the Erlang distribution with mean $k/C\xi$ and shape parameter k . Combining (5.37) and (5.38) yields

$$\Pr \{ T_{GB} \leq t \} = \sum_{i=0}^L \pi_i Y_{L-i+1}(t). \quad (5.39)$$

Combining equations (5.34), (5.36) and (5.39) gives us the desired distributions (for $k = 0, 1, \dots$)

$$\Pr \{ T_P \leq t \} = \Pr \{ N_P = k \} \cdot \Pr \{ T_{GB} \leq t \}, \quad kR \leq t \leq (k+1)R, \quad (5.40)$$

and

$$\Pr \{ T_G \leq t \} = \Pr \{ N_G = k \} \cdot \Pr \{ T_{GB} \leq t \} \quad kR \leq t \leq (k+1)R. \quad (5.41)$$

It may be more realistic to allow repair to be initiated on a server as soon as it fails as opposed to waiting until the end of the repair period. Then conceptually, if we consider the system at lattice points only (every R time units), the analysis could be very simple. Suppose P_G is the probability of being in the GOOD state at the end of a repair period, independent of the starting service state. Then the distribution for N_{GB} , the number of repair periods in the GOOD state before reentering the BAD state, would be

$$\Pr \{ N_{GB} = k \} = P_G^k (1 - P_G), \quad k = 0, 1, \dots \quad (5.42)$$

Unfortunately, P_G is dependent upon the starting state which complicates the analysis considerably.

In the previous analysis, the service state transitions could only be in one direction between lattice points. If repair on a failed server begins immediately and not at lattice points, then the service state can wander in either direction. Figure 5.8 shows how the system can conceivably enter and exit the BAD state during a period of length R . This transition would change the server failure rate and thus affect

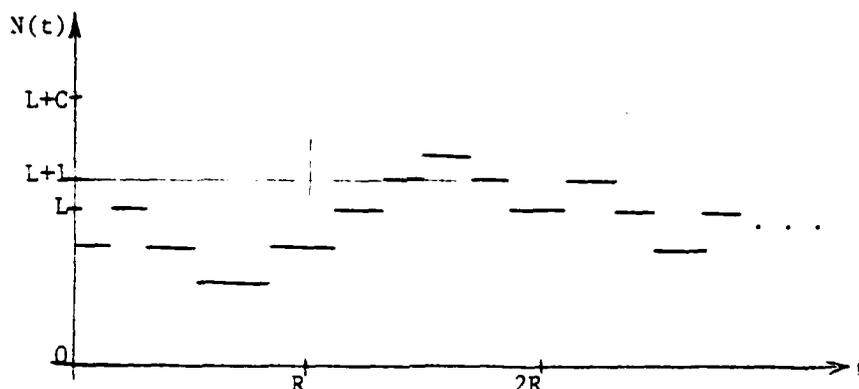


Figure 5.8: Transitions When Repairs Initiated Immediately

the number of server failures in the period R . Since the repair time is R , the number of servers in repair at the end of the period is precisely the number that failed during the period. To find the distribution for this number, we need to know (1) how many servers are being repaired at the beginning of the period and (2) how much repair remains to be done on each server. Given certain considerations, these problems may not be significant factors.

Suppose that the service channel failure rate is small compared to the repair rate and the chance of entering and exiting the BAD state in a single period is negligible. Define P_G as the probability of staying in the GOOD state in a time interval R . If $N(t)$ is the number of server failures in time t while in the GOOD state, then

$$P_G = \Pr \{N(R) \leq L\} = \sum_{i=0}^L \Pr \{N(R) = i\} \quad (5.43)$$

Since we remain in the GOOD state for the entire period, then $N(R)$ is Poisson distributed (equation (5.28)) and we can apply equation (5.42) to get a simple result.

We have now completely described the passage time distributions for the service states given exponential, Erlang, and deterministic distributed repair times. These distributions can be integrated with actual or

approximate time dependent line length distributions to describe the inventory system's time dependent behavior.

CHAPTER VI
CONCLUDING REMARKS

This thesis presented stationary and non-stationary (time dependent) analyses for a recoverable item inventory system. It differs from the previous work done in this area because the capacity of the service facility was limited and variable due to server breakdowns.

The mathematical modeling of the system in Chapter II demonstrated the need to derive a representation for the distribution of the number of units in the system. In Chapter III we showed that analytic attempts to find this distribution yield neither comprehensible closed form solutions nor a means for comparison between systems. The development of a diffusion approximation, however, did give us some insight into the nature of the system's performance. The approximate distribution was found to have a geometric tail which led to a simple way to obtain a solution to an optimization problem. The solution of this problem can be used by managers to allocate limited resources optimally among item spares and repair facilities.

The time dependent analysis of Chapter V provides a basis for examining the short term or non-stationary behavior of the recoverable item inventory system. A study of the distribution for the times between service channel failures gives a technique to analyze the effects of different service facility designs. For example, the frequency and duration of changes in the service facility's performance can be calculated. This information can also be integrated with transient line length distributions to provide a measure of overall system performance through time.

There are several different areas where this work can be extended. One obvious extension would be to consider multi-echelon systems. The result would be a modification of Sherbrooke's METRIC model [32] to include a finite number of servers. In the same manner, items which require service could be composed of sub-units which need to be replaced, repaired, and stocked. This multi-indentured concept is considered for the adequate server case in Muckstadt's MOD-METRIC model [23].

Another possible extension is to model the service channel failures as partial breakdowns due to sub-unit failures. Typically, the service channel sub-units can be removed and the service station's ability to perform is only partially affected. An example of this is an electronics test station which can independently repair radio and radar units. Failure of the radio section may not affect the ability to service radars. Thus an investment in service channel spares should be considered for sub-components and not for entire servers.

This situation requires a complex reliability analysis. Spare sub-units can be considered as elements in parallel with the installed sub-unit. Consequently, the entire repair station could be modeled as a system composed of sub-components linked in series and parallel. The purchase of an additional sub-unit would provide a certain marginal improvement of the service channel's reliability. If this phenomenon can be quantified, then performance curves could be derived for fixed levels of investment in server sub-units and the remaining analysis should be similar to that performed in Chapter IV.

Chapter V suggests one final area for future research with its discussion of the dearth of comparisons between the different time-

dependent distributions for queues found in the literature. It would be instructive to have a standard set of common time-dependent problems (including non-stationary inputs to the queue) to provide a means for comparing the various proposed approximation methods.

BIBLIOGRAPHY

- [1] B. Avi-Itzhak and P. Naor, "Some Queuing Problems With The Service Station Subject To Breakdown", Opns. Res. 11, p.303-320 (1963).
- [2] Z. Barzily and D. Gross, "Some Practical Considerations In The Application Of Finite Source Queueing Models", Technical Paper Serial T-360, Institute For Management Science and Engineering, George Washington University, Washington, D.C. (1977).
- [3] J.D. Blackburn, "Optimal Control of Queueing Systems With Intermittent Service", Technical Report 8, Department of Operations Research, Stanford University, Stanford, Calif. (1971).
- [4] D. Blackwell, "A Renewal Theorem", Duke Math. Journal 15, p.145-151 (1948).
- [5] G.F. Feeney and C.C. Sherbrooke, "A System Approach to Base Stockage of Recoverable Items", The RAND Corporation, RM-4720-PR (1965).
- [6] M.J. Fischer, "An Approximation To Queueing Systems With Interruptions", Management Science 24, p.338-344 (1977).
- [7] S.C.Graves and J. Keilson, "A Methodology for Studying The Dynamics of Extended Logistic Systems", Working Paper Series 7729, Graduate School of Management, University of Rochester, Rochester, N.Y. (1977).
- [8] D. Gross and C.M. Harris, Fundamentals of Queueing Theory, Wiley, New York (1974).
- [9] D. Gross, H.D. Kahn, and J.D. Marsh, "Queueing Models for Spares Inventory and Repair Capacity", Technical Paper Serial T-344, Institute for Management Science and Engineering, George Washington University, Washington, D.C. (1977).
- [10] B. Halachmi and W.R. Franta, "On Diffusion Approximate Solutions to Open Multi-Server Queueing Systems", Technical Report 74-17, Computer Information and Control Sciences, University of Minnesota, Minneapolis, Minn. (1974).
- [11] D. Iglehart, "Limiting Diffusion Approximations for the Many Server Queue and The Repairman Problem", J. Applied Prob. 2, p.429-441 (1965).

- [12] D.L. Iglehart and W. Whitt, "Multiple Channel Queues in Heavy Traffic:I", Adv. in Applied Probability 2, p.150-177 (1970).
- [13] J. Keilson, "Markov Chain Models- Rarity and Exponentiality", Working Paper Series 7737, Graduate School of Management, University of Rochester, Rochester, N.Y. (1977).
- [14] J. Keilson and H.F. Ross, "Passage Time Distributions For Gaussian Markov (Ornstein-Uhlenbeck) Statistical Processes", Reprint Series R-216, Graduate School of Management, University of Rochester, Rochester, N.Y. (1975).
- [15] J.F.C. Kingman, "The Single Server Queue in Heavy Traffic", Proceedings of The Cambridge Philosophy Society 57, p.902-904 (1961).
- [16] J.F.C. Kingman, "The Heavy Traffic Approximation in the Theory of Queues", p.137-159, Proceedings of The Symposium on Congestion Theory, W.L. Smith and W.E. Wilkinson eds., University of North Carolina Press, Chapel Hill, N.C. (1965).
- [17] L. Kleinrock, Queueing Systems, Volume II: Computer Applications, p.62-100, Wiley, New York (1976).
- [18] H. Kobayashi, "Application Of The Diffusion Approximation To Queueing Networks, I. Equilibrium Queue Distributions", Journal of the Association for Computing Machinery 21, p.316-328 (1974).
- [19] T.C.T. Kotiah, "Approximate Transient Analysis of Some Queueing Systems", Opns. Res. 26, p.333-346 (1978).
- [20] I.L. Mitran and B. Avi-Itzhak, "A Many-Server Queue With Service Interruptions", Opns. Res. 16, p.628-638 (1968).
- [21] S.C. Moore, "Approximating The Behavior of Nonstationary Single-Server Queues", Opns. Res. 23, p.1011-1032 (1975).
- [22] J. Muckstadt, "A Model For A Multi-Item, Multi-Echelon, Multi-Indenture Inventory System", Management Science 20, p.472-481 (1973).
- [23] J. Muckstadt, "Some Approximations In Multi-Item, Multi-Echelon Inventory Systems for Recoverable Items", The RAND Corporation, P-5763 (1976).

- [24] G.F. Newell, "Approximation Methods For Queues With Application To The Fixed Cycle Traffic Light", SIAM Review 7, p.223-239 (1965).
- [25] G.F. Newell, "Queues With Time Dependent Arrival Rates.I: The Transition Through Saturation", J. of Applied Prob. 5, p.436-451 (1968).
- [26] G.F. Newell, "Queues With Time Dependent Arrival Rates.II: The Maximum Queue", J. of Applied Prob. 5, p.579-590 (1968).
- [27] G.F. Newell, Applications of Queueing Theory, p.101-118, Chapman and Hall, London (1971).
- [28] E. Parzen, Stochastic Processes, Holden Day, New York (1962).
- [29] R.L. Pokress, "The $M(t)/M/\infty$ Queue As An Approximation To The $M(t)/M/r$ Queue", Unpublished Research Notes, The RAND Corporation, (1977).
- [30] S. Ross, "Average Delay In Queues With Nonstationary Poisson Arrivals", Technical Report ORC 77-13, Operations Research Center, University of California, Berkeley, Calif. (1977).
- [31] T.L. Saaty, "Time Dependent Solution of The Many-Server Poisson Queue", Opns. Res. 8, p.755-772 (1960).
- [32] C.C. Sherbrooke, "METRIC: A Multi-Echelon Technique for Recoverable Item Control", The RAND Corporation, RM-5078-PR (1966).
- [33] A.W. Shogan, "A Queueing System Subject To Breakdown and Having Non-Stationary Poisson Arrivals", Technical Report 188, Department of Operations Research, Stanford University, Stanford, Calif. (1977).
- [34] H. White and L.S. Christie, "Queueing With Preemptive Priorities Or With Breakdown", Opns. Res. 6, p 79-95(1958).
- [35] U. Yechiali, "A Queueing-Type Birth and Death Process Defined On A Continuous-Time Markov Chain", Opns. Res. 21, p.604-609 (1973).
- [36] U. Yechiali and P. Naor, "Queueing Problems With Heterogeneous Arrivals and Service". Opns. Res. 19, p. 722-734 (1971).

END

DATE
FILMED

1-82

DTIC