An Investigation of the Transient Behavior of Stationary Queueing Systems

by

Emily Jane Roth

S.B., Massachusetts Institute of Technology

(1977)

SUBMITTED TO THE SLOAN SCHOOL
OF MANAGEMENT IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1981

Signature of Author__________________________________
Sloan School of Management
March 11, 1981

Certified by ____________________________________
Amedeo R. Odoni
Thesis Supervisor

Accepted by____________________________________
Chair, Departmental Graduate Committee

An Investigation of the Transient Behavior of Stationary Queueing Systems

by

Emily Jane Roth

Submitted to the Sloan School of Management on March 11, 1981
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Operations Research

ABSTRACT

In this dissertation, we develop a collection of techniques and approximations that provide information about the transient behavior of many stationary queueing systems for which tractable theoretical transient solutions are not available. The primary contribution is a set of bounds to estimate the effective duration of the transient period of the expected queue length.

We describe and extend a method due to Koopman for obtaining both exact numerical solutions to stationary Markovian queueing systems and approximate solutions to partially deterministic systems in which the embedded chain is a first-order Markov process. A truncated set of state equations is solved numerically yielding the state probabilities as a function of time. In addition to k-server, single-queue systems, multiple-queue systems with priority schemes can be solved.

Solutions generated through use of this numerical technique are used to empirically demonstrate that for ergodic, infinite-capacity, single-queue, single-server systems, the expected queue length decays in an approximately exponential manner for large t. We suggest a closed-form expression for estimating the time constant of the exponential function used to approximate this behavior. In addition, empirical results indicate four categories of initial behavior of the expected queue length as a function of the initial state of the system. For each category, an upper bound is determined empirically for the amount of time required for the transient effects of the initial conditions to become negligible.

Finally, we propose a technique for approximating the transient expected queue length of ergodic, infinite-capacity, single-queue, single-server systems that begin at rest. Based on the above observation that the expected queue length can, in many systems, be approximated by an exponential function, the exact numerical transient solution of an M/M/1 queueing system is scaled, using simple arithmetic operations, yielding approximate solutions for more complex systems. Comparison with numerical solutions indicates that, except for small values of t, the accuracy of this approximation is good, and that solution costs will typically be significantly lower than those for numerical solution of the original system.

Thesis Supervisor: Amedeo R. Odoni

Title: Professor of Aeronautics and Astronautics

ACKNOWLEDGEMENTS

I am extremely indebted to Professor Amedeo Odoni for his guidance and encouragement and for his constructive comments on many drafts of this dissertation. I am very fortunate to have such an exceptional advisor.

I also owe thanks to Professors Al Drake, Joe Ferreira, and Steve Graves for serving on my committee. Their comments were essential in my struggle to develop a sense of perspective.

I would like to express my appreciation to Dean Toby Davis of the School of Urban and Public Affairs at Carnegie-Mellon University for adjusting my responsibilities in order to allow me additional time in a superb working environment to complete this thesis. I am also grateful to Mary Joyce Airgood of SUPA for her hours of skillful typing.

A special thanks is due to my father for providing a role model, and to both of my parents for encouraging a sense of self-reliance.

This dissertation would not have been completed without the support of my friends. I thank Neil Buttimer, Bruce Lamar, and Peg Hume for being my family at Ware Street, and Dr. Elisabeth Drake for her companionship in the pursuit of our common obsessions. Each of these people was responsible for easing me over many mental hurdles.

I thank my friend and colleague Professor Denise DiPasquale of SUPA for helping me through the final stretch. In addition to the squash games and MGB rides, Denise was immeasurably helpful in maintaining the proper atmosphere of humor and cynicism.

Finally, two special acknowledgements. To Al for the racketball and video game era; also, for introducing me to Fungleton and to Oscar and the dog--I haven't been the same since. And to Vicki Bier, friend and intellectual sparring partner, whose invaluable editorial advise caused several hundred commas to be added to the weight of this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

"...as a practical decision aid, there appears to be little value in queueing theory models."
Jack Byrd Jr., "The Value of Queueing Theory."

"The fact that some of the queueing literature is arcane and useless is irrelevant, for much of the literature is exactly what we practitioners need."
Peter Kolesar, "A Quick and Dirty Response to the Quick and Dirty Crowd; Particularly to Jack Byrd's 'The Value of Queueing Theory.'"

Most of the extensive research on queueing systems to date has been geared primarily toward determination of exact results for steady-state conditions. These results are frequently of questionable use to the practitioner , either because of their complexity or due to the fact that they are derived under special sets of assumptions which are not often satisfied in practice. In fact, the usefulness of queueing theory as a whole has been the subject of rather heated debate in the literature (e.g., [3, 4, 24, 40]).

One important example in which existing theoretical results are inadequate is the situation where behavior of a queueing system before it reaches steady-state is of particular interest. In this work, we focus on the development of useful techniques and approximations for this situation--i.e., we are concerned with the transient behavior of stationary queueing systems.

Due to severe difficulties encountered in attempting to derive exact, closed-form solutions, there exist virtually no useful general results for time-varying behavior of queueing systems. Even the simple case of a stationary, infinite—capacity M/M/1 queueing system falls into this category. The exact transient solution of this system includes an infinite sum of weighted modified Bessel functions; it is seldom practical to evaluate this type of expression even using numerical techniques [18].

Since theoretical approaches have met with little success in analyzing transient behavior of queueing systems, we have chosen here to approach the problem almost entirely in an empirical manner. Throughout, our intent is to develop results and insights which will be of use in applications. Rather than considering specific applications in depth, we focus instead on the development of methods general enough to be of use in a broad class of applications. We illustrate such potential applications through the following two examples.

Consider a production line which is subject to machine failures. After each failure, the production process is halted until the machine is repaired. The firm might be interested in assessing the cost of these failures: for example, such information would be useful when deciding on the number of repairpersons that should be available.

The time period influencing the cost of any machine failure can be separated into two distinct segments; the repair time during which production is at a complete halt, and the start-up period during which the production line returns to full operating capacity. The cost due to the former will depend on system characteristics such as the availability of repair personnel and the length of the actual repair time. This thesis, however, is more concerned with the second time segment, the start-up period. Quite often these start-up periods are assumed to have a negligible effect on the overall production. Under this assumption, equilibrium analysis will be sufficient to determine production levels and,in many cases, existing closed-form results can then be used.

Whether or not this approach is valid is clearly dependent on the accuracy of the assumption that the system is, for all practical purposes, in steady-state at all times. We will suggest here a method for testing the validity of this assumption.

Our work can also be useful in the solution of certain nonstationary systems. We illustrate with the following application pertaining to airport runway operations. Airport planners frequently use steady-state queueing models to estimate the delays experienced by arriving and departing aircraft. The demand profile for runway use at a major airport typically exhibits considerable variation over time, with peaks during morning and evening rush hours and little demand during the late-night/early-morning period. The standard approach to estimating airport delays has been to approximate the time-varying demand profile by a piecewise constant curve[1] and to use theoretical steady-state results to calculate the expected delay during each time period [39]. These values provide an estimate of the expected delay to an aircraft arriving or departing during each time period.

This type of analysis once again relies heavily on the assumption that transient effects are of negligible importance. However, with approximations introduced in this dissertation, we will show that in many typical airport situations the transient period is in fact long enough to cause the analysis standardly used to be invalid.

[1] Time points t_o, t_1, ... are chosen. Then, the demand profile $\lambda(t)$ is estimated by $\lambda(t) = \lambda_i$, i=1,2,... where

$$\lambda_i = \frac{\lambda(t_i) - \lambda(t_{i-1})}{2} ,$$

i.e., the arrival rate is constant during each time period $(t_i - t_{i-1}]$, i=1,2,...

In this work we consider two types of systems: "Markovian" systems, which will be defined as queueing systems that have a first-order Markov process representation, and "partially deterministic" systems in which either the interarrival or service time distribution is deterministic and the embedded chain is a first-order Markov process.

We restrict our work to these two classes as they are the only types of systems for which we can obtain accurate numerical transient solutions on which to base our empirical analysis. This will be discussed in more detail in Chapter 2.

We consider only stationary processes, i.e., processes for which the arrival and service rates are independent of time during the period of interest. Systems are further restricted to be ergodic. Clearly, any system which is not ergodic (e.g., an infinite-capacity system with traffic intensity greater than 1) will never attain an equilibrium condition. The expected number of customers in queue (not including those in service) at time t, and the expected delay at time t, are measures which are frequently of interest in applications. This analysis focuses on the former measure; comparable results can be derived for the expected delay in an analogous manner.

Our strategy is to examine and compare the transient behavior of the expected queue length of many queueing systems in an attempt to characterize similarities in their responses. In Chapter 2, we address the problem of obtaining the transient solutions for stationary queueing systems which will be needed for our empirical work. We begin with a survey of many of the solution techniques developed in the literature. We want to observe the dominant features of the behavior of the expected queue length over time in order to develop a reasonably good approximation to its actual functional form. Therefore, accuracy is a particularly important attribute to consider in our evaluation of existing solution techniques. Also, we will compare many different queueing systems; thus, the technique must have the flexibility to handle various levels of traffic intensity, initial conditions, and a range of interarrival and service time distributions. With these considerations in mind, a particular numerical solution technique is selected with which to solve the class of systems examined in this dissertation. This technique is described in detail in Section 2.2.

Chapter 3 begins with a brief examination of the few available theoretical transient solutions. A particular form is postulated for $Q(t)$, the expected queue length as a function of time. In Section 3.1, we present empirical confirmation of this hypothesis for infinite-capacity, single-queue, single-server systems, considering a range of levels of traffic intensity and several specific forms of interarrival and service time distributions. In Section 3.2, the amount of time until the transient effects become negligible is examined. Specifically, a closed-form approximation is obtained which can be used to estimate the amount of time until the transients of $Q(t)$ have been reduced by $\eta\%$ (for any $0 \leq \eta < 100$).

In Chapter 4, we consider the effect of initial conditions on the decay of transients. In particular, we parallel the analysis in Chapter 3, first making general comments on the form of $Q(t)$, and then developing a means to estimate the amount of time required until the transient effects are negligible.

In Chapter 5, we present a new approximate solution technique based on the results of Chapter 3. For a rather broad class of stationary queueing systems which begin at rest, our results suggest that this method can be used to determine transient solutions to an accuracy close to that of numerical solution techniques but at a significantly lower computation cost.

Finally, in Chapter 6 we summarize the work and indicate several directions for further research.

CHAPTER 2

SOLUTION TECHNIQUES FOR STATIONARY QUEUEING SYSTEMS

The main purpose of this work is to investigate the manner in which ergodic, stationary queueing systems approach steady-state. In particular, we seek a closed-form expression for estimating the amount of time required for the transient effects to become negligible. In order to accomplish this goal, it is necessary to examine transient solutions to many types of queueing systems.

In Section 2.1, various techniques for determining transient response of stationary queueing systems are summarized. These techniques may be grouped into four categories:

(i) exact, closed-form solutions

(ii) simulation

(iii) approximations

(iv) numerical techniques

As a result of this discussion, we argue that, given the nature of the problem being examined, a particular numerical technique is preferred. In Section 2.2, a thorough description of this technique is provided.

2.1 Review of Alternative Approaches

For our purposes, a solution technique is needed which will allow us to compare the manner in which different queueing systems approach equilibrium. In particular, we must have the ability to trace the expected queue length (our representative measure of system behavior) from time t = 0 until the system has reached equilibrium. In addition to being sufficiently flexible to solve many types of finite- and infinite-capacity stationary queueing systems at moderate cost, the solution technique must be capable of handling varying initial conditions and the entire range of traffic intensities ($0 < \rho < 1$).

2.1.1 Exact, Closed-Form Solutions

Most early research on transient behavior of stationary queueing systems was geared towards finding exact solutions to the equations representing the evolution of the system over time. Given that these equations only rarely yield closed-form solutions, this approach is of little use unless combined with more recent, computer-oriented methods (see Section 2.1.4).

Several examples which illustrate the complexity of transient solutions for even the simplest queueing systems are provided in Table 2.1. In each case, the expression for $P_i(t)$, the probability that there are i customers ($i = 0,1,...$) in the system at time t is listed.

Note that, in general, the expressions in the table are so complex that their usefulness in applications is questionable. For example, note the infinite sum containing Bessel functions in the expression for $P_i(t)$ for an infinite-capacity M/M/1 system. These Bessel functions must be recomputed for each new value of t--a process which is time consuming. Also, as there is no closed-form expression for $I_n(y)$, numerical errors are introduced.

In view of the fact that the queueing systems covered by the expressions listed in Table 2.1 are among the simplest known, one can justifiably be pessimistic about the likelihood that future attempts to obtain closed-form, exact solutions for the transients of more complex systems will be successful. Except for the M/M/∞ case, available closed-form results will not yield useful solutions to our problem.

Table 2.1: Transient Solutions to Some Stationary Queueing Systems

System	Capacity	Reference	$P_i(t)$	Notes
M/M/1*	N	[31]	$\frac{\rho^i(1-\rho)}{1-\rho^{N+1}} + \rho^{i/2} \sum_{k=1}^{N} C_k \left[\sin \frac{ik\pi}{N+1} - \sqrt{\rho} \sin \frac{(i+1)k\pi}{N+1}\right] e^{-[\lambda-\mu-2\sqrt{\lambda\mu} \cos (\frac{k\pi}{N+1})]t}$	C_k depends on initial conditions
M/M/1	∞	[12]	$e^{-(\lambda+\mu)t}\left[\rho^{(i-s)/2} I_{i-s}(2\sqrt{\lambda\mu}t)+\rho^{(i-s-1)/2} I_{i+s+1}(2\sqrt{\lambda\mu}t)+(1-\rho)\rho^i \sum_{k=i+s+2}^{\infty} \rho^{-k/2} I_K(2\sqrt{\lambda\mu t})\right]$	s customers in system at t=0 $I_n(y) = \sum_{k=0}^{\infty} \frac{(\frac{y}{2})^{n+2k}}{k!(n+k)!}$, n>-1
M/M/∞	∞	[12]	$\frac{1}{i!} [\rho(1-e^{-\mu t})]^i e^{-\rho(1-e^{-\mu t})}$	Initially empty system
M/G/∞	∞	[12]	$\frac{1}{i!} [\lambda qt]^i e^{-\lambda qt}$	Initially empty system $q = \frac{\int_o^t P(s>s_o)ds_o}{t}$

*Keilson [14], has derived a similar result for ergodic, finite-capacity queueing systems which are "time-reversible", i.e., $a_{ij}P_i(\infty) = a_{ji}P_j(\infty)$ where a_{ij} is the $(i,j)^{th}$ element of the one-step transition matrix of the Markov chain representing the queueing system. This class of systems includes the finite-capacity M/M/k queueing system.

2.1.2 Simulation

Simulation of a queueing system involves sampling from the distribution of all possible sequences of events in the system. Given an infinite number of independent sample paths from this distribution, one can determine actual system behavior (e.g., the expected queue length, Q(t)). In principle then, simulation can be used to determine Q(t) for virtually any queueing system. A major problem with the technique is that in practice, one must use only a finite number of these sample paths and thus obtain, at best, point estimates and confidence intervals for Q(t).[1]

Simulation can be useful in estimating steady-state statistics, such as Q(∞), in systems for which exact expressions do not exist or are intractable. An important consideration here, is that the initial portion of each sample path is affected by the initial conditions of the simulation run. Wilson and Pritsker [43,44] survey various policies for selecting initial conditions for the purpose of minimizing the duration of this transient period and identifying the "truncation point", the time after which the simulation can be considered to be of a system in steady-state.

One way to lessen the importance of the start-up policy in the estimation of Q(∞) is to use the results from one long simulation run rather than aggregating those of several shorter replications (each having a start-up period). If the initial portion of a single long run is discarded (to remove the transients) and the remainder is divided into many segments, the mean queue lengths of all segments may be averaged to provide an estimate of Q(∞).[2]

[1] A good general discussion of techniques for simulating queueing systems is provided in the two-volume text by Kleijner [19,20].

[2] Law and Carson [27] discuss and compare several methods to obtain confidence intervals for the mean of a stochastic process. They include several examples of queueing systems for which they estimate the steady-state expected delay.

Care must be used in choosing the length of these segments. Ideally, we want to aggregate many independent sample paths. Since the segments are all from the same simulation run, there will be correlation between points at the end of one segment and those at the start of the next segment. Therefore, the segments must be long to reduce this correlation.

In this work, our interest is to determine the functional form of Q(t), not Q(∞). For this purpose, we must have very close estimates of Q(t) for all values of t. In simulation, this corresponds to achieving a very narrow confidence band for Q(t) for each t. As the number of computations increases (roughly) with the square of the desired precision[25], obtaining these narrow confidence bands leads to prohibitive costs. In addition, it has been shown by Daley [6] that the variance of the queue length for a GI/M/1 system increases as $1/(1-\rho)^2$.[3] Thus, as $\rho \to 1$, more simulation runs are needed to achieve the same level of accuracy.

Therefore, while simulation may be useful in many situations (e.g., calculating steady-state measures of system behavior), for our purposes other methods may be preferable.

2.1.3 Approximations

The solution techniques in this category cover a wide range and are geared toward obtaining estimates of one or more aspects of queueing system behavior through use of simplifying assumptions. These assumptions

[3] Although it has not been proven, it is likely that a similar type of relationship holds for more general queueing systems, e.g., GI/G/1. This conjecture is based on the result that the mean queue length, Q(∞), is proportional to $1/(1-\rho)$ for queueing systems ranging from M/M/1 to GI/G/k (see Kleinrock [21]).

may be based on prior knowledge or intuitive notions of how the queueing system works: in some cases, simplifying assumptions are made primarily for purposes of mathematical convenience. Frequently, no bounds can be determined to measure the extent of error introduced by these assumptions. In such cases, the validity of the approximate technique would have to be reevaluated for each new system by comparing the approximate solution with results provided by an alternative method (typically simulation).

Most of the methods cited in this section were actually developed to solve nonstationary systems. As time-varying parameters frequently cause the system to be continuously in the transient state, such solution techniques can certainly be used to determine the approximate transient behavior of stationary queueing systems.

The following two methods, due to Moore and Newell, apply to rather general systems. Each simplifies the solution procedure by removing some of the randomness present in the original system.

Moore [30] proposes an approximate technique for solving a finite-capacity $M^X/G/1$ queueing system in which the single server processes each customer according to a service time distribution which is given by a probabilistic choice among Erlang random variables. As a direct function of time, the $M^X/G/1$ system is not Markovian since the number of customers in the system at any time is specified by the number at the instant before, *and* the length of time the current customer has been in service. This is due to the fact that the service time distribution has memory--i.e., the amount of time remaining in a particular service interval depends on how long ago the service began. Moore first eliminates the memory in the system by examining it only at the instant following a service completion (the start of an epoch). He then recursively solves the embedded chain equations to determine the state probabilities at these points in time, and then approximates the length of an epoch by the expected service time.

Thus, through use of this procedure, one can compute exact, numerical estimates of the state probabilities at the start of each epoch as well as obtain an approximate idea of where these epochs occur along the time axis. From the state probabilities it is then a simple matter to obtain values for such measures as the expected number in queue.

There are two errors which arise during use of this method. The first is due to numerical solution of the embedded chain equations and should be negligible.[4] The second is caused by the approximation of the length of each epoch by its expected value--in a sense, by approximating the service time as a deterministic random variable. Moore does not determine a method to bound the magnitude of this error except by comparison to simulation results. Thus, the accuracy of the solution for each new system must be examined by comparison to results from an alternative solution technique.

For queueing systems which are heavily congested, (i.e., utilization factor near or greater than 1), past results (e.g., Kingman [15,16]) suggest that the specific forms of the interarrival and service time distributions do not heavily influence system behavior. Work by Newell [33,34] uses this notion and the idea that, in a system under heavy traffic, it is permissible to approximate the behavior of the discrete-state queueing system by that of a continuous-state diffusion process. A

[4] The magnitude of this type of error will be discussed further in Section 2.2.4.

diffusion equation can then be written and solved to determine the expected queue length or the expected delay as a function of time. (We study the diffusion equation in more detail in Chapter 3.) Gaver [8] and Kobayashi [23] discuss use of this diffusion approximation for transient analysis of queueing systems.

This solution procedure has promise for many applications, but the two assumptions that form the basis for this method are not valid when traffic intensity is low. Thus, it is not particularly suitable for our purposes.

Several more recent investigations deal with systems which are described completely by an infinite set of Chapman-Kolmogorov equations. In all of these cases, this infinite set of differential equations is reduced, by means of a "closure assumption", to a finite set which is then solved. These closure assumptions are often motivated more by mathematical convenience than by intuitive arguments.

Among researchers following this approach, Rider [35] combines the state equations for an infinite-capacity M/M/1 queueing system into a single differential equation for the expected queue length. Since this equation is dependent on the probability of an empty system, it is necessary, for solution, to relate the idle probability to the expected queue length. This is achieved by means of a closure assumption.

A similar closure technique has been introduced by Rothkopf and Oren [37]. In their method both the mean and variance of the queue length are related to the probability of an empty system by means of a closure assumption (different from that of Rider) in order to obtain an approximate solution for an M/M/k system. Rothkopf and Oren also report that Chang [5] uses a similar method to analyze networks of M/M/1 queues and that Wang [41] is trying to extend Chang's method to handle more general queueing systems.

As mentioned above, the closure assumptions for the above methods seem to be chosen more for computational reasons rather than to exploit known attributes of the systems. For instance, Rothkopf and Oren [37] assume that the state probabilities have a negative binomial distribution. This leads to an efficient solution technique but one whose accuracy must be verified by comparison to simulation.

In applications, a mathematical model is typically only a rough representation of the actual system. In these instances, additional error incurred through use of an approximate technique to solve this model may not be important due to the approximate nature of the model itself. For our purposes, the mathematical models are assumed to be exact--i.e., the only error is due to the solution of a given mathematical model; therefore, we seek a solution technique which will accurately determine the behavior of the model. Since most approximate solution techniques do not allow us to bound this error, for our purposes these techniques are not particularly attractive.

2.1.4 Numerical Techniques

Unlike the approximate techniques, exact numerical solutions require no assumptions once the particular queueing system to be solved has been specified and the equations describing system behavior have been written. These equations are then solved to any desired degree of accuracy using an appropriate numerical method.

When numerical solution is feasible, it is often significantly less expensive than simulation. In addition, the class of systems for which such numerical solution is feasible will certainly become progressively larger with the development of more sophisticated computer hardware and software.

All of the numerical techniques we consider[5] produce the state probabilities, $P_i(t)$--the probability that there are i customers in the system at time t (i=0,1,...). From these probabilities the expected queue length can be determined.

The first approach we consider was developed by Kotiah [26] and is different from the other techniques presented in this section. Kotiah uses numerical techniques to invert the transforms of the desired probabilities. This method applies to systems for which obtaining closed-form expressions for the exact transforms of the number of customers in the system and the expected queue length require finding a root to a polynomial. (Examples include infinite-capacity M/M/1 and $M/E_k/1$ systems.) Kotiah compares three numerical techniques to determine a sequence of rational approximations to this root--the method of successive approximations, Newton's method, and the series method. Given an approximate root, the transforms are inverted yielding $P_i(t)$, i=0,1,..., and Q(t).

Comparison of exact values of $P_0(t)$, $P_1(t)$ and Q(t) for an M/M/1 queueing system with those obtained through use of Kotiah's technique suggest that Newton's method will yield good approximations for $P_i(t)$, i=0,1,..., for small t. It is not clear that the approximations will remain close as t increases. In particular, it is not apparent that accuracy will remain good until steady-state is reached. In addition, while the method can, in principle, be used for Erlangian systems, computational complexity would be greatly increased.

[5] As in the methods discussed in the previous section, most of these techniques were developed to solve nonstationary queueing systems.

The remaining solution approaches in this category are similar in that each relies upon a numerical method to directly solve the set of state equations. In order to solve this set through use of a computer, there must be a finite number of state equations or equivalently, system capacity must be finite. Frequently, this is not restrictive as in most cases it is possible to choose a number N, such that the probability of the system having N or more customers is negligible. Thus solving the system with a finite capacity N is effectively the same as solving the corresponding infinite-capacity system.

Neuts [32] developed a numerical technique which can determine state probabilities of finite-capacity GI/G/1 systems. By approximating the random variables describing the interarrival and service times as bounded, discrete random variables, the state probabilities of the resulting discrete-state, discrete-time Markov chain can be obtained recursively. These calculations are very complex unless the system is small. Also, the method is intended only for determining short-range behavior in systems which have interarrival and service time distributions that are concentrated about one value. Since we require knowledge of Q(t) from time t=0 until the system reaches steady-state, this solution technique is not sufficiently flexible for our purposes.

Finally, we consider two numerical techniques which can be used to solve systems in which the state equations are a finite set of simultaneous, first order-differential equations. In each case, these systems are solved numerically yielding the state probabilities as a function of time.

The first of these techniques is used in an important paper by Koopman [25] to solve finite-capacity M/M/1 queueing systems.[6] Koopman employs a standard Runge-Kutta numerical method to solve the state equations. Hengsbach and Odoni [13] extended this solution technique to handle systems with k servers and greatly improved the efficiency of the computer programs.[7]

The second approach to solving finite-capacity systems defined by a set of Chapman-Kolmogorov equations is presented in the work of Grassmann [9,10,11]. Many standard numerical techniques (including Runge-Kutta) rely on the calculation of powers of the transition matrix (infinitesimal generator) of the queueing system. Computationally, the presence of both positive and negative elements in this matrix can lead to high round-off errors. Grassmann, using a technique he calls "randomization", eliminates this source of error by introducing a new matrix which has only positive elements. By expressing the original solution in terms of this new matrix, it is no longer necessary to calculate powers of the infinitesimal generator.

Either of these last two techniques can be used to determine the transient behavior (e.g., Q(t)) for the same class of queueing systems.

[6] Finite capacity M/D/1 systems are solved in a similar way--the embedded chain equations are solved recursively yielding the state probabilities at the start of each epoch. In this case this is strictly an approximate technique as it is necessary to add the assumption that all arrivals and service completions occur at the instant before the start of an epoch. This will be discussed further in Section 2.2.1.

[7] It is interesting to note that in a 1965 paper, Leese and Boyd [28] had dismissed this technique as numerically intractable. Extensive improvements in computer hardware and software since 1966 have, in fact, permitted tractable numerical solution.

They will handle any level of traffic intensity and any specified set of initial conditions. In each case, the error due to the numerical routine can be made arbitrarily small. If the capacity of the system is large, randomization is preferred as storage requirements are not as great.[8] Grassmann [10] shows that if solutions are not needed at frequent time points, randomization is less expensive than Runge-Kutta. Also, he states that accuracy is improved. On the other hand, if solutions are required at a large number of time points, Runge-Kutta would probably be more efficient.

We have elected to use the method developed by Koopman for our study. In Section 2.2, this technique is described in detail and extended to handle any system in which behavior can be described by a set of Chapman-Kolmogorov equations.

2.2 Description of a Numerical Solution Technique

In this section, we provide a detailed description of the numerical solution technique which will be used in Chapter 3 for our empirical study of the manner in which stationary, ergodic queueing systems approach equilibrium. As mentioned in Section 2.1.4, this solution technique was first used successfully by Koopman. Recall that the basic strategy is to:

(i) set up the state equations for the particular system under consideration, and

(ii) solve this set of equations numerically to obtain the state probabilities as a function of time.

[8] Grassmann [10] has solved systems with over 10,000 states.

Given the particular problem addressed in this thesis, this discussion will be limited to stationary queueing systems. However, it should be noted that this solution technique may be applied to many non-stationary queueing systems. (See Appendix 1 for details.)

First, we apply this numerical solution technique to single-queue systems. (This material will be used in Chapters 3-5.) Then, in Section 2.2.2, an extension to multiple-queue systems under priority schemes is illustrated. In Section 2.2.3 we summarize the ways in which error is introduced through use of this solution technique. Finally, in Section 2.2.4 we discuss computational characteristics.

2.2.1 Single-Queue Systems

2.2.1.1 Markovian Systems

This section is an examination of a technique for solving a variety of finite-capacity queueing systems which can be characterized as "Markovian." We define a Markovian queueing system as a system which at all times has a discrete-state description in which behavior is according to a first-order Markov process.[9] (Supplementary variables may be needed to represent a higher-order process as a first-order Markov chain.) This class includes $M/M/k$, $M^X/M/1$, $M/M^X/1$, $M/H_k/1$, and $E_k/E_k/1$ queueing systems. In all cases, the behavior of the queueing system can be described through the Chapman-Kolmogorov equations, the set of simultaneous, first-order differential equations that describe the rate of transitions for each network state.[10] Recall that closed-form solutions have been obtained for

[9] This definition excludes systems in which only the embedded chain is first-order (e.g., M/G/1).

[10] For details on deriving the state equations, see [12] or [21].

only very special cases. Given the arrival and service rates and any specified set of initial conditions, these system equations can be solved using standard numerical techniques that will be discussed in Section 2.2.4. Solving the Chapman-Kolmogorov equations yields the state probabilities as a function of time. From the state probabilities one can easily determine many characteristics of system behavior.

As an example, consider a finite-capacity, M/M/k queueing system in which the interarrival and service time distributions are independent and exponential with means $1/\lambda$ and $1/\mu$, respectively. A maximum of N customers are allowed in the system at any time. For this queueing system, a convenient state descriptor is the number of customers in the system (in queue plus in service). The state transition diagram is shown in Figure 2.1. By defining

$$P_i(t) = P(i \text{ customers in the system at time } t),$$

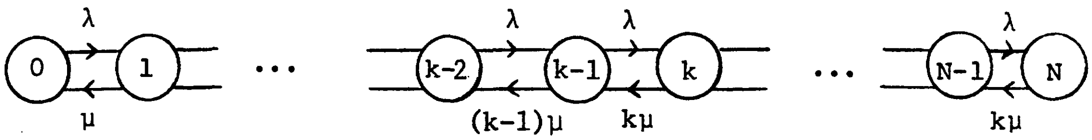
$i = 0,1, \ldots N$, the Chapman-Kolmogorov equations can be derived directly from the state transition diagram. They also appear in Figure 2.1.

For any given set of initial conditions $P_i(0)$, $i=0,1,\ldots,N$, these N+1 simultaneous differential equations can then be solved numerically for the state probabilities $P_i(t)$, $i = 0,1,\ldots,N$, as a function of time. Examples of the time-dependent characteristics of system behavior obtainable from the state probabilities include:

(i) the expected number of customers in the system at time T,

$$L(T) = \sum_{i=1}^{N} iP_i(T) \ , \tag{2.1}$$

State Transition Diagram



state i: i customers in the system

Chapman-Kolmogorov Equations

$$\dot{P}_0(t) = -\lambda P_0(t) + \mu P_1(t)$$

$$\dot{P}_i(t) = -(\lambda + i\mu)\ P_i(t) + \lambda P_{i-1}(t) + (i+1)\mu P_{i+1}(t)$$

$$i = 1,2,\ldots,k-1$$

$$\dot{P}_i(t) = -(\lambda + k\mu)\ P_i(t) + \lambda P_{i-1}(t) + k\mu P_{i+1}(t)$$

$$i = k,k+1,\ldots,N-1$$

$$\dot{P}_N(t) = -k\mu\ P_N(t) + \lambda P_{N-1}(t)$$

Figure 2.1: State Transition Diagram and Chapman-Kolmogorov Equations for a Finite-Capacity M/M/k Queueing System

(ii) the expected number of customers in the queue at time T,

$$Q(T) = \sum_{i=1}^{N} (i-1)P_i(T)$$

$$= L(t) - [1-P_0(t)], \qquad (2.2)$$

and

(iii) the expected delay at time T,

$$W(T) = \frac{1}{\mu} \sum_{i=2}^{N-1} (i-1)P_i(T) \qquad (2.3)$$

where the expected delay at time T is the average amount of time a customer would have to wait if she were to enter the system at time T.

2.2.1.2 Partially Deterministic Systems

We next consider a class of queueing systems in which either the interarrival or service time distribution is deterministic. A further requirement is that the embedded chain have a first-order Markov process representation. Queueing systems in this category include M/D/k, D/M/k, E_k/D/1, D/E_k/1, and bulk arrival/bulk service systems. The basic solution technique is the same: derive the system equations and solve them numerically to obtain the state probabilities as a function of time. Unfortunately, for these systems the state transition equations represent only the embedded chain behavior--deriving these equations requires the assumption that all arrivals and service completions occur either the instant before or after the start of an epoch.

As an example, we consider a finite-capacity, M/D/k queueing system. We define epochs to begin at $t = t_0, t_1, \ldots$, where $(t_{i+1} - t_i)$, $i = 0,1,\ldots$, is the length of the deterministic service time and $t_0 = 0$ is the specified initial starting time. With the assumption that all service completions and customer arrivals occur only at the instant before the start of an epoch (at times t $= t_0^-$, t_1^-, ...), we can derive the state transition equations for the embedded chain. This set of simultaneous difference equations is shown in Figure 2.2. Given arrival rate λ, service rate μ, and a set of initial conditions $P_i(0)$, i = 0,1,...,N, the (N+1) equations can be solved recursively to determine the state probabilities at times $t = t_0, t_1, \ldots$. As with Markovian systems, we can obtain many of the desired measures of system behavior from these state probabilities.

The added assumption that the state of the system changes only at the end of an epoch leads to an implicit error in these models of partially deterministic systems. Consider, for instance, an M/D/k system in which each service time lasts exactly one minute. Our model essentially assumes that service completions and customer arrivals can occur only at the instant before the start of an epoch. This implies that if the system empties at time t=4 and the next customer arrival actually occurs at time t=4.5, this customer, according to our model, will not begin service until time t=5. In reality, however, service would have begun at t=4.5. Since the error is introduced only when the system is in the empty state, this difference becomes negligible as the traffic intensity, ρ, approaches 1.

2.2.2 Multiple-Queue Systems Under Priorities

In the previous section we described a numerical solution technique for a wide variety of simple single-queue systems. It can also be used to solve multiple-queue versions of many of these systems provided system capacity is small.

$$P_0(t+1) = e^{-\rho} \sum_{j=0}^{k} P_j(t)$$

$$P_i(t+1) = e^{-\rho} \left[\frac{\rho^i}{i!} \sum_{j=0}^{k} P_j(t) + \frac{\rho^{i-1}}{(i-1)!} P_{k+1}(t) + \frac{\rho^{i-2}}{(i-2)!} P_{k+2}(t) + \ldots + P_{k+i}(t) \right] \qquad 1 \le i \le N-k$$

$$P_i(t+1) = e^{-\rho} \left[\frac{\rho^i}{i!} \sum_{j=0}^{k} P_j(t) + \frac{\rho^{i-1}}{(i-1)!} P_{k+1}(t) + \frac{\rho^{i-2}}{(i-2)!} P_{k+2}(t) + \ldots + \frac{\rho^{i+k-N}}{(i+k-N)!} P_N(t) \right] \qquad N-k+1 \le i \le N-1$$

$$P_N(t+1) = \sum_{\ell=N}^{\infty} \frac{\rho^\ell}{\ell!} \sum_{j=0}^{k} P_j(t) + \sum_{\ell=N-1}^{\infty} \frac{\rho^\ell}{\ell!} P_{k+1}(t) + \sum_{\ell=N-2}^{\infty} \frac{\rho^\ell}{\ell!} P_{k+2}(t) + \ldots + \sum_{\ell=k}^{\infty} \frac{\rho^\ell}{\ell!} P_N(t)$$

$P_i(t)$ = P(i customers in the system at time t) $i = 0,1,\ldots,N$

maximum of N customers allowed in the system at any time, $N \ge k$.

$\rho = \lambda/\mu$

Figure 2.2: The Embedded Chain Equations for the Finite-Capacity M/D/k Queueing System

2.2.2.1 Description of the Basic Multiple-Queue System

Figure 2.3 illustrates the basic system under consideration. Customers of class c, $c = 1,2,\ldots,C$, enter the corresponding queue c according to a probabilistic process with mean interarrival time $1/\lambda_c$. Arrivals to each queue are assumed to be statistically independent. Once a customer enters a queue she maintains her position ensuring first-come, first-served (FCFS) treatment within each queue.

The service facility contains k independent identical servers, serving all types of customers in either a preemptive or nonpreemptive manner. (Preemptive service allows an incoming customer of one type to interrupt the service of another type.) The service time for a customer of class c is probabilistic with mean $1/\mu_c$.

The physical capacity of the system can be specified either by the maximum number of spaces in each queue (not including those in service) or by the total number of each type of customer allowed in the system (in queue and in service). We have chosen to use the latter definition.

Each time a server becomes idle, the type of customer to next begin service must be determined by a set of priority rules. Examples include strict priority queueing (also known as head-of-the-line priority) and alternating priority (as described in the following subsection).

2.2.2.2 Example

To illustrate use of the technique and possible priority schemes, we will discuss in detail the solution technique for a finite-capacity, two-queue ($C = 2$), nonpreemptive, M/M/1 system under four priority schemes described below. At the same time, we will make comments on the analysis of more general systems.

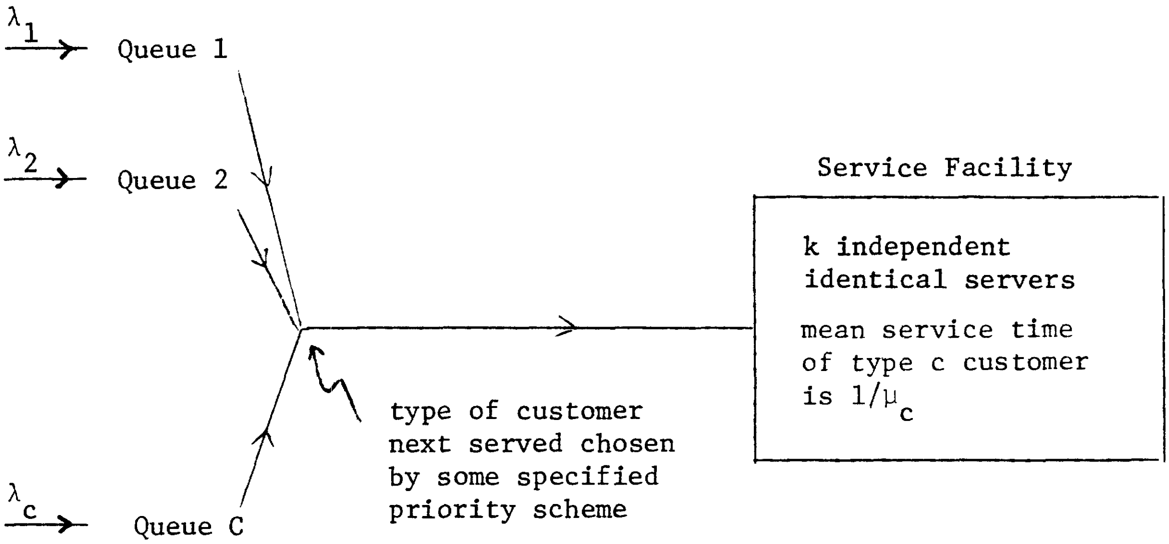

λ_1
Queue 1
λ_2
Queue 2
λ_c
Queue C
Service Facility
k independent identical servers
mean service time of type c customer is $1/\mu_c$
type of customer next served chosen by some specified priority scheme

Figure 2.3: Basic Multiple-Queue System With Priority Schemes

2.2.2.2.1 Priority Schemes

The "strict type 1" scheme specifies absolute priority for type 1 customers. Whenever the server completes a service, type 1 customers, if present, are given preference on a FCFS basis. If queue 1 is empty the first customer in queue 2 begins service. In the "alternating" scheme, as soon as a customer completes service, a customer of the other type, if there is one in the system, will begin service.

The "strict type 1/alternating threshold" scheme combines these two sets of priority rules. As long as the number of type 2 customers in the system remains at or below an arbitrary threshold, type 1 customers are given absolute priority. If the number of type 2 customers exceeds the threshold value, an alternating priority scheme is used. As soon as the number of type 2 customers in the system is at or below the threshold level, type 1 customers are once again given absolute priority.

The "strict type 1/strict type 2 threshold" scheme also gives type 1 customers absolute priority unless the number of type 2 customers in the system exceeds the threshold value. If the threshold is exceeded, type 2 customers are given absolute priority, until the number of type 2 customers is at or below the threshold level.

In summary, the schemes are outlined below:

- strict type 1: Absolute priority given to type 1 customers.
- alternating: Priority is given to the type of customer other than that which last completed service.
- strict type 1/ alternating threshold: In intervals during which the number of type 2 customers in the system exceeds the threshold value, alternating priority is used; otherwise type 1 customers have absolute priority.

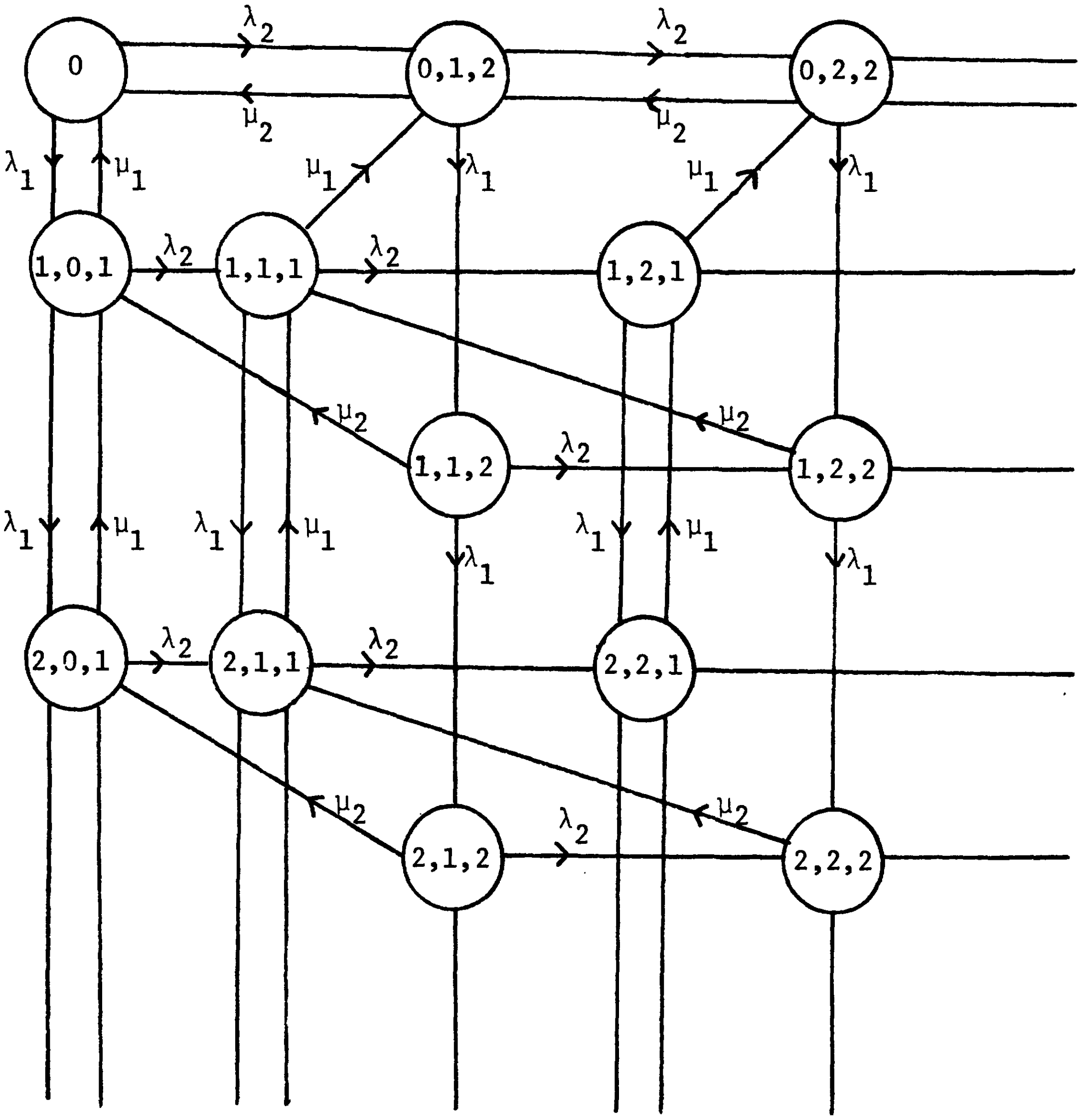
- strict type 1/ strict type 2 threshold: As in strict type 1/alternating threshold, except that, while the number of type 2 customers in the system exceeds the threshold value, type 2 customers have absolute priority.

2.2.2.2.2 State Equations

The state transition equations for this system with the listed queue disciplines are first-order, differential equations describing the rate of change of the probability of the number of each type of customer in each state of the system as a function of time. (See [36] for a description of a technique for deriving these equations.) Each of the priority schemes listed above entails solving a set of $2(N1+1)(N2+1) + 1$ of these simultaneous, first-order equations. As an example, the state transition diagram and state equations for strict type 1 priority are presented in Figures 2.4 and 2.5.

The analysis for k-server, C-queue (C > 2) M/M/1 systems is similar; the number of equations in this case will be $2^k \prod_{I=1}^{C} (NI+1) + 1$ where NI is the maximum number of type 1 customers allowed in the system (I = 1,2,...,C). In principle, this technique can be used to solve multiple-queue versions of the Markovian and partially deterministic systems discussed in the single queue case. For many systems, however, the number of required equations is large enough to be an obstacle in numerical solution; this issue will be explored in Section 2.2.4.

A straightforward extension of the basic C-queue model under priorities includes so-called "changeover costs;" service times which depend also on the type of customer previously in service. This allows more realistic modeling of systems such as many production lines where a



state (i,j,k): i type 1 and j type 2 customers in the system and a type k customer in service

Figure 2.4: Nonpreemptive, Finite-Capacity, Two-Queue M/M/1 System Under Strict Type 1 Priority--State Transition Diagram

Notation

N1 = maximum number of type 1 customers allowed in the system.

N2 = maximum number of type 2 customers allowed in the system.

R(t) = probability that the system is empty at time t.

$P_{i,j}(t)$ = probability that i type 1 and j type 2 customers are in the system at time t *and* a type 1 customer is in service.

$Q_{i,j}(t)$ = probability that i type 1 and j type 2 customers are in the system at time t *and* a type 2 customer is in service.

State Transition Equations

$$\dot{R}(t) = -(\lambda_1 + \lambda_2)R(t) + \mu_1 P_{1,0}(t) + \mu_2 Q_{0,1}(t)$$

$$\dot{P}_{1,0}(t) = -(\lambda_1 + \lambda_2 + \mu_1)P_{1,0}(t) + \lambda_1 R(t) + \mu_1 P_{2,0}(t) + \mu_2 Q_{1,1}(t)$$

$$\dot{P}_{i,0}(t) = -(\lambda_1 + \lambda_2 + \mu_1)P_{i,0}(t) + \lambda_1 P_{i-1,0}(t) + \mu_1 P_{i+1,0}(t) + \mu_2 Q_{i,1}(t) \qquad i = 2,3,\dots,N1-1$$

$$\dot{P}_{i,j}(t) = -(\lambda_1 + \lambda_2 + \mu_1)P_{i,j}(t) + \lambda_1 P_{i-1,j}(t) + \lambda_2 P_{i,j-1}(t) + \mu_1 P_{i+1,j}(t) + \mu_2 Q_{i,j+1}(t)$$
$$i = 2,3,\dots,N1-1;\ j = 1,2,\dots,N2-1$$

$$\dot{P}_{i,N2}(t) = -(\lambda_1 + \mu_1)P_{i,N2}(t) + \lambda_1 P_{i-1,N2}(t) + \lambda_2 P_{i,N2-1}(t) + \mu_1 P_{i+1,N2}(t) \qquad i = 2,3,\dots,N1-1$$

$$\dot{P}_{N1,0}(t) = -(\lambda_2 + \mu_1)P_{N1,0}(t) + \lambda_1 P_{N1-1,0}(t) + \mu_2 Q_{N1,1}(t)$$

$$\dot{P}_{1,N2}(t) = -(\lambda_1 + \mu_1)P_{1,N2}(t) + \lambda_2 P_{1,N2-1}(t) + \mu_1 P_{2,N2}(t)$$

$$\dot{P}_{1,j}(t) = -(\lambda_1 + \lambda_2 + \mu_1)P_{1,j}(t) + \lambda_2 P_{1,j-1}(t) + \mu_1 P_{2,j}(t) + \mu_2 Q_{1,j+1}(t) \qquad j = 1,2,\dots,N2-1$$

$$\dot{P}_{N1,j}(t) = -(\lambda_2 + \mu_1)P_{N1,j}(t) + \lambda_1 P_{N1-1,j}(t) + \lambda_2 P_{N1,j-1}(t) + \mu_2 Q_{N1,j+1}(t) \qquad j = 1,2,\dots,N2-1$$

$$\dot{P}_{N1,N2}(t) = -\mu_1 P_{N1,N2}(t) + \lambda_1 P_{N1-1,N2}(t) + \lambda_2 P_{N1,N2-1}(t)$$

$$\dot{P}_{0,j}(t) = 0 \qquad j = 1,2,\dots,N2$$

$$\dot{Q}_{0,1}(t) = -(\lambda_1 + \lambda_2 + \mu_2)Q_{0,1}(t) + \lambda_2 R(t) + \mu_1 P_{1,1}(t) + \mu_2 Q_{0,2}(t)$$

$$\dot{Q}_{i,1}(t) = -(\lambda_1 + \lambda_2 + \mu_2)Q_{i,1}(t) + \lambda_1 Q_{i-1,1}(t) \qquad i = 1,2,\dots,N1-1$$

$$\dot{Q}_{i,j}(t) = -(\lambda_1 + \lambda_2 + \mu_2)Q_{i,j}(t) + \lambda_1 Q_{i-1,j}(t) + \lambda_2 Q_{i,j-1}(t)$$
$$i = 1,2,\dots,N1-1;\ j = 2,3,\dots,N2-1$$

$$\dot{Q}_{i,N2}(t) = -(\lambda_1 + \mu_2)Q_{i,N2}(t) + \lambda_1 Q_{i-1,N2}(t) + \lambda_2 Q_{i,N2-1}(t) \qquad i = 1,2,\dots,N1-1$$

$$\dot{Q}_{N1,1}(t) = -(\lambda_2 + \mu_2)Q_{N1,1}(t) + \lambda_1 Q_{N1-1,1}(t)$$

$$\dot{Q}_{0,N2}(t) = -(\lambda_1 + \mu_2)Q_{0,N2}(t) + \lambda_2 Q_{0,N2-1}(t) + \mu_1 P_{1,N2}(t)$$

$$\dot{Q}_{0,j}(t) = -(\lambda_1 + \lambda_2 + \mu_2)Q_{0,j}(t) + \lambda_2 Q_{0,j-1}(t) + \mu_1 P_{1,j}(t) + \mu_2 Q_{0,j+1}(t) \qquad j = 2,3,\dots,N2-1$$

$$\dot{Q}_{N1,j}(t) = -(\lambda_2 + \mu_2)Q_{N1,j}(t) + \lambda_1 Q_{N1-1,j}(t) + \lambda_2 Q_{N1,j-1}(t) \qquad j = 2,3,\dots,N2-1$$

$$\dot{Q}_{N1,N2}(t) = -\mu_2 Q_{N1,N2}(t) + \lambda_1 Q_{N1-1,N2}(t) + \lambda_2 Q_{N1,N2-1}(t)$$

$$\dot{Q}_{i,0}(t) = 0 \qquad i = 1,2,\dots,N1$$

Figure 2.5: Nonpreemptive, Finite-Capacity, Two-Queue M/M/1 System Under Strict Type 1 Priority--State Transition Equations

setup time is required when a server switches from one type of customer to another. This situation is illustrated in Figure 2.6.

In the nonpreemptive,two-queue M/M/1 example, the basic difference is in the state description; we now define state (i,j,k,ℓ) to represent i type 1 and j type 2 customers in the system ($i = 0,1,\ldots,N1$; $j = 0,1,\ldots,N2$), a type $k(k = 1,2)$ customer currently in service and a type $\ell(\ell = 0,1,2)$ customer last in service (ℓ is set to zero if the current customer arrived to find an idle server). Average service rates are now specified by $\mu_{k\ell}$ where k is the type of customer currently in service and ℓ is the type previously served. The number of equations increases to $6(N1+1)(N2+1) + 1$. (See [36] for more detail on solving these systems, including listings of the corresponding Chapman-Kolmogorov equations for each of the four priority schemes described in Section 2.2.2.2.1.

2.2.2.2.3 Performance Measures

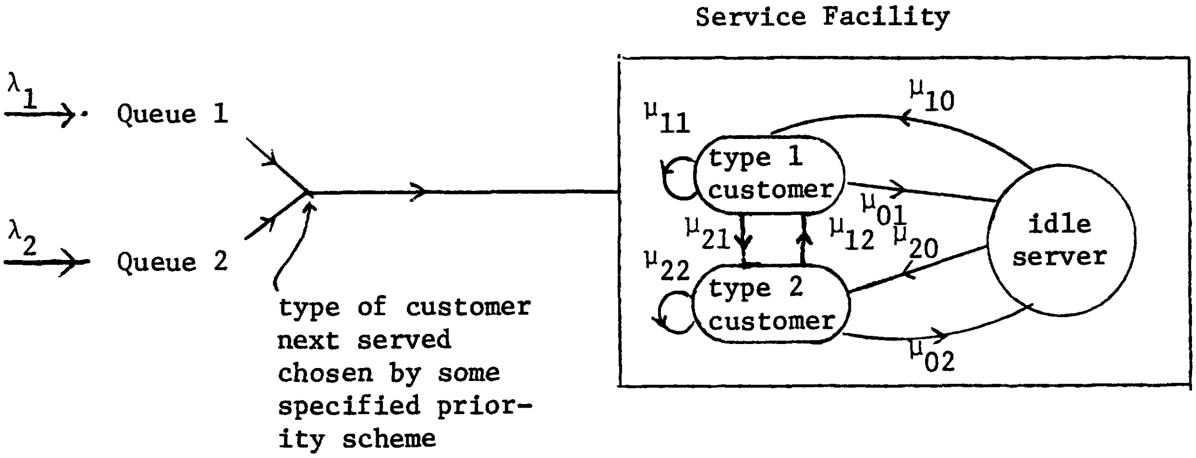
As before, solving the set of differential equations for multiple queue systems will yield the state probabilities as a function of time. These can in turn be used to obtain statistics to describe system behavior. Examples of expressions for the two-queue M/M/1 system are presented below:

(i) The probability that a type 1 customer is in service at time T,

$$\text{PROB1}(T) = \sum_{i=1}^{N1} \sum_{j=0}^{N2} P_{i,j}(T). \tag{2.4}$$

(ii) The probability that a type 2 customer is in service at time T,

$$\text{PROB2}(T) = \sum_{i=0}^{N1} \sum_{j=1}^{N2} Q_{i,j}(T). \tag{2.5}$$



$\mu_{k\ell}$ = service rate for a type k customer which follows a type ℓ customer, $k = 1,2$; $\ell = 0,1,2$.

$\mu_{0\ell}$ = transition rate for the time to reach the idle state from the instant the last customer in the system completes service (frequently $\mu_{0\ell} = \infty$), $\ell = 1,2$.

Figure 2.6: Two-Queue, Single-Server System With Priority Schemes and Changeover Costs

(iii) The expected number of type 1 customers in the system at time T,

$$EL1(T) = \sum_{i=1}^{N1} \sum_{j=0}^{N2} i[P_{i,j}(T) + Q_{i,j}(T)]. \qquad (2.6)$$

(iv) The expected number of type 2 customers in the system at time T,

$$EL2(T) = \sum_{i=0}^{N1} \sum_{j=1}^{N2} j[P_{i,j}(T) + Q_{i,j}(T)]. \qquad (2.7)$$

(v) The expected number of customers in queue 1 at time T,

$$EQ1(T) = \sum_{i=1}^{N1} \sum_{j=0}^{N2} [(i-1)P_{i,j}(T) + iQ_{i,j}(T)]. \qquad (2.8)$$

(vi) The expected number of customers in queue 2 at time T,

$$EQ2(T) = \sum_{i=0}^{N1} \sum_{j=1}^{N2} [jP_{i,j}(T) + (j-1)Q_{i,j}(T)]. \qquad (2.9)$$

(vii) The total expected number of customers in queue at time T,

$$EQ(T) = EQ1(T) + EQ2(T) \qquad (2.10)$$

2.2.2.2.4 Expected Delay

One measure which is usually of great interest in evaluating different priority schemes is the expected time delay faced by a customer entering the system at any given time. Unfortunately, because of the interaction between queues, this is a very difficult quantity to determine for most multiple-queue systems under priority schemes, and a topic to which little research has been directed. This quantity can be evaluated for the nonpreemptive, two-queue M/M/1 system under strict priority; the derivation is presented below.

Assume that type 1 customers are allowed absolute priority. Consider first a type 1 customer who enters the system at time T. Assuming FCFS service within each queue, this customer must wait for the customer currently in service and those ahead of her in queue 1 to complete service before she can begin her service. Thus,her expected wait in queue, $W_1(T)$, is given by

$$W_1(T) = \sum_{i=1}^{N1-1} \sum_{j=0}^{N2} E \left[\begin{array}{l}\text{time to}\\ \text{serve the}\\ \text{type 1 cus-}\\ \text{tomers al-}\\ \text{ready in}\\ \text{system at}\\ \text{time T}\end{array} \middle| \begin{array}{l}\text{arriving}\\ \text{customer}\\ \text{finds}\\ \text{system in}\\ \text{state}\\ \text{(i,j,1) at}\\ \text{time T}\end{array}\right] P \left[\begin{array}{l}\text{arriving}\\ \text{customer}\\ \text{finds}\\ \text{system in}\\ \text{state}\\ \text{(i,j,1) at}\\ \text{time T}\end{array}\right]$$

$$+ \sum_{i=0}^{N1-1} \sum_{j=1}^{N2} E \left[\begin{array}{l}\text{time to}\\ \text{serve the}\\ \text{type 1 cus-}\\ \text{tomers al-}\\ \text{ready in}\\ \text{system } \underline{\text{and}}\\ \text{the type 2}\\ \text{customer in}\\ \text{service at}\\ \text{time T}\end{array} \middle| \begin{array}{l}\text{arriving}\\ \text{customer}\\ \text{finds}\\ \text{system in}\\ \text{state}\\ \text{(i,j,2) at}\\ \text{time T}\end{array}\right] P \left[\begin{array}{l}\text{arriving}\\ \text{customer}\\ \text{finds}\\ \text{system in}\\ \text{state}\\ \text{(i,j,2) at}\\ \text{time T}\end{array}\right]$$

$$= \sum_{i=1}^{N1-1} \sum_{j=0}^{N2} \frac{i}{\mu_1} P_{ij}(T) + \sum_{i=0}^{N1-1} \sum_{j=1}^{N2} \left(\frac{i}{\mu_1} + \frac{1}{\mu_2}\right) Q_{i,j}(T) \qquad (2.11)$$

If the arriving customer is of type 2, her waiting time has three components--the remaining time for the customer currently in service, the time for the type 1 and type 2 customers already in queue at time T to be served, and the service time for all type 1 arrivals occurring while the arriving customer waits. Thus, the expected delay faced by a customer of type 2 arriving at time T, $W_2(T)$, can be found by solving the following equation:

$$W_2(T) = E[\text{time to serve the customers already in the system at time T}]$$

$$+ E\left[\begin{array}{l}\text{time to serve all type 1 customers that}\\ \text{arrive while the type 2 customer waits}\end{array}\right]$$

$$= \sum_{i=0}^{N1} \sum_{j=0}^{N2-1} \left\{ \left[\frac{i}{\mu_1} + \frac{j}{\mu_2}\right] \left[P_{i,j}(T) + Q_{i,j}(T)\right] \right\} + \frac{\lambda_1}{\mu_1} W_2(t) \qquad (2.12)$$

Thus

$$W_2(t) = \left(\frac{1}{1-\lambda_1/\mu_1}\right) \left[\sum_{i=0}^{N1} \sum_{j=0}^{N2-1} \left(\frac{i}{\mu_1} + \frac{j}{\mu_2}\right) P_{i,j}(t) + \sum_{i=0}^{N1} \sum_{j=0}^{N2-1} \left(\frac{i}{\mu_1} + \frac{j}{\mu_2}\right) Q_{i,j}(t)\right]. \qquad (2.13)$$

The expected delay under other priority schemes for this system is an open research topic at this time. One observation is that the expected delay faced by a virtual customer of either type will be bounded from above and below by the expected delays to that type of customer under the strict type 1 and strict type 2 priority rules. Other directions for research include the derivation of expressions for the expected delay in multiple-queue M/M/1 systems and in non-M/M/1 systems.

2.2.3 Sources of Error

The solution technique outlined in the previous sections can be used, in principle, to solve any system which can be expressed as a first-order Markov process. There are, however, three sources of error which must be noted:

(i) error due to the numerical solution of the state equations (round-off error),

(ii) for partially deterministic systems in which the embedded chain is a first-order Markov process, error due to the assumption that all arrivals and service completions occur the instant before the start of an epoch, and

(iii) for infinite-capacity systems and for finite-capacity systems defined by a large number of state equations, error caused by solving a truncated set of these equations.

The maximum size of the error due to the numerical solution of the state equations is specified by the user and will be discussed further in Section 2.2.4. The error introduced when solving partially deterministic systems is a function of the traffic intensity, ρ. As mentioned in Section 2.2.1.2, this error is negligible on a percentage basis for large ρ, but may be significant when ρ is close to 0.

Finally, we mention a modification which can reduce the error which arises when solving a truncated set of state equations. After each iteration, the probability of a saturated system, $P_N(t)$, is examined. If $P_N(t)$ is larger than some small specified value ε (e.g., we have used $\varepsilon = 10^{-8}$), the computer program will increase N by some fixed number n (we have used n=5), and continue. Conversely, if $P_N(t)$ is smaller than ε, and $P_{N-n}(t)$ is less than ε, N will be reduced by n for the next iteration. Thus, by maintaining a negligible probability of saturation, we can solve a system which has effectively infinite capacity. This will be discussed further in Section 2.2.4 along with the limitations on the number of equations which can be solved.

2.2.4 Computational Characteristics

The IMSL[11] subroutine DVERK was used to obtain numerical solutions to the set of first-order, differential equations which specify behavior in a Markovian system. This program employs a fifth- and sixth-order Runge-Kutta method to solve for the state probabilities as a function of time. The global error (total error per call to DVERK)[12] is proportional to a user-specified tolerance level (we used 10^{-6}).

This subroutine is available in both single and double precision arithmetic. Use of single precision results in a significant reduction of CPU time. However, our experience indicates that the additional accuracy afforded through use of double precision arithmetic is necessary in solving systems with more than forty equations.

The number of equations necessary to define the system was found to be the major contributing factor to computation cost. To illustrate, in Figure 2.7 we show the CPU time required to solve an M/M/1 system with $\rho = .75$, $\mu = 1$, and a finite capacity of N customers as a function of N. (Note that (N+1) Chapman-Kolmogorov equations are required to define behavior in this system. Each case was run for 30 units of model time. These results suggest that time increases in an approximately linear manner with system capacity (or, equivalently, number of equations). Although the size of the matrix of coefficients for the system equations increases as N^2, this matrix is sparse -- in fact it is tridiagonal -- and therefore the number of equations is the dominant factor contributing to computation cost.

[11] International Math-Science Library.

[12] One call to DVERK is required for each point in time in the model when output is desired.

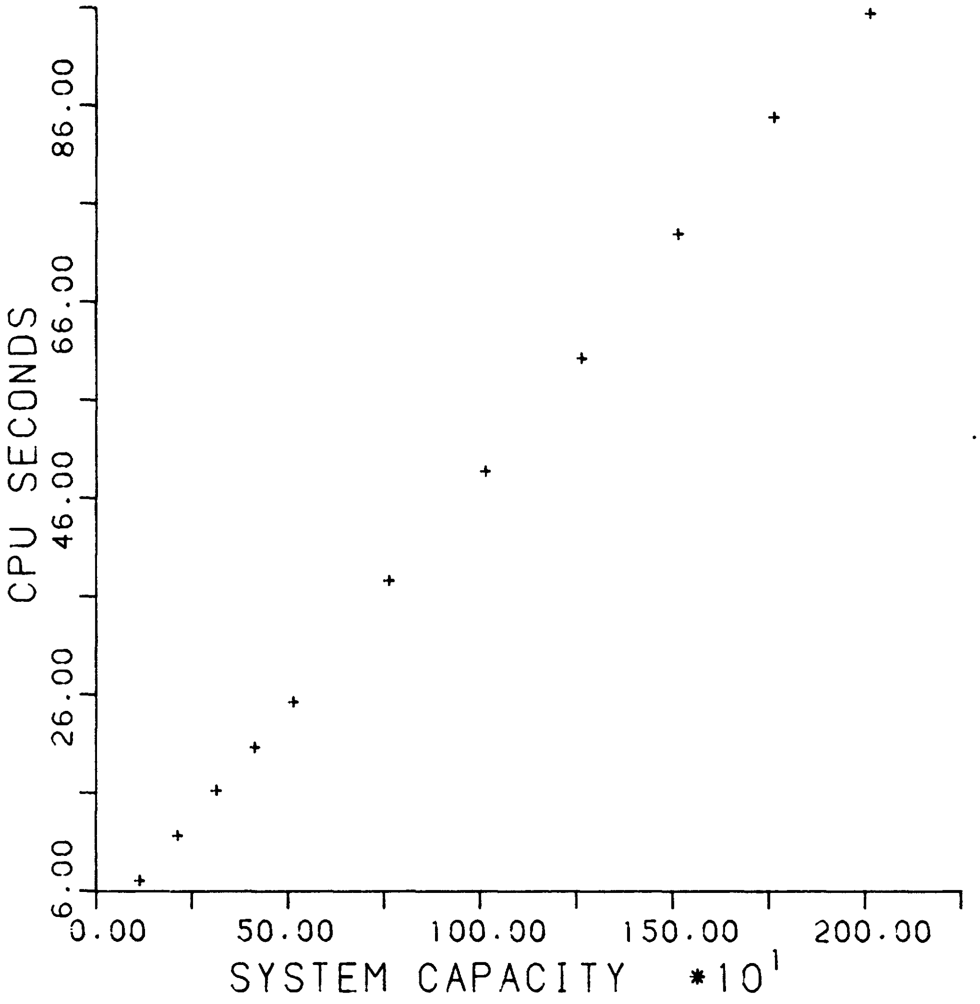


Figure 2.7: Computation Time for Numerical Solution of a Finite-Capacity M/M/1 System with ρ = .75 and μ = 1 for 30 Units of Model Time

Computation cost is also dependent on the sparsity of the equations being solved. For instance, solving a single-queue M/M/1 system will require less CPU time than solving a single-queue $M/H_2/1$ system even if the number of equations is identical. Assuming systems that begin at rest, the solution of an M/M/1 system with $\rho = .75$, $\mu = 1$, and $N = 350$ for 30 units of model time will require about 17.8 CPU seconds, less than the 27.6 CPU seconds required to solve an $M/H_2/1$ system with $\rho = .75$, $\alpha = .2$, $\mu_1 = 1$, $\mu_2 = 2$, and $N = 175$ for an equivalent amount of model time. Note that in each case, 351 first-order, differential equations are required to specify system behavior.

In Table 2.2 we illustrate, for several single-queue systems, the manner in which the number and sparsity of the system equations varies with the interarrival and service time distributions. For comparison purposes, we indicate the sparsity of an equation by the total number of state probabilities required to compute the derivative.

As mentioned in Section 2.2.3, this numerical solution technique can often be utilized to solve infinite-capacity queueing systems and finite-capacity systems defined by a large number of state equations by varying N throughout the computer run to maintain a negligible probability of system saturation. We illustrate the cost of this through several examples with Erlangian interarrival or service times. All systems have $\rho = .75$, $\mu = 1$, begin at rest, and were solved for 30 units of model time. Figures 2.8 and 2.9 show how CPU time varies with k in the solution of infinite-capacity $M/E_k/1$ and $E_k/M/1$ queueing systems.

Recursive solution of the difference equations for partially deterministic systems in which the embedded chain is a first-order Markov process is straightforward and significantly less

Table 2.2: A Characterization of the Sets of State Equations for Several Finite-Capacity Markovian Queueing Systems

System	Total Number of System Equations	Number of Equations With 2 Operations	Number of Equations With 3 Operations	Number of Equations With 4 Operations
M/M/k	N+1	2	N-1	0
E_k/M/1	kN+k	k	kN	0
M/E_k/1	kN+1	k+1	kN-k	0
E_{k_1}/E_{k_2}/1	$k_1k_2N+k_1$	2	$2k_1+2k_2-4$	$k_1k_2N-k_1k_2-k_1-k_2+2$
M/H_2/1	2N+1	2	1	2N-2
M/$\hat{G}$/1*	4N+1	4	2N-1	2N-2

*where the interarrival time, s, has pdf

$$f_s(s_o) = \beta\mu_1 e^{-\mu_1 s_o} + (1-\beta)\,\frac{\mu_2^3}{2}\,s_o^2\,e^{-\mu_2 s_o}\ , \quad s_o \geq 0.$$

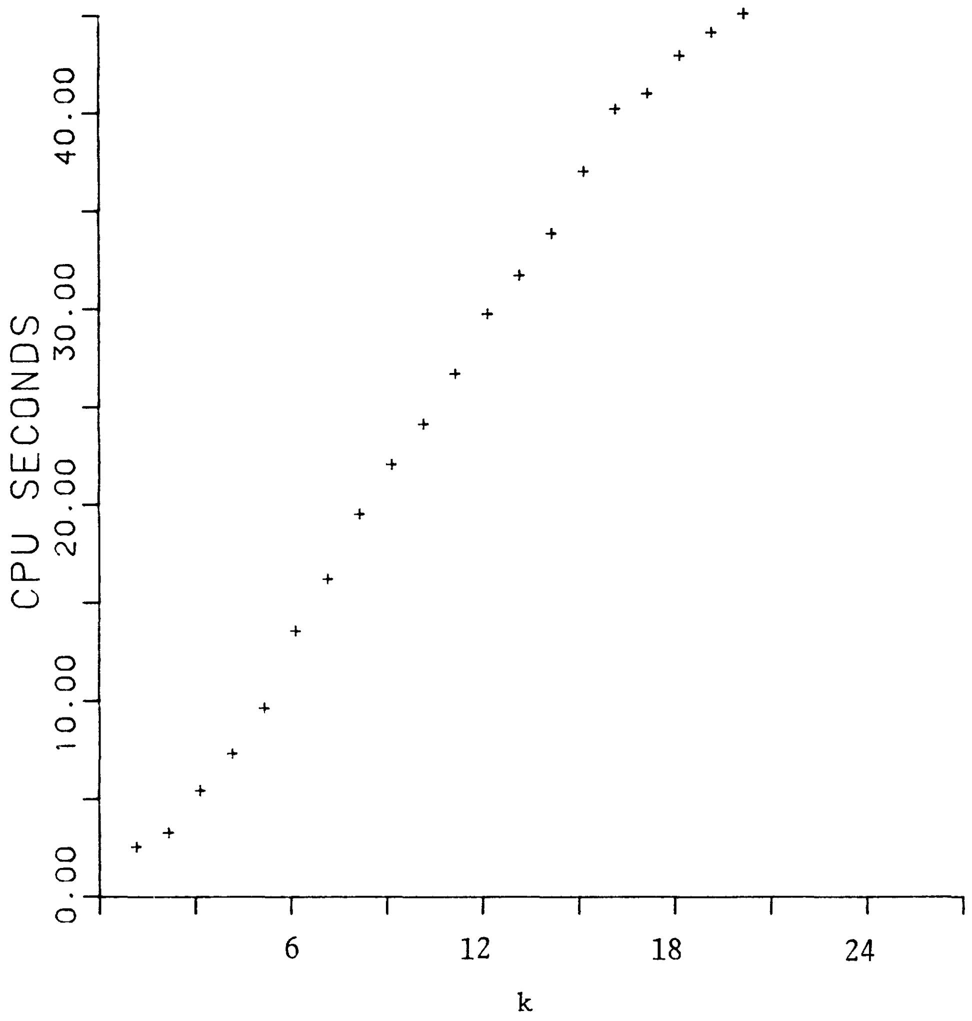


Figure 2.8: Computation Time for Numerical Solution of an Infinite-Capacity M/E_k/1 System With ρ=.75, μ=1, and $P_0(0)$=1, for 30 Units of Model Time

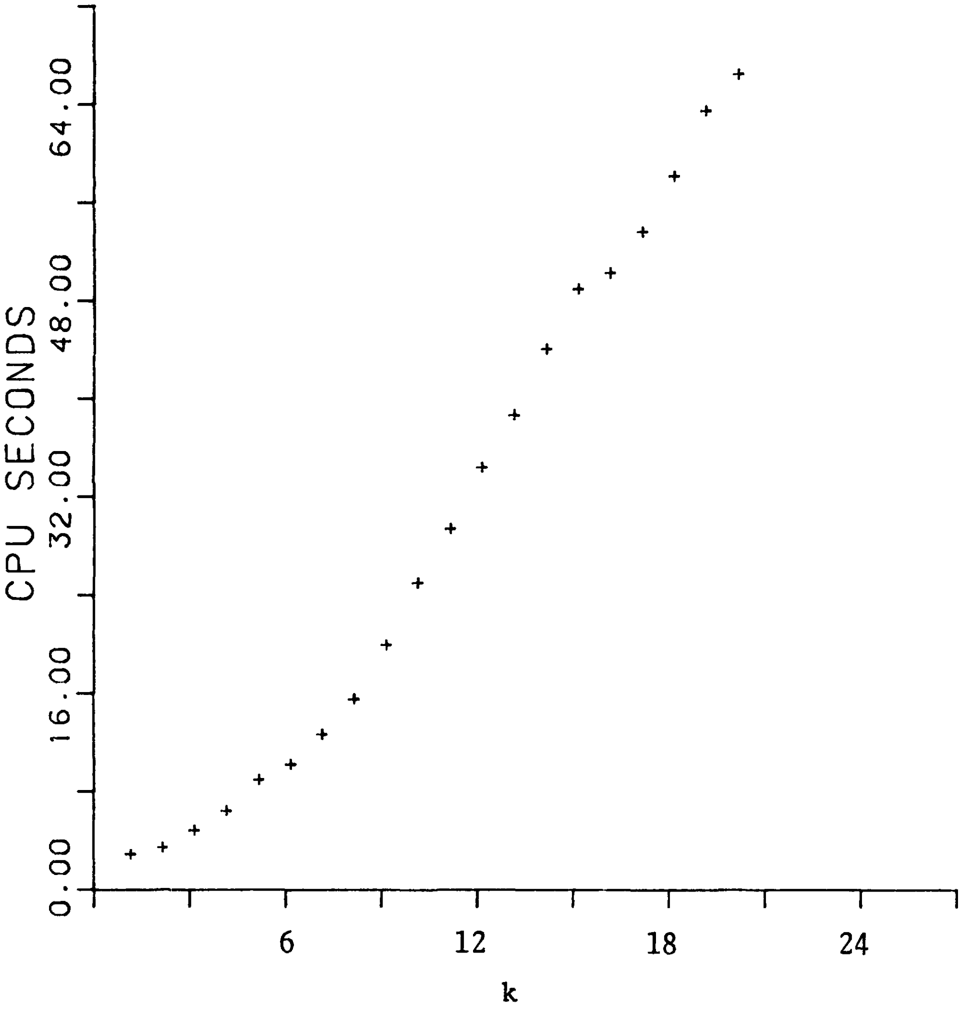


Figure 2.9: Computation Time for Numerical Solution of an Infinite-Capacity $E_k/M/1$ System With ρ=.75, μ=1, and $P_0(0)$=1, for 30 Units of Model Time

costly than solution of Markovian systems. As an example, to solve an infinite-capacity M/M/1 system with $\rho = .75$ and $\mu = 1$ for 30 minutes of the model time requires about 2.0 CPU seconds -- if the service time is deterministic only .4 CPU seconds are needed.

Listings of all computer programs used in this dissertation appear in Appendix 2. Computations were performed on an IBM 370/168 computer at the MIT Information Processing Center and on a DEC20 computer at Carnegie-Mellon University.

CHAPTER 3

TRANSIENT BEHAVIOR OF THE EXPECTED QUEUE LENGTH OF INFINITE-CAPACITY, SINGLE-QUEUE, SINGLE-SERVER SYSTEMS WHICH BEGIN AT REST

The primary purpose of this dissertation is to investigate empirically the transient response of stationary queueing systems. The technique described in Section 2.2 can be used to obtain numerical transient solutions to many systems. In this chapter, we examine a large sample of these solutions in an attempt to determine, for these same systems, a closed-form expression to approximate transient behavior. Such a closed-form expression is very attractive since it may provide a rough indication of system behavior inexpensively and without need of a computer. In addition, with this approximation we hope to estimate the amount of time required for the transient effects (caused by the initial state of the system) to become negligible. This information can be very useful in applications in determining whether the system is essentially in equilibrium. If so, for many stationary systems existing steady-state theoretical results can then be used to easily calculate the equilibrium values of the desired performance measures (e.g., the expected queue length or the expected delay).

Despite the importance of transient effects, we are unaware of any previous attempt to characterize their general form. Given the techniques available for obtaining accurate transient solutions, it is not surprising that research has been focused on other issues. As discussed in Section 2.1, numerical techniques seem to be best suited for this purpose. The emphasis on these methods is, however, a recent phenomenon, and most work to date has focused on improving the efficiency and accuracy of the numerical techniques rather than using their results to draw conclusions on some general attributes of transient behavior.

There is, however, a small amount of literature addressing the problem of the length of the transient period. A brief review follows.

Morse [31], in his analysis of the finite-capacity, single-queue M/M/1 system shows that the transient behavior is governed by exponential decay. He also suggests an approximate time constant for this system but does not check its validity.

Newell [33] refers briefly to the problem of time to steady-state in his work on the diffusion approximation for GI/G/1 queueing systems under heavy traffic. A closed-form expression is obtained that he considers to be an order of magnitude estimate of the time required for the transient effects to become negligible. The accuracy of this expression is not verified. (This work will be discussed in greater detail in Section 3.2.)

In his work on solution techniques for nonstationary M/G/1 queueing systems, Kivestu [17] gave considerable attention to the study of time constants for these systems. The thesis provides a summary of the work of Morse and Newell, but we found the remainder of Kivestu's work to be confusing and of little value for our purposes.

Barzily and Gross [1] examine the transient response of the stationary, finite-source M/M/k queueing system. Their particular concern is measurement of the amount of time until the system reaches equilibrium. Four measures of the "distance" of the system from steady-state are compared. This report contains some interesting intuitive observations as well as several numerical examples, but the work does not progress to the point of specifying a procedure for predicting the time to equilibrium.

Finally, a recent paper by Marks [29], applies regression techniques to study the manner in which the time to steady-state depends on the traffic intensity. He fits linear, quadratic, and parabolic regression models to simulation results of an infinite-capacity, single-queue M/M/k system. The work depends on an accurate determination of the point, t_∞, after which the system is in equilibrium. This is a difficult problem in simulation (see Section 2.1.2), but Marks does not present a particularly convincing argument justifying his choice of t_∞. Thus, his work is of questionable validity.

In our work, we examine the transient behavior of the expected queue length as a representative indicator of system response. The analysis for the expected delay is comparable. All systems are assumed to be ergodic, i.e., regardless of initial conditions, the system will eventually be in equilibrium.

Our strategy here is to:

(i) postulate a functional form with which to approximate the expected queue length as a function of time,

(ii) examine the validity of this functional form through an empirical analysis, and

(iii) use these results in an attempt to determine a closed-form expression with which to estimate the amount of time until a system is essentially in equilibrium.

3.1 Characteristics of the Functional Form

In order to characterize dominant features of the transient behavior, we examine, first, two representative theoretical results. In both cases we assume the system is empty at time t=0. This is frequently an appropriate assumption in applications and is the least complicated way in which to begin our analysis.

The simplest available result is for a stationary M/M/∞ system. Since, by definition, there is never a queue when the number of servers is infinite, we use the expected number of customers in the system, L(t), as our indicator of system behavior. L(t) is given by (see Table 2.1)

$$L(t) = \frac{\lambda}{\mu} - \frac{\lambda}{\mu} e^{-\mu t} \quad , \quad t \geq 0.^{1} \qquad (3.1)$$

The second theoretical result is as follows: for a stationary M/M/1 system which has a finite capacity of N customers, Morse's work [31] (see Table 2.1) implies that Q(t), the expected queue length at time t, is given by

$$Q(t) = \sum_{i=0}^{N} iP_i(\infty) + \sum_{i=0}^{N} i \left(\frac{\lambda}{\mu} \right)^{i/2} \sum_{k=1}^{N} C_k \left[\sin \frac{ik\pi}{N+1} - \sqrt{\frac{\lambda}{\mu}} \sin \frac{(i+1)k\pi}{N+1} \right] e^{-\delta_K t} \; ,$$

$$t \geq 0, \qquad (3.2a)$$

[1] This result can be seen by the following, rather intuitive derivation. By definition, to order dx, in the increment (x, x + dx] there can be at most one (Poisson) customer arrival to the system. Thus, if the system is at rest at t=0,

$$L(t) = \int_{o}^{t} P \begin{pmatrix} \text{customer arriving in } (x,x+dx] \\ \text{still in service at time } t \end{pmatrix} \; P \begin{pmatrix} \text{there is a customer} \\ \text{arrival in } (x,x+dx] \end{pmatrix} , \; t \geq 0,$$

or

$$L(t) = \int_{o}^{t} e^{-\mu(t-x)} \lambda dx$$

$$= \frac{\lambda}{\mu} - \frac{\lambda}{\mu} e^{-\mu t} \; , \; t \geq 0 \; .$$

where

$$\delta_k = \lambda + \mu - 2\sqrt{\lambda\mu} \ \cos\left(\frac{k\pi}{N+1}\right), \ k = 1,2,\ldots,N, \tag{3.2b}$$

$P_i(\infty)$, $i = 0,1,\ldots,N$, are the equilibrium state probabilities, and the coefficients C_k, $k = 1,2,\ldots,N$, are determined from the initial conditions.

Note that these two theoretical results have a common form--each is the sum of a constant term and one or more additional terms which decay in an exponential manner. A finite sum of exponential terms (e.g., expression (3.2a)) will asymptotically be dominated by a single exponential term , i.e., the exponential term with minimal decay rate. These observations suggest that for many practical applications an adequate and simple approximation for the expected queue length might be the sum of a constant (equal to the steady-state expected queue length) and a decaying exponential.

In this section, we test our hypothesis that the expected queue length, Q(t), approaches its equilibrium value in an approximately exponential manner for many stationary, ergodic queueing systems which have a single, infinite capacity queue served by a single server. All systems are assumed to be empty at time $t = 0$. Under our hypothesis, there exist parameters $\tau(\tau > 0)$ and $A(-\infty < A < 0)$ such that Q(t) can be expressed as

$$Q(t) \cong Q(\infty) + A\, e^{-t/\tau} \quad , \ t \geq 0. \tag{3.3}$$

The time constant τ is the amount of time required for the expected queue length to get $1/e \cong 37\%$ closer to its final value. This constant τ is also frequently referred to as the "relaxation time" of the system [31]. It will be examined in Section 3.2. If (3.3) is strictly true, the parameter A accounts for the difference between the initial and steady-state expected queue lengths. Thus, for a system which begins at rest,

$$A = Q(0) - Q(\infty) = -Q(\infty). \tag{3.4}$$

Due to the complexity of the theoretical solutions, we must rely on experimental confirmation of the hypothesis that the expected queue length can be approximated by (3.3) for some $A < 0$. To accomplish this, we use the numerical methods of Chapter 2 to compute the expected queue length over time from time $t = 0$ until $Q(t)$ has reached its equilibrium value $Q(\infty)$.

Intuitively, for a queueing system starting at rest, we expect the form of $Q(t)$ to be influenced by the traffic intensity, $\rho(0 < \rho < 1)$, and the forms of the interarrival and service time distributions.

By fixing one of these attributes at a time we will examine the influence of the remaining attribute on the form of $Q(t)$ by plotting $\log|Q(\infty)-Q(t)|$ versus time. In all cases $Q(t)$ is determined by numerically solving the state equations for a finite-capacity system. Care is taken to ensure that the probability the system is saturated is negligible (less than 10^{-8}). Thus, we solve a system which, for all practical purposes, is equivalent to an infinite-capacity system.

In order to plot $\log|Q(\infty)-Q(t)|$, it is necessary that we know the value of $Q(\infty)$, the steady-state expected queue length. In many instances $Q(\infty)$ can be calculated using exact, closed-form expressions. For example, for M/G/1 systems the well-known Pollaczek-Khintchine formula yields

$$Q(\infty) = \frac{\rho^2+\lambda^2\sigma_s^2}{2(1-\rho)} \quad , \tag{3.5}$$

where σ_s^2 is the variance of the service time.

For GI/M/1 systems, the equilibrium expected queue length is given by

$$Q(\infty) = \rho\left(\frac{\sigma}{1-\sigma}\right) \quad , \tag{3.6a}$$

where σ solves the transcendental equation

$$\sigma = f_a^T \, (\mu-\mu\sigma) \ , \tag{3.6b}$$

and $f_a^T(s)$ is the Laplace transform of the interarrival time, a. We have obtained either closed-form or numerical solutions of (3.6) for particular forms of the pdf $f_a(a_o)$, $a_o \geq 0$, including k^{th}-order Erlang, deterministic, and k^{th}-order hyperexponential distributions.

For systems in which $Q(\infty)$ cannot be calculated easily, one can frequently estimate its value through simulation or by continuing to apply the numerical solution technique until the system has reached equilibrium. (Our working definition of "steady-state" is all $\bar{t}$ such that

$\frac{|Q(t)-Q(\bar{t})|}{Q(t)} \leq 10^{-4}$ for all $t \geq \bar{t}$.)

In the next several pages we exhibit graphs of $\log|Q(\infty)-Q(t)|$ for several specific examples which show the dependence of Q(t) on the traffic intensity and on the type of the interarrival and service time distributions. A comprehensive list of the queueing systems we tested is presented in Table 3.1, later in this section.

First, we examine the dependence of the $\log|Q(\infty)-Q(t)|$ curve on the traffic intensity, ρ, by fixing the forms of the interarrival and service time distributions (both negative exponential) and then varying ρ. Figures 3.1 and 3.2 are plots of $\log|Q(\infty)-Q(t)|$ versus time for an M/M/1 system with ρ = .25, .5, .75, .85, and .9. The expected service time is fixed at $1/\mu = 1$. Figures 3.3 and 3.4 are plots of $\log|Q(\infty)-Q(t)|$ for the same five values of ρ and an expected service time of $1/\mu = 2$.

In each case, after an initial period $\log|Q(\infty)-Q(t)|$ appears, for all practical purposes, to vary linearly with time. This linear relationship

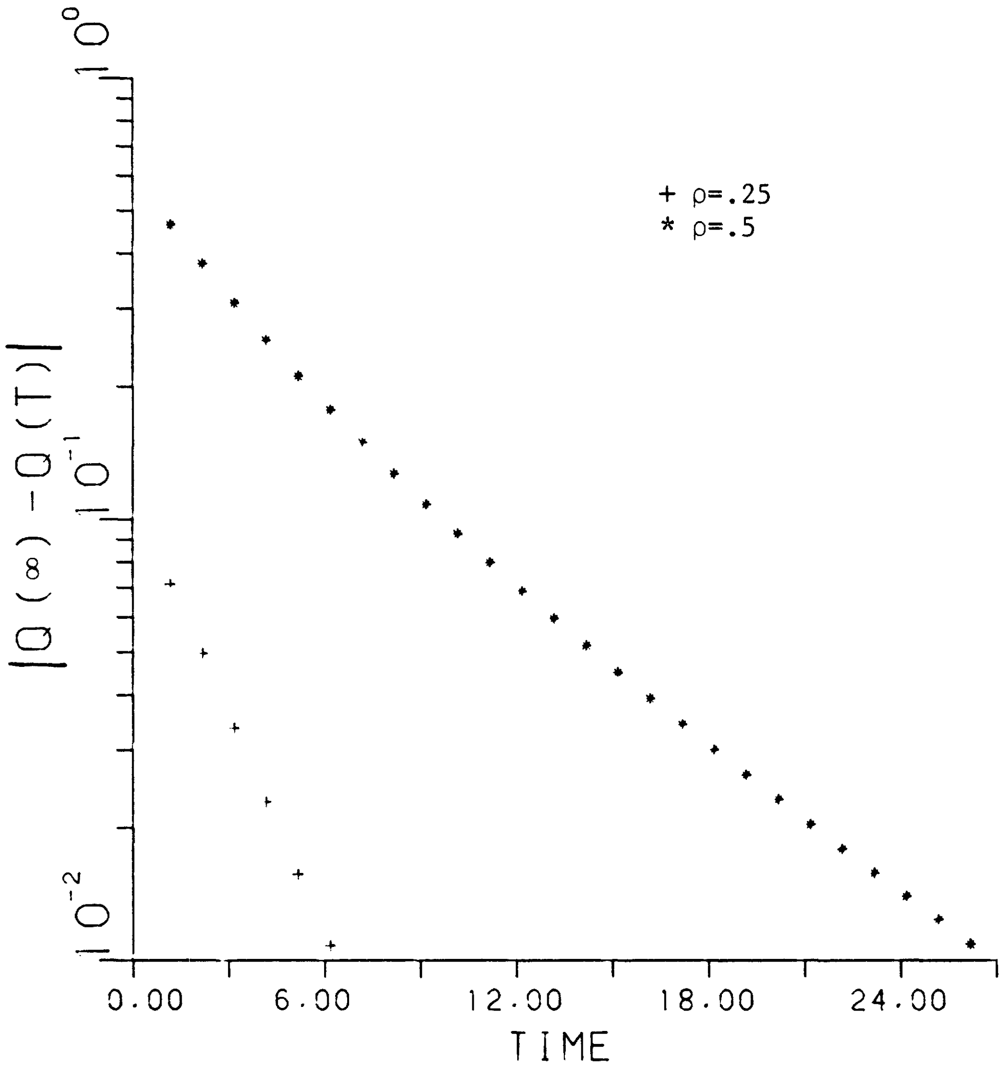


Figure 3.1: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an M/M/1 System With μ=1 and $P_0(0)$=1; ρ=.25 and .5

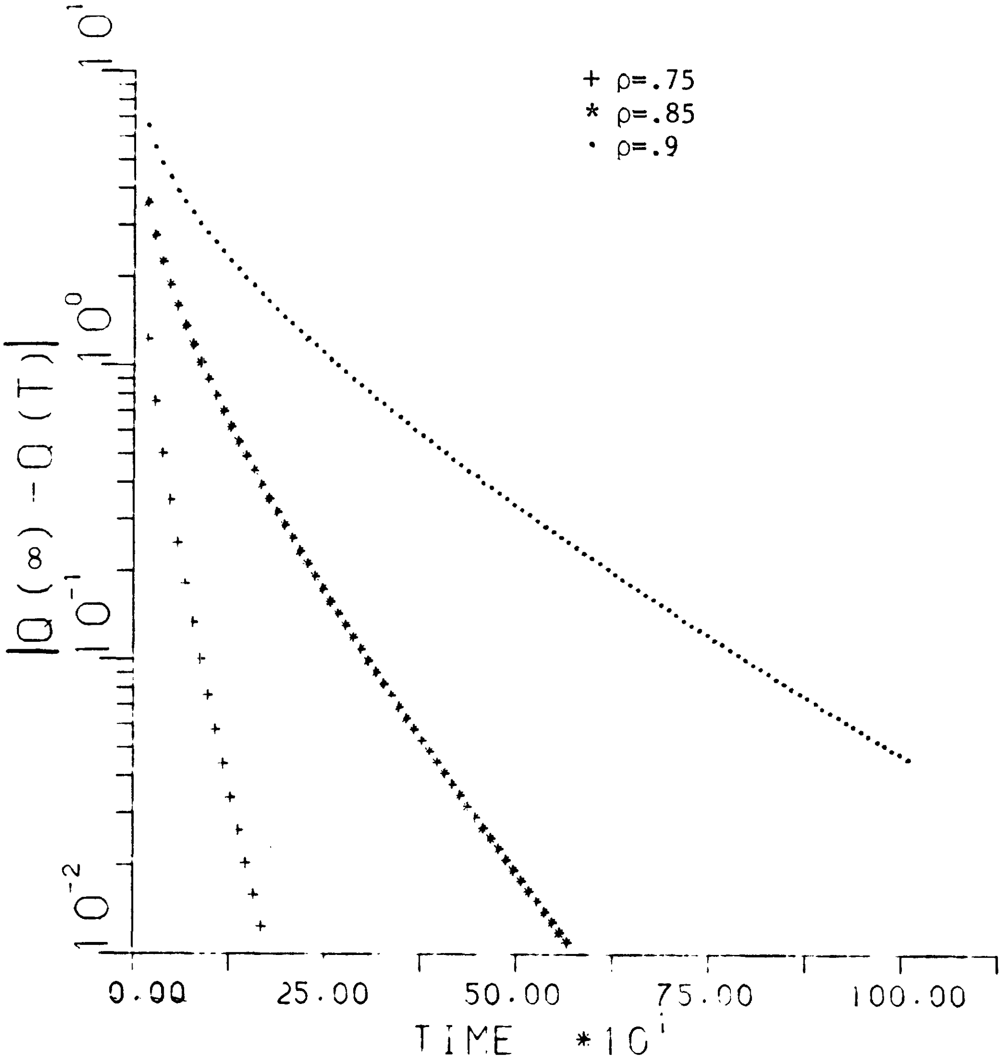


Figure 3.2: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an M/M/1 System With $\mu=1$ and $P_0(0)=1$; $\rho=.75$, $.85$, and $.9$

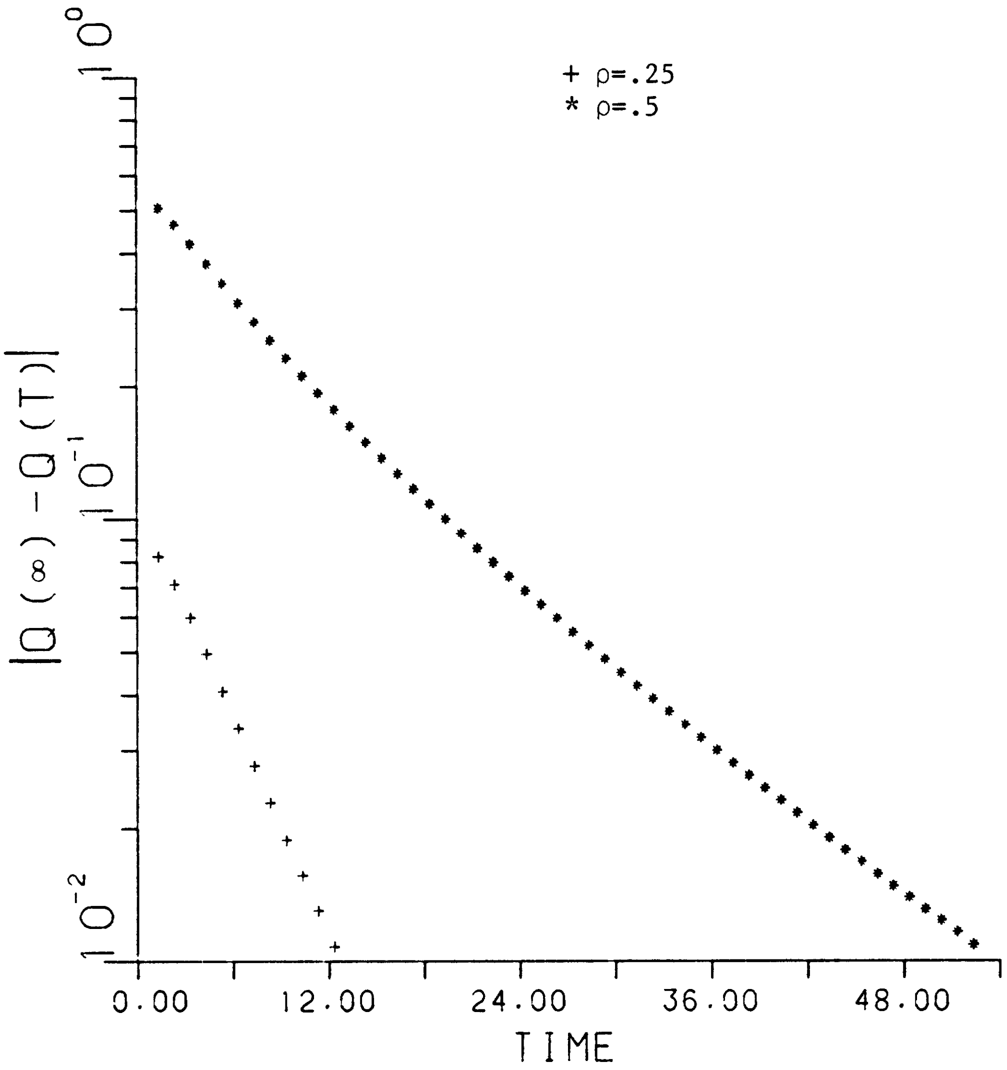


Figure 3.3: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an M/M/1 System With $\mu=.5$ and $P_0(0)=1$; $\rho=.25$ and .5

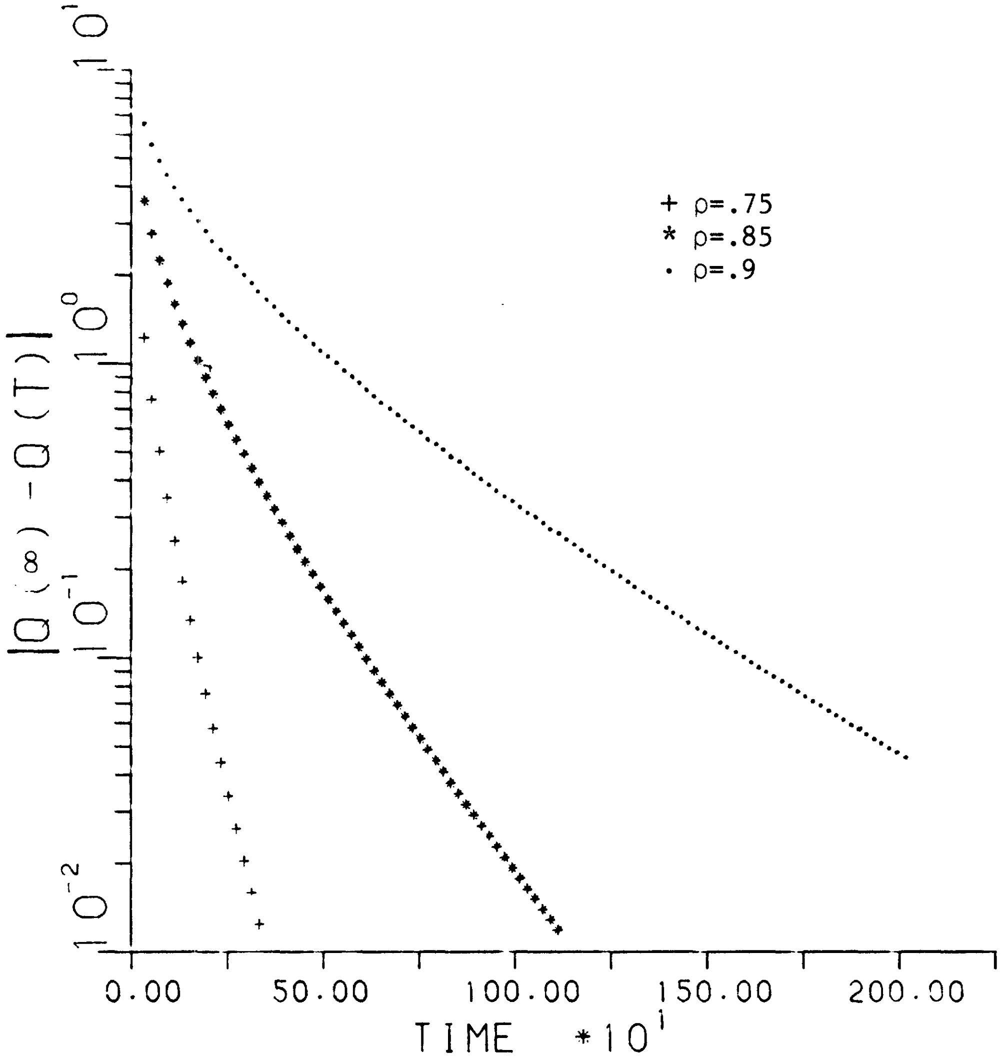


Figure 3.4: Semilogarithmic Plots of |Q(∞) - Q(T)| Versus Time for an M/M/1 System With μ=.5 and $P_0(0)$=1; ρ=.75, .85, and .9

between $\log|Q(\infty)-Q(t)|$ and t implies exponential decay. For small t the curves are convex. This implies that initial decay occurs at a rate faster than that of the eventual exponential function. We will discuss this behavior in more detail later.

To determine the effect of the particular types of interarrival and service time distributions on the general form of Q(t), we examine four other queueing systems-- M/D/1, $M/H_2/1$ (a second-order hyperexponential service time), a particular M/G/1 system with a service time given by a weighted sum of Erlang random variables, and the $E_2/E_2/1$ system. We also solved several types of $E_k/M/1$ systems; see Table 3.1.

The M/D/1 system differs from those discussed previously in that it cannot be converted into a first-order Markov process. Since only the embedded chain is first-order, we cannot obtain an exact numerical solution for all t. As discussed in Chapter 2, we solve this system (approximately) using the assumption that all arrivals and service completions occur only at the instant before the start of an epoch. This approximation is good for systems with large ρ once the queue is sufficiently long (insuring a negligible probability of an arriving customer entering an empty system).

Figure 3.5 is a plot of $\log|Q(\infty)-Q(t)|$ versus time for an M/D/1 system with $\rho = .9$ and $\mu = 1$. Once again, for large t, Q(t) decays in an approximately exponential manner.

The $M/H_k/1$ system is "more random" (i.e., exhibits greater variability) than corresponding M/M/1 or $M/E_k/1$ systems since the hyperexponential random variable which specifies the service time has coefficient of variation greater than 1.[2] (A hyperexponential random variable

[2] The coefficient of variation is defined as the ratio of the standard deviation to the mean. It can be interpreted as a normalized measure of spread.

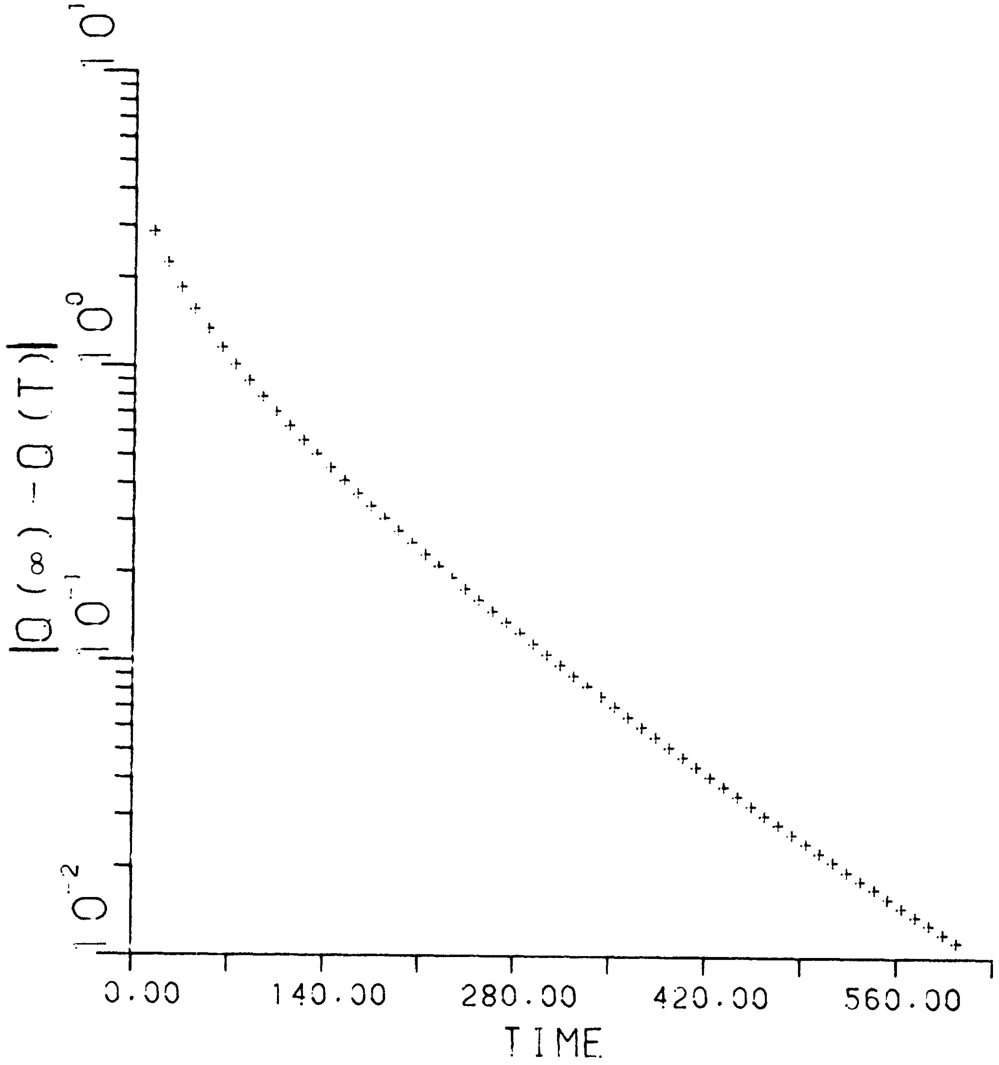


Figure 3.5: A Semilogarithmic Plot of $|Q(\infty) - Q(T)|$ Versus Time for an M/D/1 System With $\rho=.9$, $\mu=1$, and $P_0(0)=1$

is specified as a probabilistic choice among negative exponential random variables.) To illustrate the behavior of this type of system, we consider the $M/H_2/1$ system depicted in Figure 3.6. Customers arrive according to a Poisson process with parameter λ and have a service time given by a negative exponential random variable. The mean service time for any particular customer is chosen through an independent Bernoulli trial: with probability α, $0 < \alpha < 1$, the expected service time is $1/\mu_1$; with probability $(1-\alpha)$ the expected service time is $1/\mu_2$. The probability density function for this hyperexponential service time, s, is

$$f_s(s_o) = \alpha\, \mu_1 e^{-\mu_1 s_o} + (1-\alpha)\mu_2 e^{-\mu_2 s_o} \quad , \; s_o \geq 0, \tag{3.7}$$

and the unconditional expected service time for this system is

$$1/\mu = \alpha\, 1/\mu_1 + (1-\alpha)\, 1/\mu_2 \quad . \tag{3.8}$$

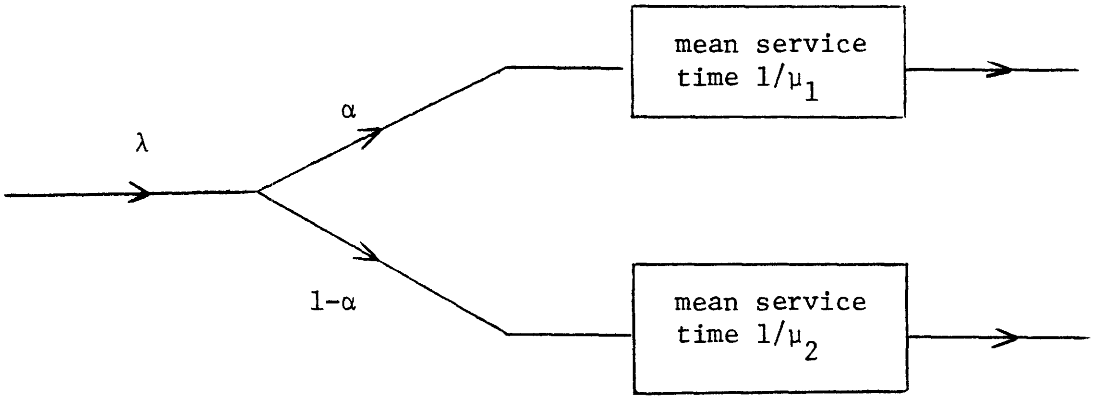
Figure 3.7 is a graph of $\log|Q(\infty)-Q(t)|$ versus time for an $M/H_2/1$ system with $\lambda = 1.25$, $\alpha = .2$, $\mu_1 = 1$, $\mu_2 = 2$, and $\rho = .75$. Once again, after an initial period decay is approximately exponential.

We now examine a single-server system with Poisson arrivals and a particular "phase type" probability density function for the service time, s, given by

$$f_s(s_o) = \beta\mu_1 e^{-\mu_1 s_o} + \frac{(1-\beta)\mu_2^3}{2} s_o^2\, e^{-\mu_2 s_o} \quad , \; s_o \geq 0, \tag{3.9}$$

where $0 < \beta < 1$. Clearly, random variable s is a "weighted combination" of (or, probabilistic choice between) negative exponential and third-order Erlang random variables. This implies that

$$E(s) = \frac{\beta}{\mu_1} + \frac{3(1-\beta)}{\mu_2} \tag{3.10}$$



interarrival and service times given by independent exponential random variables.

Figure 3.6: The $M/H_2/1$ Queueing System

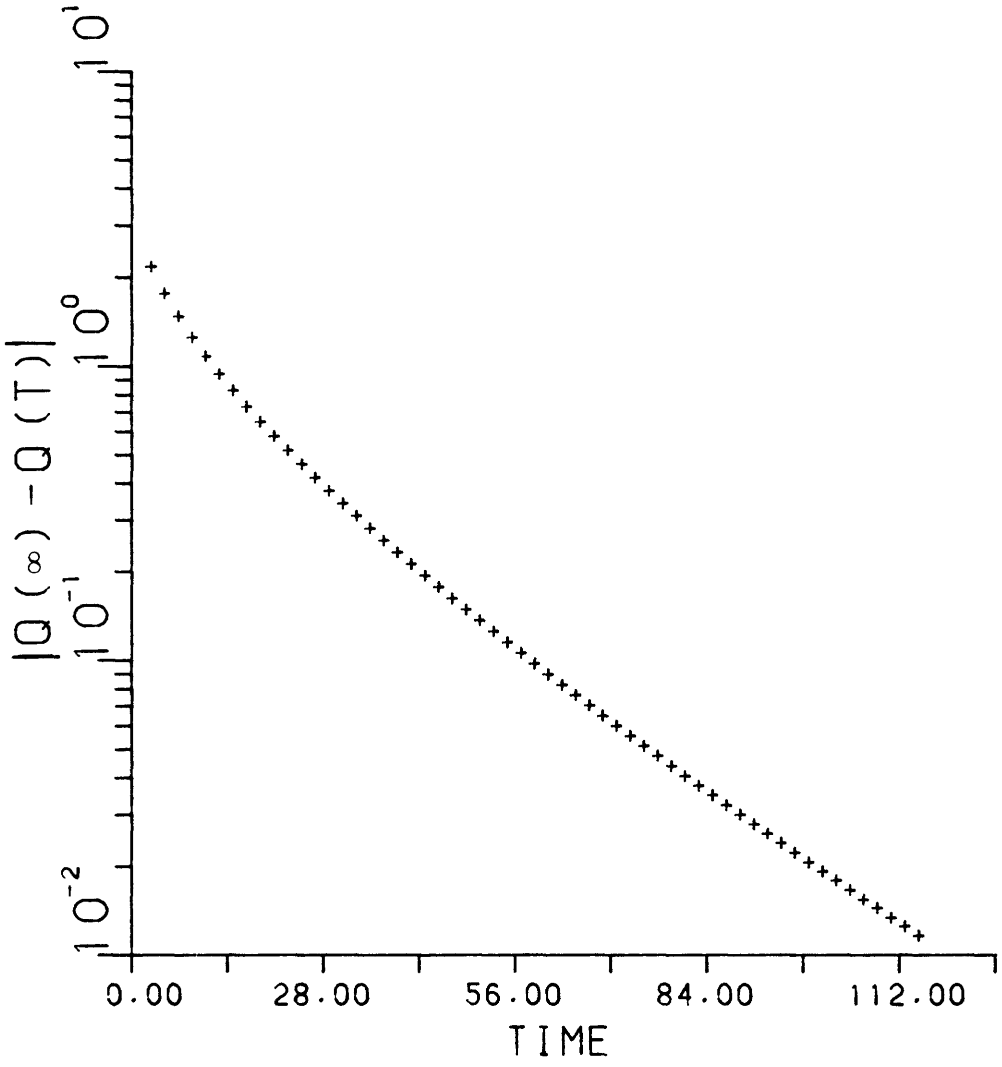


Figure 3.7 A Semilogarithmic Plot of $|Q(\infty) - Q(T)|$ Versus Time for an $M/H_2/1$ System With $\rho=.75$, $\alpha=.2$, $\mu_1=1$ $\mu_2=2$, and $P_0(0)=1$

and

$$\sigma_s^2 = \frac{2\beta}{\mu_1^2} + \frac{12(1-\beta)}{\mu_2^2} - [E(s)]^2 \ . \qquad (3.11)$$

To illustrate, Figure 3.8 is a plot of $f_s(s_o)$, $s_o \geq 0$, for $\beta = 1/6$ and $\mu_1 = \mu_2 = 1$. Note that this probability density function is slightly bimodal with maxima at $s_o = 0$ and $s_o = 1.775$ and a local minimum at $s_o = .225$.

We solved a queueing system with a service time given by this random variable s and $\lambda = .3$ (note that $\rho = .8$). Figure 3.9 is a plot of $\log|Q(\infty)-Q(t)|$ versus time for this system. As before, except for small t, decay is approximately exponential.

To test our hypothesis for a system which has nonexponential interarrival _and_ service time distributions we examined an $E_k/E_k/1$ queueing system. Figure 3.10 is a plot of $\log|Q(\infty)-Q(t)|$ for an $E_2/E_2/1$ system with $\rho = .75$ and $\mu = 1$.

Finally, Table 3.1 is a complete listing of the systems we examined. In all cases behavior is similar--the expected queue length eventually approaches its equilibrium value in an approximately exponential manner. These results seem to confirm that, for some $\hat{t} \geq 0$,

$$Q(t) \cong Q(\infty) + Ae^{-t/\tau} \ , \ t \geq \hat{t} \ .$$

3.2 Examination of the Time to Equilibrium

The experimental results presented in the previous section suggest that after an initial time period, the transients of the expected queue length decay in an approximately exponential manner for many queueing systems which begin at rest. Now we will examine the exponential time constant τ. Deriving a closed-form expression for τ would allow estimation of the amount of time

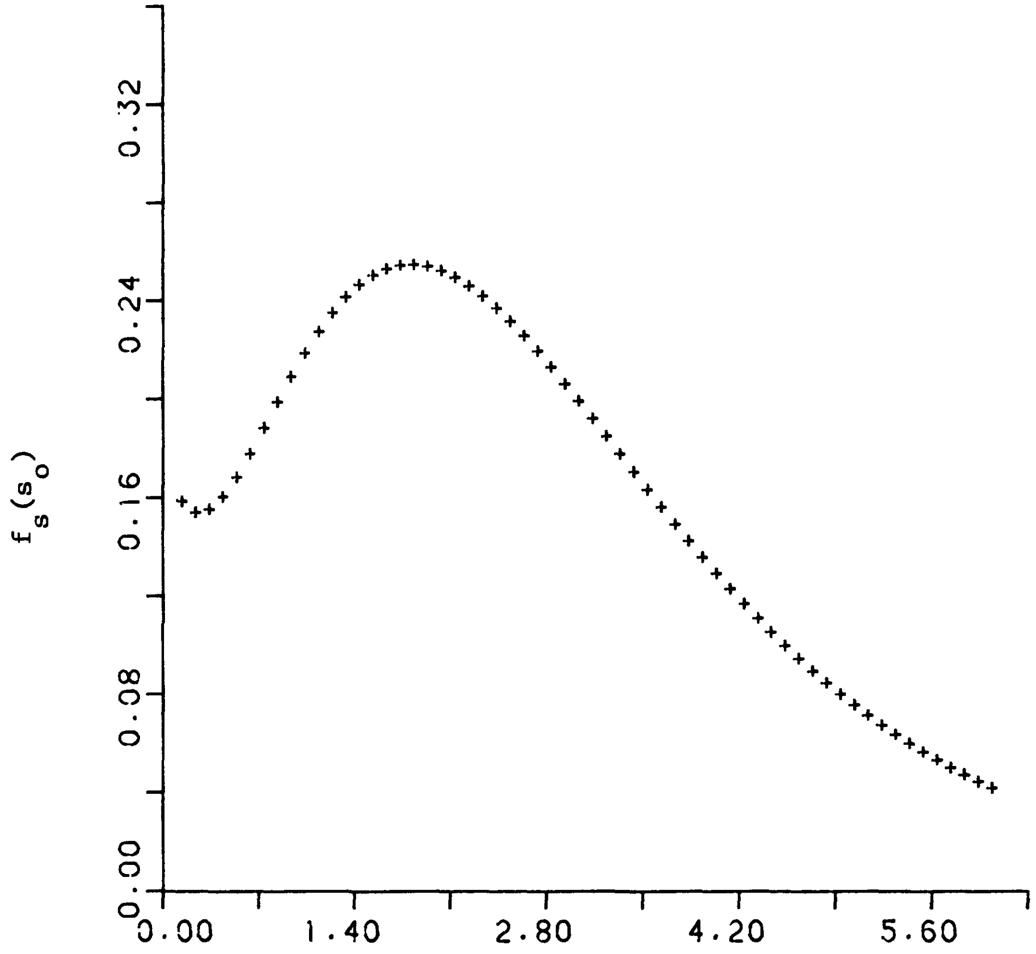


Figure 3.8: The Probability Density Function of a Random Variable that is a Probabilistic Choice Between Exponential and Third-Order Erlang Random Variables*

$$^{*}f_s(s_o) = \frac{1}{6} e^{-s_o} + \frac{5}{12} s_o^2 e^{-s_o} \quad , \; s_o \geq 0$$

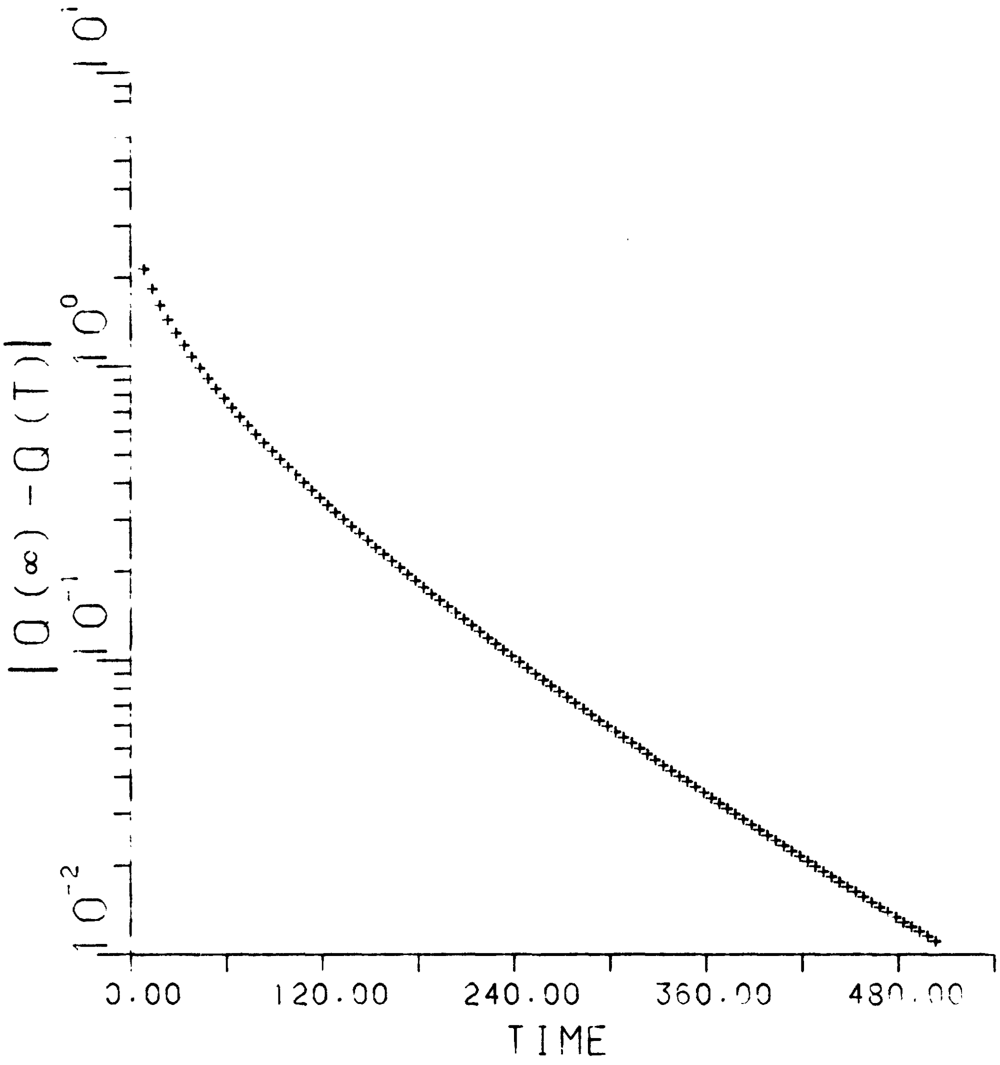


Figure 3.9: A Semilogarithmic Plot of $|Q(\infty) - Q(T)|$ Versus Time for an $M/\hat{G}/1$* System With $\rho=.8$, $\beta=1/6$, $\mu_1=\mu_2=1$, and $P_0(0)=1$

*the service time probability density function is

$$f_s(s_o) = \beta\mu_1 e^{-\mu_1 s_o} + (1-\beta)\,\frac{\mu_2^3}{2}\,s_o^{\,2}\,e^{-\mu_2 s_o}\,, \quad s_o \geq 0.$$

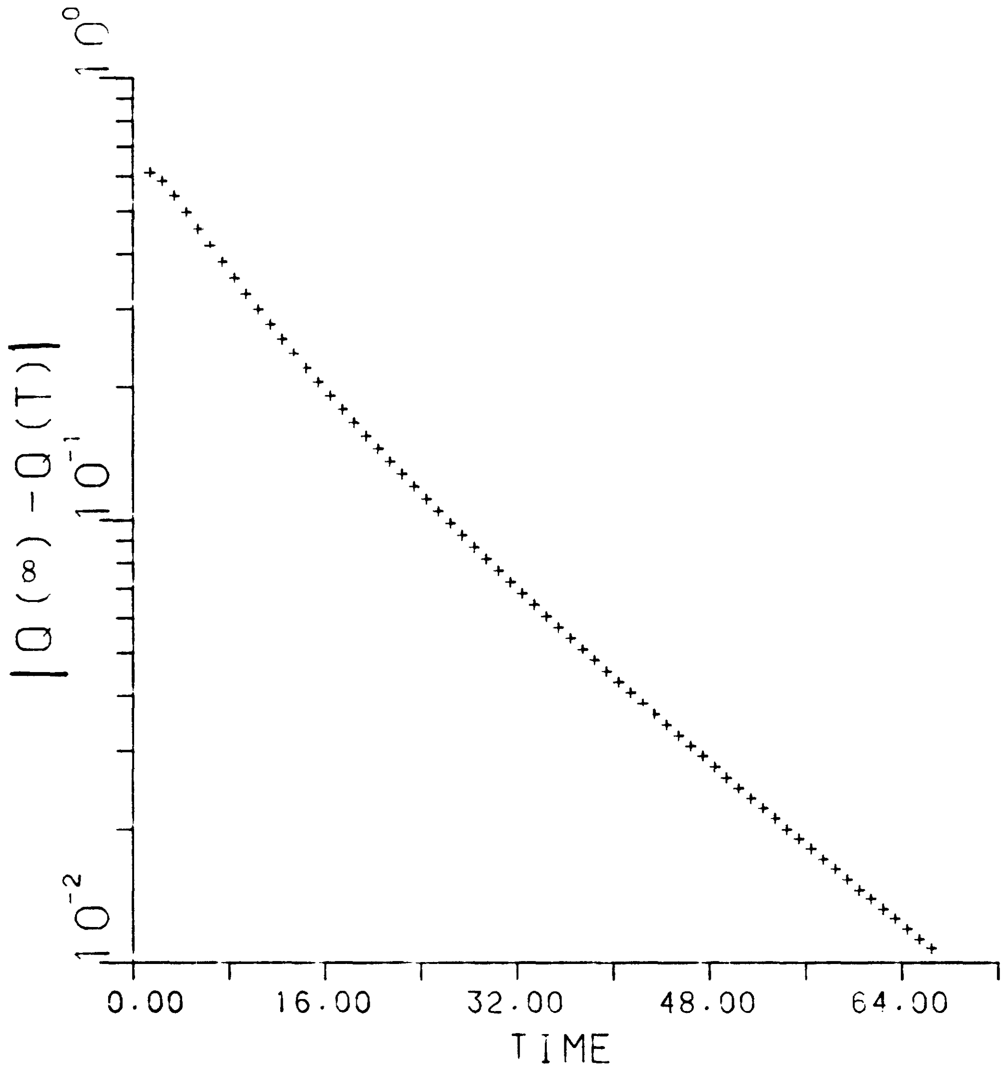


Figure 3.10: A Semilogarithmic Plot of |Q(∞)-Q(T)| Versus Time for an $E_2/E_2/1$ System With $\rho=.75$, $\mu=1$, and $P_o(0)=1^2$

Table 3.1: List of Test Cases

		ρ				
System	μ	.25	.5	.75	.85	.9
M/M/1	.5	✓	✓	✓	✓	✓
	1	✓	✓	✓	✓	✓
	2	✓	✓	✓	✓	✓
E_2/M/1	.5		✓	✓	✓	
	1	✓	✓	✓	✓	
	2		✓		✓	
E_3/M/1	.5		✓	✓	✓	
	1		✓	✓	✓	
	2		✓		✓	
E_4/M/1	.5			✓	✓	
	1			✓	✓	
	2				✓	
E_{10}/M/1	1			✓		
M/E_2/1	.5		✓	✓	✓	
	1	✓	✓	✓	✓	✓
	2			✓	✓	
M/E_3/1	.5		✓	✓	✓	
	1	✓	✓	✓	✓	✓
	2			✓	✓	
M/E_4/1	.5		✓	✓	✓	
	1		✓	✓	✓	✓
	2			✓	✓	
M/E_{10}/1	1			✓		
E_2/E_2/1	1		✓	✓	✓	
M/D/1	.5	✓	✓	✓	✓	✓
	1	✓	✓	✓	✓	✓

Table 3.1 (Continued)

System	α	μ_1	μ_2	ρ		
				.3	.75	.8
$M/H_2/1$	.2	1	2		✓	
	.4				✓	
	.6				✓	
	.8				✓	
	.2	1	.5		✓	
	.4				✓	
$M/\hat{G}/1$*	1/6	1	1			✓
	1/6	1	3	✓		

*where the service time s, has pdf

$$f_s(s_o) = \alpha\mu_1 e^{-\mu\, s_o} + (1-\alpha)\,\frac{\mu_2^3\, s_o^2\, e^{-\mu_2 s_o}}{2}\,, \quad s_o \geq 0.$$

required by a system until the transient effects can be ignored (e.g., after four time constants only $e^{-4} \cong 1.8\%$ of the transients remain).

To measure τ experimentally we need $Q(\infty)$ and two values $Q(t_1)$ and $Q(t_2)$ for any $\hat{t} \leq t_1 \leq t_2$, where $\hat{t}$ is the time after which decay is exponential. We determined$\hat{t}$ by inspection from the numerical results. This is adequate for the purposes of this work. In future work, if more accuracy is required, one more formal manner in which to estimate $\hat{t}$ would be to perform a log-linear regression on $|Q(\infty)-Q(t)|$ for large t only and then to extrapolate backward to include $t = 0$. Then, the smallest value of time which has a small residual could be chosen as the estimate of $\hat{t}$. It is important that the initial regression be a good fit, i.e., a very high R^2 value.

Once $\hat{t}$ has been chosen, using expression (3.3) we have

$$Q(t_j) \cong Q(\infty) + Ae^{-t_j/\tau}, \; j = 1, 2, \tag{3.12}$$

and therefore,

$$\frac{Q(\infty)-Q(t_1)}{Q(\infty)-Q(t_2)} = e^{-(t_1-t_2)/\tau} . \tag{3.13}$$

Rewriting expression (3.13) yields

$$\tau = \frac{t_2-t_1}{\log\left[\frac{Q(\infty)-Q(t_1)}{Q(\infty)-Q(t_2)}\right]} = \frac{t_2-t_1}{\log|Q(\infty)-Q(t_1)| - \log|Q(\infty)-Q(t_2)|} \tag{3.14}$$

for all $\hat{t} \leq t_1 \leq t_2$. Thus, under our hypothesis, τ is the negative reciprocal of the slope of $\log|Q(\infty)-Q(t)|$ versus time. Therefore, if we know $Q(t)$ for all $t \geq \hat{t}$, we can calculate τ.

Now that we can measure τ empirically, we use the experimental results of Section 3.1 to determine a closed-form approximation for τ which can be used without prior knowledge of Q(t). First, we postulate a form for τ using intuition and a result by Newell. Then, we refine this expression using the empirical values of τ for a large number of queueing systems.

Intuitively, the time constant should depend on system characteristics such as the arrival and service rates and the coefficients of variation for both the interarrival and service times. The coefficient of variation is a measure of the relative "variability" of a random variable (the larger the coefficient of variation, the greater the "variability"). Therefore, we expect τ to vary directly with powers of C_a and C_s, the coefficients of variation for the interarrival and service times, respectively, since greater variability in a system would be expected to increase the time to equilibrium.

We also expect that, as the system approaches saturation ($\rho \to 1$), or as the expected service time becomes long, the system will require more time to reach equilibrium. Thus, it is reasonable to expect τ to vary directly with powers of $\frac{1}{(1-\rho)}$ and $1/\mu$. In fact, for $\rho = \lambda/\mu$ fixed, τ must depend directly on $1/\mu$.

The simplest way to understand this last relationship is through an example. Consider a queueing system with given interarrival and service time distributions and fixed traffic intensity, ρ. In case 1, let λ and μ be specified in terms of customers/minute and in case 2, in terms of customers/hour. Thus, if λ_i and μ_i are the arrival and service rates for case i (i=1,2), $\lambda_1 = \lambda_2/60$ and $\mu_1 = \mu_2/60$. Let $Q_1(t)$ be defined as the expected queue length at t minutes, $Q_2(t)$ as the expected queue length at t hours.

Then, since the systems are equivalent with the exception of the units in which time is measured,

$$Q_2(t) = Q_1(\gamma t), \qquad t \geq 0, \tag{3.15a}$$

where

$$\gamma = \mu_2/\mu_1 = 60. \tag{3.15b}$$

Since τ_1 and τ_2 are measured in the units of their respective systems,

$$Q_2(\tau_2) = Q_1(\tau_1). \tag{3.16}$$

But, by equation (3.15b), we also have

$$Q_2(\tau_2) = Q_1(\gamma\tau_2). \tag{3.17}$$

Therefore,

$$\tau_1 = \gamma\tau_2 = (\mu_2/\mu_1)\ \tau_2\ , \tag{3.18}$$

or

$$\mu_1\tau_1 = \mu_2\tau_2 = \text{constant}. \tag{3.19}$$

This implies that τ must be proportional to $1/\mu$. Note that this result holds for general queueing systems; we made no assumptions about system characteristics in the derivation.

A result from the diffusion approximation for queues under heavy traffic suggests a form for the time constant τ. A fundamental assumption of the diffusion approximation is that queue length can be treated as a continuous rather than a discrete random variable. Defining F(x,t) as the cumulative distribution function of the queue length at time t, Newell [33] derives the following diffusion equation for the queue length:

$$\frac{\partial F(x,t)}{\partial t} = -(\lambda-\mu)\,\frac{\partial F(x,t)}{\partial x} + \frac{(\lambda C_a^2+\mu C_s^2)}{2}\,\frac{\partial^2 F(x,t)}{\partial x^2} \qquad (3.20)$$

where F(x,t) → 1 as x → ∞, and F(x,t) → 0 as x → 0.[3] Newell than makes the substitutions

$$x' = \frac{x}{d} \quad \text{and} \quad t' = \frac{t}{\tau_N} \;,$$

where

$$d = \frac{\lambda C_a^2+\mu C_s^2}{\mu-\lambda} \qquad (3.21)$$

and

$$\tau_N = \frac{\lambda C_a^2+\mu C_s^2}{\mu^2(1-\rho)^2}$$

$$= \frac{\rho C_a^2+C_s^2}{\mu(1-\rho)^2} \;. \qquad (3.22)$$

With these substitutions, (3.20) reduces to

$$\frac{\partial F'(x',t')}{\partial t'} = \frac{\partial F'(x',t')}{\partial x'} + \frac{1}{2}\,\frac{\partial^2 F'(x',t')}{\partial x'^2} \;. \qquad (3.23)$$

Based on the fact that (3.20) can be transformed into a dimensionless equation, Newell then comments that "...the relaxation time of the queue in the original time units must be of order $[\tau_N]$..." In this context the

[3] This equation states that the rate at which the P(≤ x customers in the queue) changes with time must be equal to a weighted sum of the density function of the number in queue at time t and the rate of change of this density function over x at time t. The weights are the negative rate of change of the mean of the number in queue at time t and one-half the rate of change of the variance of the number in queue at time t [22].

relaxation time corresponds roughly to the amount of time required to measure "significant" changes in the queue length. Note that Newell's expression for τ_N is <u>not</u> derived as an exponential time constant, but it is expressed in units of time, and varies directly with $1/\mu$ and powers of C_a, C_s, and $\frac{1}{(1-\rho)}$, as we expect of the exponential time constant τ. We used τ_N as our initial estimate for τ.

For each of the queueing systems listed in Table 3.1, we calculated the experimental time constant τ_{exp} by measuring the slope of the $\log|Q(\infty)-Q(t)|$ curve. We then compared τ_{exp} to τ_N (calculated from expression (3.22)). Close examination of these results led, by a trial-and-error approach, to a modified expression for the time constant, given by

$$\tau_R = \frac{(1+\sqrt{\rho})^2(C_a^2+C_s^2)}{2.7\ \mu(1-\rho)^2} \quad . \tag{3.24}$$

Like τ_N, τ_R varies directly with $1/\mu$ and the second powers of C_a, C_s, and $\frac{1}{(1-\rho)}$. Also note that since for many systems C_a^2 and C_s^2 will take values between 0 and 1, and since $0 < \rho < 1$, the most important contributions to τ_R will usually be from the $1/\mu$ and $\frac{1}{(1-\rho)^2}$ factors. In Table 3.2 we list, for each system in Table 3.1, the numerically obtained τ_{exp}, along with τ_N and τ_R. Also, we indicate the ratios τ_{exp}/τ_N and τ_{exp}/τ_R. Due to the approximate nature of this work, all ratios are expressed to one decimal place. In most cases τ_R is within about 10% of the experimental time constant. Based on these empirical results we propose τ_R as an approximate time constant for all values of ρ $(0<\rho<1)$, not only for heavy traffic conditions (Newell's assumption).

To improve expression (3.24) for τ_R, a log-linear regression on a collection of experimental cases could be used to check the powers of the

Table 3.2: A Comparison of Estimated and Observed Time Constants of Systems That Begin at Rest

System	μ	ρ	τ_{exp}	τ_N	τ_R	τ_{exp}/τ_R	τ_{exp}/τ_R
M/M/1	.5	.25	5.4	4.4	5.9	1.2	.9
		.5	16.4	12	17.3	1.4	.9
		.75	80	56	83	1.4	1.0
		.85	231	164	243	1.4	1.0
		.9	545	380	563	1.4	1.0
	1	.25	2.8	2.2	3.0	1.3	.9
		.5	7.9	6	8.6	1.3	.9
		.75	40	28	41	1.4	1.0
		.85	121	82	122	1.5	1.0
		.9	279	190	281	1.5	1.0
	2	.25	1.4	1.1	1.5	1.3	.9
		.5	3.7	3	4.3	1.2	.9
		.75	18.8	1	21	1.3	.9
		.85	60	41	61	1.5	1.0
		.9	133	95	141	1.4	.9
E_2/M/1	.5	.5	11.5	10	13.0	1.2	.9
		.75	61	44	62	1.4	1.0
		.85	178	127	182	1.4	1.0
	1	.25	2.4	2	2.2	1.2	1.1
		.5	6.4	5	6.5	1.3	1.0
		.75	28	22	31	1.3	.9
		.85	95	63	91	1.5	1.0
	2	.5	3.1	2.5	3.2	1.2	1.0
		.85	44	32	46	1.4	1.0
E_3/M/1	.5	.5	11.9	9.3	11.5	1.3	1.0
		.75	53	40	55	1.3	1.0
		.85	165	114	162	1.4	1.0
	1	.5	5.7	4.7	5.8	1.2	1.0
		.75	27	20	28	1.4	1.0
		.85	75	57	81	1.3	.9
	2	.5	2.7	2.3	2.9	1.2	.9
		.85	40	29	41	1.4	1.0
E_4/M/1	.5	.75	51	38	52	1.3	1.0
		.85	150	108	152	1.4	1.0
	1	.75	25	19	26	1.3	1.0
		.85	75	54	76	1.4	1.0
	2	.85	38	27	38	1.4	1.0
E_{10}/M/1	1	.75	22	17.2	23	1.3	1.0
M/E_2/1	.5	.5	12.6	8	13.0	1.6	1.0
		.75	61	40	62	1.5	1.0
		.85	200	120	182	1.7	1.1
	1	.25	1.8	1.3	2.2	1.4	.8
		.5	6.4	4	6.5	1.6	1.0
		.75	29	20	31	1.5	.9
		.85	95	60	91	1.6	1.0
		.9	200	140	211	1.4	.9
	2	.75	13.7	10	15.5	1.4	.9
		.85	47	30	46	1.6	1.0

Table 3.2 (continued)

System	μ	ρ	τ_{exp}	τ_N	τ_R	τ_{exp}/τ_N	τ_{exp}/τ_R
$M/E_3/1$	.5	.5	10.5	6.7	11.5	1.6	.9
		.75	53	40	55	1.3	1.0
		.85	159	105	162	1.5	1.0
	1	.25	1.5	1.0	2.0	1.5	.8
		.5	5.3	3.3	5.8	1.6	.9
		.75	26	17.3	28	1.5	.9
		.85	79	53	81	1.5	1.0
		.9	178	123	188	1.4	.9
	2	.75	13.2	8.7	13.8	1.5	1.0
		.85	39	26	41	1.5	1.0
$M/E_4/1$	.5	.5	10.1	6	10.8	1.7	.9
		.75	50	32	52	1.6	1.0
		.85	164	98	152	1.7	1.1
	1	.5	4.8	3	5.4	1.6	.9
		.75	24	16	26	1.5	.9
		.85	82	49	76	1.7	1.1
		.9	172	115	176	1.5	1.0
	2	.75	12.0	7	12.9	1.7	.9
		.85	38	24	38	1.6	1.0
$M/E_{10}/1$	1	.75	20	13.6	23	1.5	.9
$E_2/E_2/1$	1	.5	4.2	3	4.3	1.4	1.0
		.75	189	14	21	1.4	.9
		.85	52	41	61	1.3	.9
M/D/1	.5	.5	7.4	4	8.6	1.9	.9
		.75	40	24	41	1.7	1.0
		.85	133	76	122	1.8	1.1
		.9	278	180	281	1.5	1.0
	1	.75	20	12	21	1.7	1.0
		.85	66	38	61	1.7	1.1
		.9	127	90	141	1.4	.9

System	α	μ_1	μ_2	ρ	τ_{exp}	τ_N	τ_R	τ_{exp}/τ_N	τ_{exp}/τ_R
$M/H_2/1$	.2	1	2	.75	28	18.9	28	1.5	1.0
	.4			.75	33	22	32	1.5	1.0
	.6			.75	33	25	36	1.3	.9
	.8			.75	40	27	39	1.5	1.0
	.2	1	.5	.75	72	53	78	1.4	.9
	.4			.75	71	50	72	1.4	1.0
$M/\hat{G}/1^*$	1/6	1	1	.8	115	84	129	1.4	.9
	1/6	1	3	.3	2.3	1.5	2.6	1.5	.9

*where the service time, s, has pdf

$$f_s(s_o) = \beta\mu_1 e^{-\mu_1 s_o} + (1-\beta)\,\frac{\mu_2^3}{2}\, s_o^{\,2}\, e^{-\mu_2 s_o}\,, \quad s_o \geq 0.$$

three terms $(1 + \sqrt{\rho})$, $(C_a^2 + C_s^2)$, and $(1 - \rho)$ and the value of the constant, 1/2.7. We did not use this more formal approach to modify (3.24), as the methods used throughout this chapter (e.g., plotting curves, then reading slopes) are rough; in addition, any practical use of the time constant to characterize system behavior will be at best approximate since transient decay can be approximated by an exponential function only for large t.

The experimental results presented in Section 3.1 suggest that, for systems which begin at rest, Q(t) eventually decays in an exponential manner and that, initially, decay is faster than this exponential function. Therefore, τ_R can be used to obtain an upper bound on the amount of time required for a system which begins at rest to reach equilibrium. After a length of time equal to $\gamma\ \tau_R, \gamma \geq 0$, at most $e^{-\gamma}$ of the transients remain.

3.3 Summary

With the empirical results of this chapter, we feel it is reasonable to conjecture that, for many types of infinite-capacity, single-queue, single-server queueing systems which begin at rest, the expected queue length can be approximated by

$$Q(t) \cong Q(\infty) + Ae^{-t/\tau_R}\ , \quad t \geq \hat{t}\ , \tag{3.25a}$$

for some A < 0,

where

$$\tau_R = \frac{(1+\sqrt{\rho})^2 \left(C_a^2+C_s^2\right)}{2.7\ \mu(1-\rho)^2}\ , \tag{3.25b}$$

and $Q(\infty)$ is calculated by one of the methods discussed earlier (Section 3.1). We have not attempted to determine a reliable method for estimating $\hat{t}$, but visual examination of our experimental $\log|Q(\infty)-Q(t)|$ curves suggests that

$\hat{t}$ is less than $2\,\tau_R$. This is certainly one area for further work.

We also have not determined an expression for the parameter A, but due to the fact that Q(t) decays initially at a faster rate than the eventual exponential function, Q(t) is bounded below by (3.25) with $A = -Q(\infty)$. An important implication of this is that the amount of time until the transient effects become negligible is bounded above by exponential decay with parameter τ_R.

The random variables specifying the interarrival and service times of our test cases were either deterministic or composed of some combination of independent Erlang random variables. We suspect that our approximation (3.25) might hold for more general queueing systems but we do not yet have the evidence needed for confirmation.

A closer approximation for Q(t) is given by

$$Q(t) = Q(\infty) + \tilde{A}e^{-t/\tau} + g(t), \qquad t \geq 0 \ , \tag{3.26}$$

where the function g(t) goes to 0 at a faster rate than the exponential function $\tilde{A}e^{-t/\tau}$ as t approaches infinity. Expression (3.26) behaves in a purely exponential manner for large t, but the g(t) term could be chosen to account for the nonexponential initial behavior of our experimental $\log|Q(\infty)-Q(t)|$ curves. In particular, (3.26) agrees with the exact solution for a finite-capacity M/M/1 system. However, addition of the g(t) term in expression (3.26) is probably of limited usefulness in practice, as it provides no extra information on the amount of time needed to reach equilibrium; the exponential decay is the determining factor.

CHAPTER 4

TRANSIENT BEHAVIOR OF THE EXPECTED QUEUE LENGTH OF INFINITE-CAPACITY, SINGLE-QUEUE, SINGLE-SERVER SYSTEMS WHICH DO NOT BEGIN AT REST

In Chapter 3, we examined the transient behavior of the expected queue length, Q(t), for two classes of queueing systems which begin at rest. Our empirical results confirmed that for Markovian systems and for those deterministic systems in which the embedded chain is a first-order Markov process, the transient part of Q(t) decays in an approximately exponential manner for large t. In addition, transients appear to decay initially at a rate faster than that of the exponential function that eventually dominates. Thus, the time to equilibrium can be bounded from above by exponential decay. The time constant for this exponential function can be estimated through use of our closed-form expression (3.24). We now examine the effects of initial conditions other than the empty state on the transient response of Q(t) for the same types of queueing systems considered in Chapter 3.

Intuitively, for large t, any effects due specifically to the initial conditions will be negligible and thus, independent of the state in which the system begins, transient decay will eventually be of the same functional form as when the system starts from rest. Therefore, we expect that for large t, Q(t) will decay in an approximately exponential manner with a time constant dependent only on the queueing system at hand, not on the initial conditions. We also expect that for small t, the effect of transients on the behavior of systems which begin near equilibrium will become negligible more quickly than those of a corresponding system which is initially empty or heavily saturated.

In this chapter, we examine these issues in an attempt to obtain a closed-form expression that can be used to predict the amount of time required for a system to effectively reach equilibrium as a function of the initial state of the system. In Section 4.1, we examine systems with deterministic initial conditions, i.e., $P_i(0)=1$ for some i, and $P_j(0)=0$ for all $j \neq i$. First, we empirically confirm that, independent of initial conditions, transient effects decay in an approximately exponential manner for large t. Then, we examine the transient behavior for small t and attempt to derive bounds for the time to equilibrium as a function of the deterministic initial conditions. Finally, in Section 4.2, we consider the effect of probabilistic initial conditions on these results.

As before, we use Q(t), the expected queue length at time t, as our representative measure of system behavior. We restrict the discussion to Markovian systems and those partially deterministic systems that have an embedded chain that is a first-order Markov process. The analysis is empirical and solutions to the systems examined are obtained through the numerical technique discussed in Section 2.2.

4.1 Systems with Deterministic Specification of Initial Conditions

We begin this analysis with a discussion of queueing systems in which at time t=0, for some specific i (i=0,1,...), there are i customers in the system with probability 1. (If i>0, this implies that Q(0) = i-1). Thus, initial conditions are specified in a deterministic manner.

4.1.1 Characteristics of the Functional Form

In the previous chapter we found that, for systems which are initially idle, the expected queue length Q(t), can be approximated by

$$Q(t) \overset{\sim}{=} Q(\infty) + Ae^{-t/\tau}, \; t \geq \hat{t} \quad , \tag{4.1}$$

where A, $\tau > 0$, and $\hat{t} > 0$ are parameters that are system-specific. We now test the validity of this expression for the same classes of queueing systems but with a range of deterministic initial conditions.

To begin, we fix the traffic intensity (ρ=.75, μ=1) and the types of the interarrival and service time distributions (both negative exponential). Note that for this M/M/1 system, $Q(\infty) = 2.25$.

Figures 4.1-4.4 illustrate $\log|Q(\infty) - Q(t)|$ versus t for nine examples with i ranging from 0 to 34. In each case, after an initial period, $\log|Q(\infty) - Q(t)|$ varies in a linear manner with t. This implies that for large t, Q(t) approaches $Q(\infty)$ in an approximately exponential manner regardless of the number of customers in the system at time t=0. Thus, for any specified deterministic initial conditions, there exist parameters A, τ, and $\hat{t}$ such that expression (4.1) is valid.

For small t, however, it is clear that the functional form of Q(t) varies greatly with i. To study this behavior in more detail we also examine plots of Q(t) versus time. In Figures 4.5 - 4.8 we illustrate Q(t) versus t for the same nine cases.

Consider Figures 4.1 and 4.5. They cover cases in which Q(0)=0. For these systems, Q(t) approaches $Q(\infty)$ in a monotonic manner. In addition, as $\log|Q(\infty) - Q(t)|$ is initially convex, we can see that Q(t) is bounded from below by (4.1) with $A = -Q(\infty)$ and $\hat{t}$=0. This is the case we studied previously in Chapter 3.

We now examine the transient response when the initial number of customers in queue is in the range from one to slightly above $Q(\infty)$. Three such cases are illustrated in Figures 4.2 and 4.6. In each example

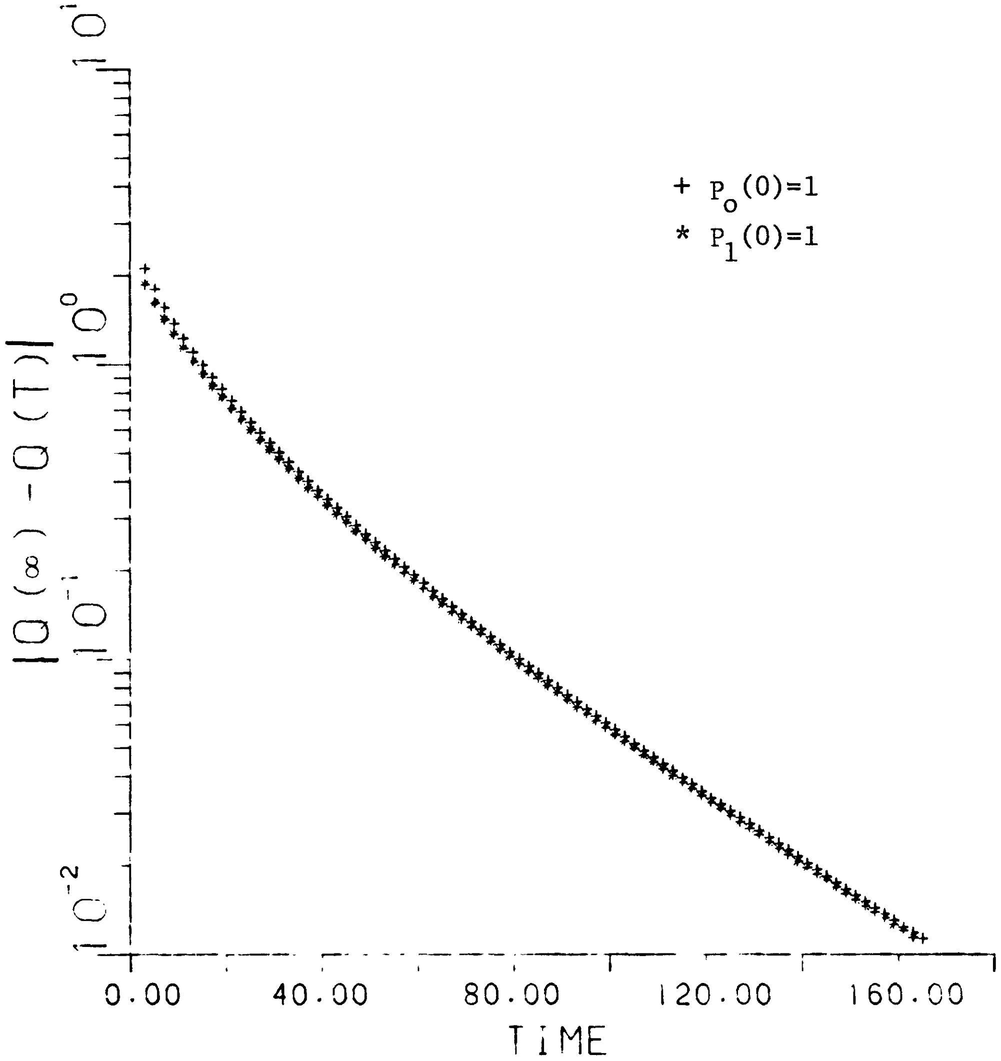


Figure 4.1: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an M/M/1 System With $\rho=.75$ and $\mu=1$; $P_0(0)=1$ and $P_1(0)=1$

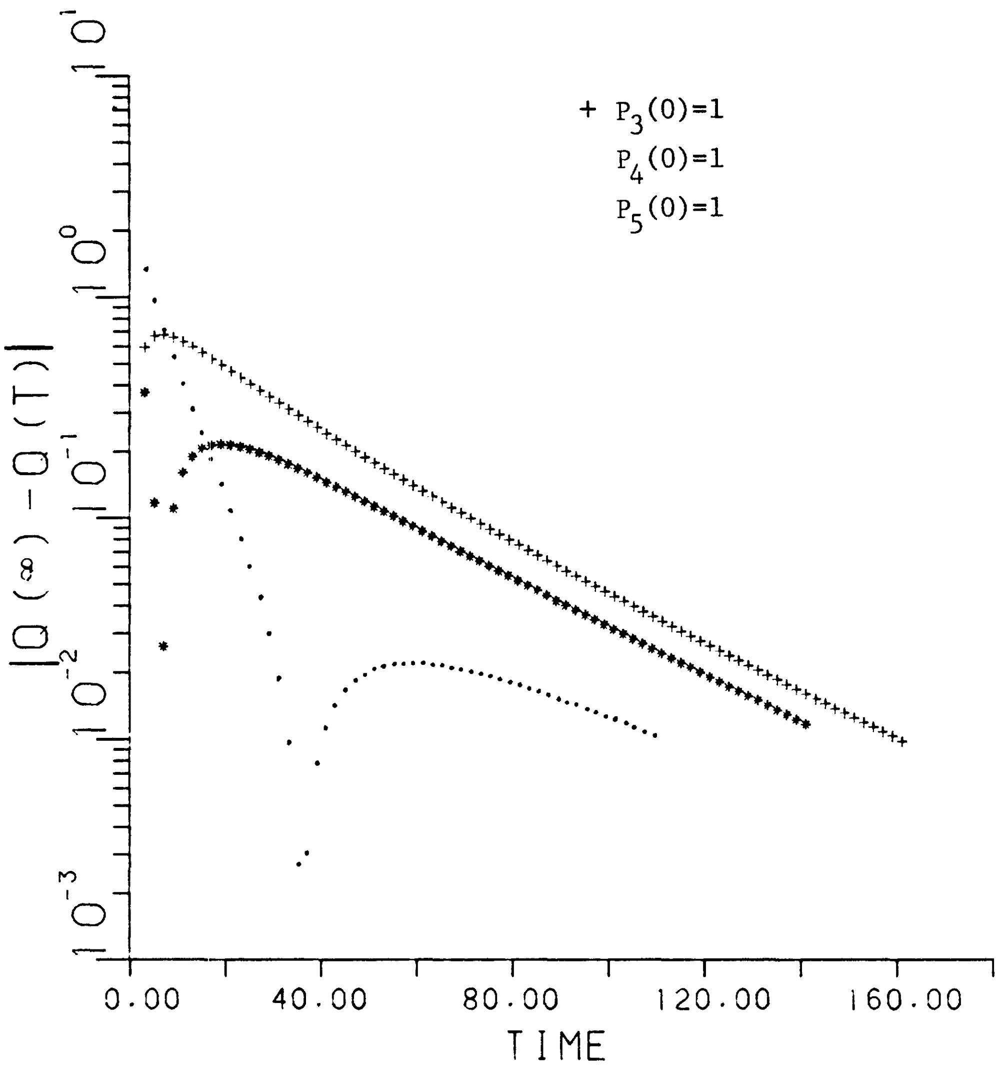


Figure 4.2: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an M/M/1 System With $\rho=.75$ and $\mu=1$; $P_3(0)=1$, $P_4(0)=1$, and $P_5(0)=1$

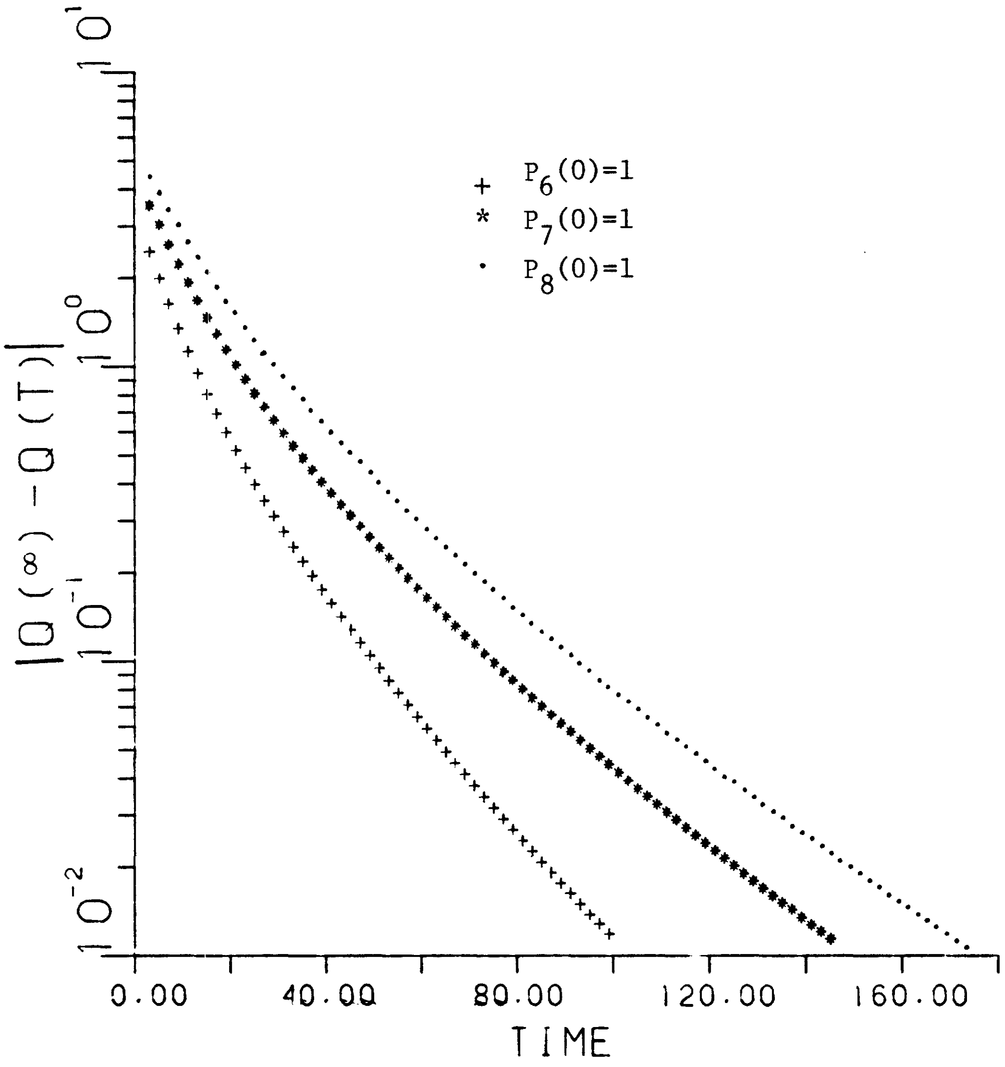


Figure 4.3: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an M/M/1 System With $\rho=.75$ and $\mu=1$; $P_6(0)=1$, $P_7(0)=1$, and $P_8(0)=1$

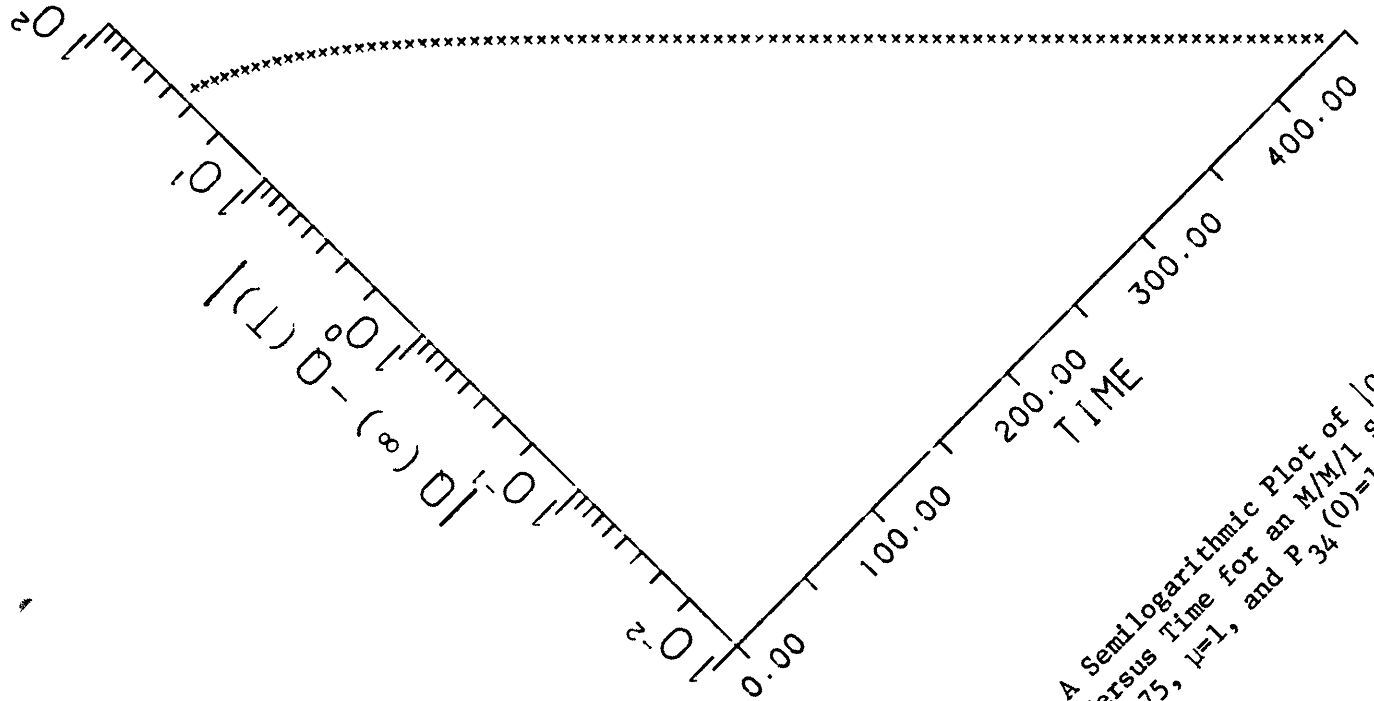


Figure 4.4: A Semilogarithmic Plot of $|Q(\infty) - Q(T)|$ Versus Time for an M/M/1 System With $\rho=.75$, $\mu=1$, and $P_{34}(0)=1$

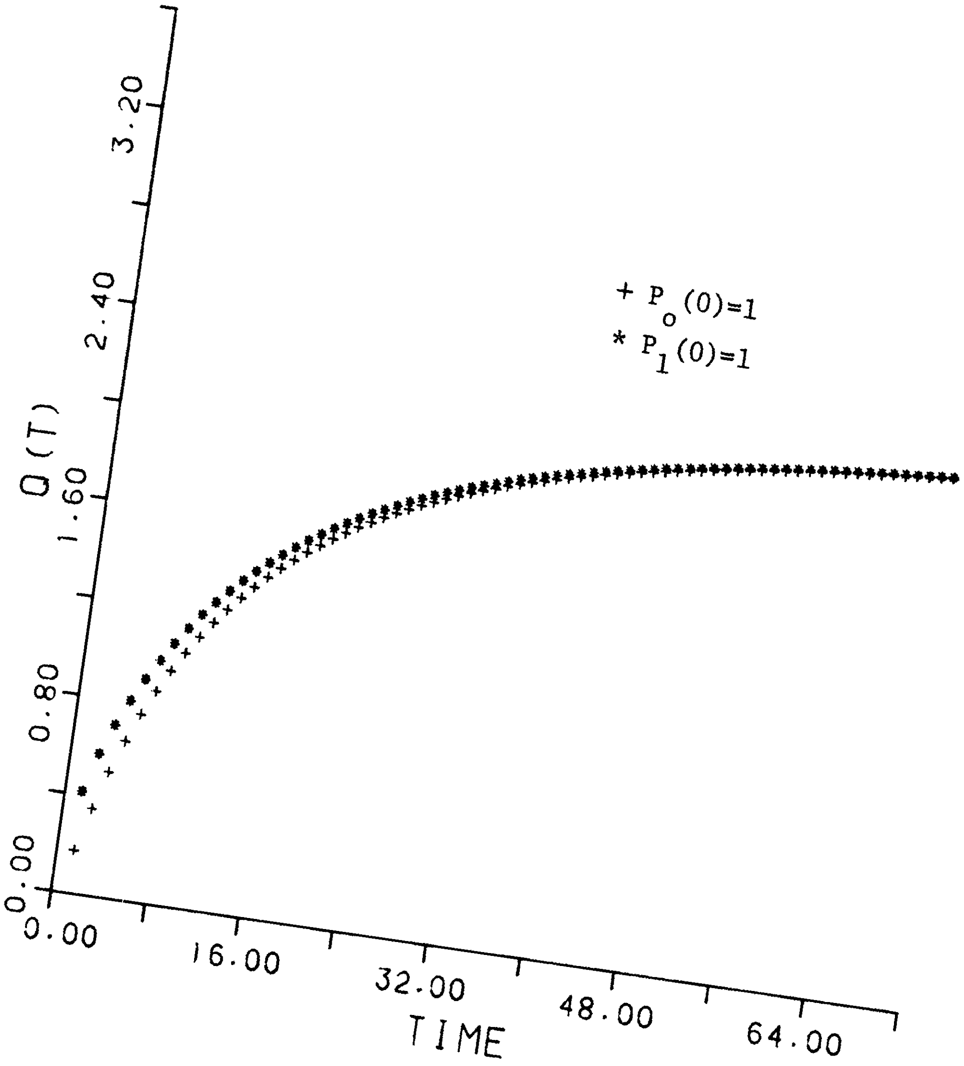


Figure 4.5: The Expected Queue Length Versus Time for an M/M/1 System With ρ=.75 and μ=1; $P_0(0)=1$ and $P_1(0)=1$

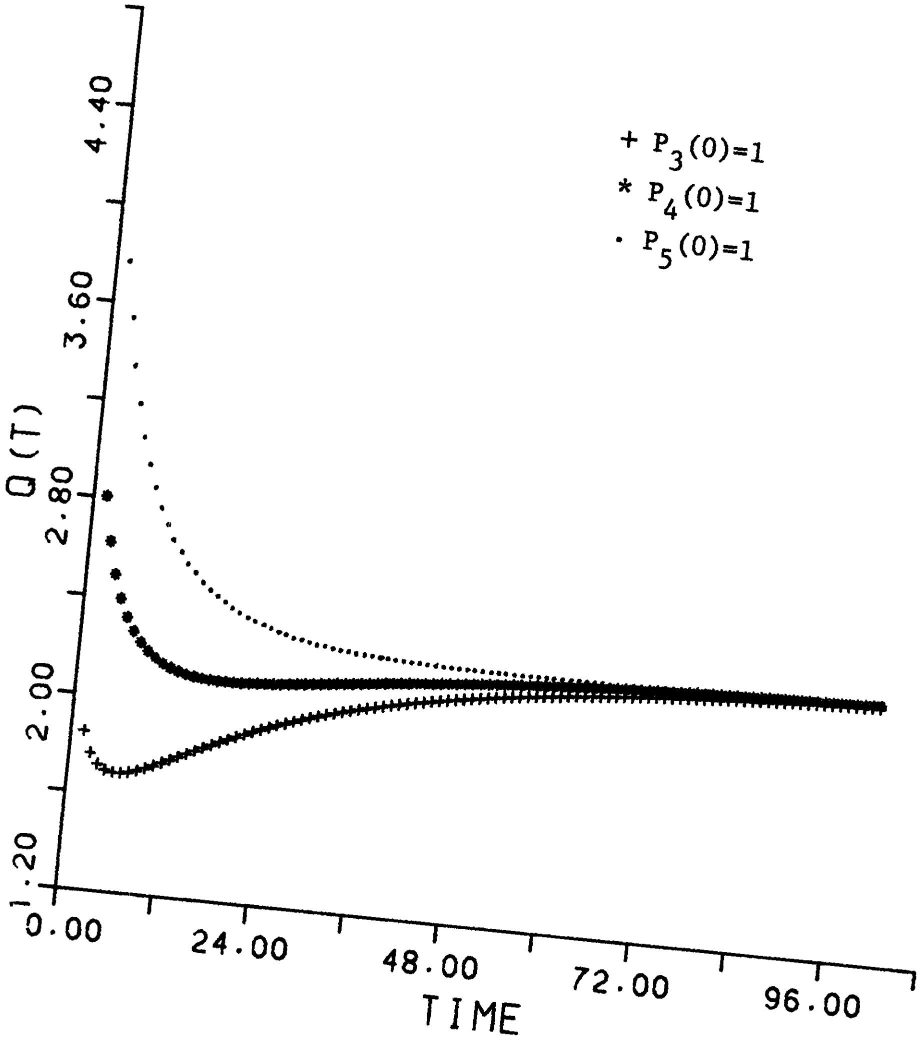


Figure 4.6: The Expected Queue Length Versus Time for an M/M/1 System With $\rho=.75$ and $\mu=1$; $P_3(0)=1$, $P_4(0)=1$, and $P_5(0)=1$

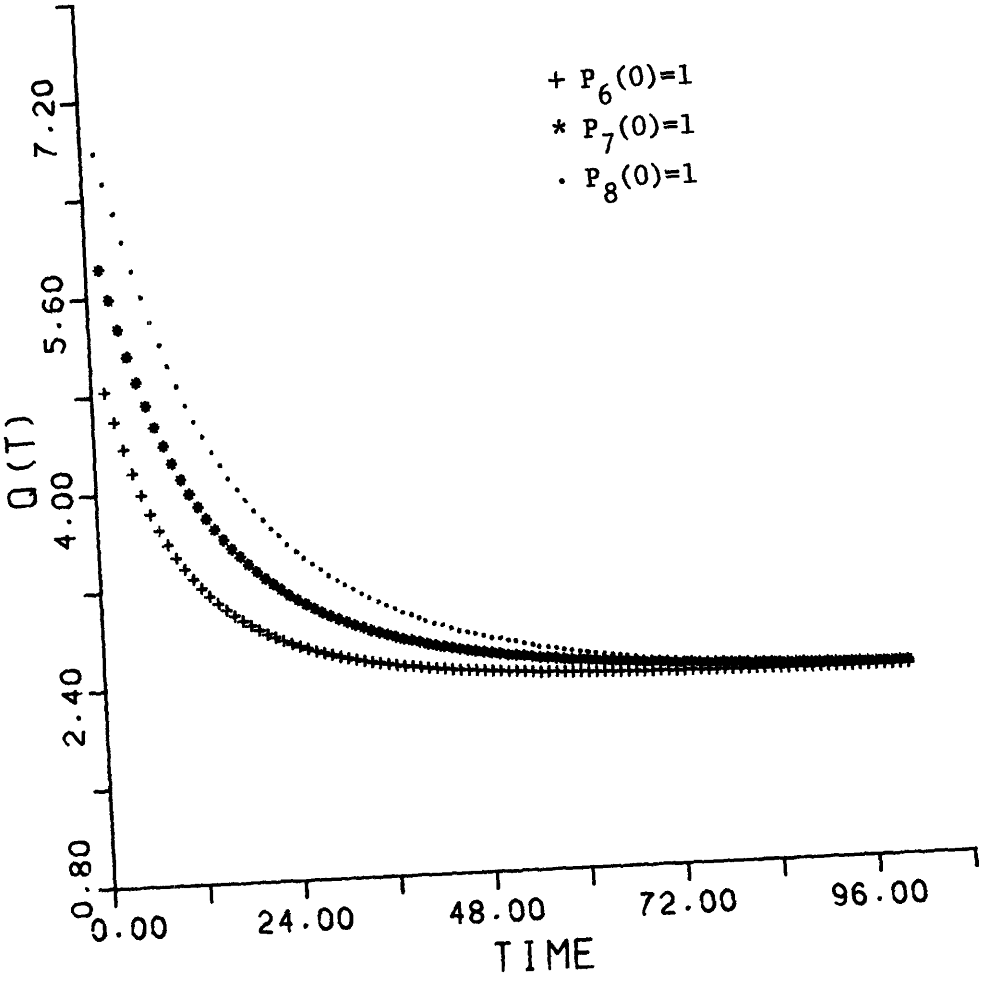


Figure 4.7: The Expected Queue Length Versus Time for an M/M/1 System With $\rho=.75$ and $\mu=1$; $P_6(0)=1$, $P_7(0)=1$ and $P_8(0)=1$

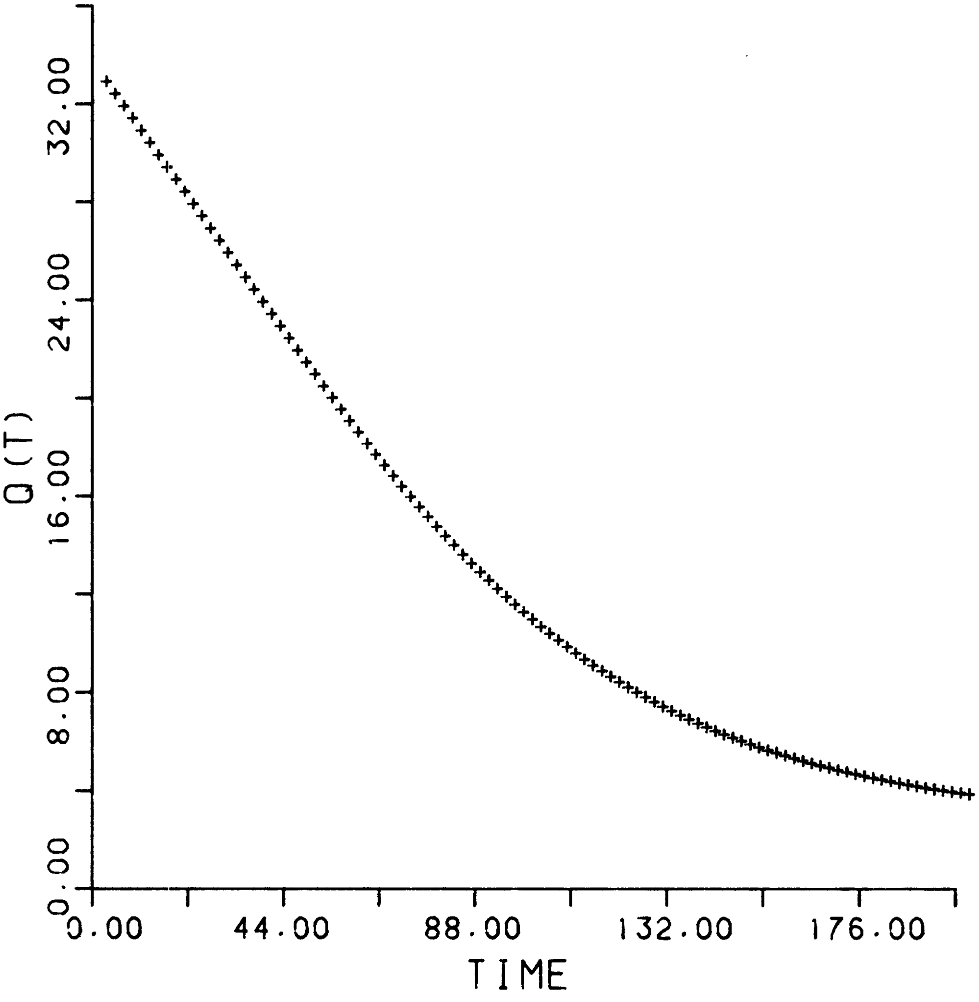


Figure 4.8: The Expected Queue Length Versus Time for an M/M/1 System With $\rho=.75$, $\mu=1$, and $P_{34}(0)=1$

Q(t) decreases initially to a minimum value, then increases monotonically to Q(∞). Note that there is never more than one critical point for t>0.

The initial drop in Q(t) is due to the fact that $\mu>\lambda$. (This is a requirement for ergodicity in an infinite-capacity, single-queue, single-server system.) For example, consider the case in which there are two customers in queue and a customer in service at t=0(i=3). Since $\mu>\lambda$, the server is likely to complete the current service before a new customer enters the system. In fact, since both the interarrival and service time distributions are given by negative exponential random variables, for this system

$$P\begin{pmatrix}\text{current service is completed}\\ \text{before first customer arrival}\end{pmatrix} = \frac{\mu}{\mu+\lambda} = .57 \qquad (4.2)$$

Thus, there is a .57 chance that the first change of state will be caused by a service completion, and thus result in the system containing two customers (equivalently, one customer in queue).

This argument is valid for any i>1. If, however, ρ is large and Q(0) << Q(∞), the initial decrease in Q(t) will be small in magnitude and occur within a very short time period. As an example, we consider an M/M/1 system with ρ=.9 and μ=1 (Q(∞) = 8.1). If there are initially 2 customers in the system with probability 1, Q(t) reaches a global minimum after roughly .12 time units. This behavior is illustrated in Figure 4.9. Thus, if Q(t) is observed only once or twice per time unit, Q(t) will appear to be a monotonically increasing function of t.

In Figures 4.3 and 4.7, we present examples of systems in which Q(0) is large enough to prevent the initial decrease in Q(t) from "overshooting" Q(∞), but not large enough to heavily saturate the system. In each case,

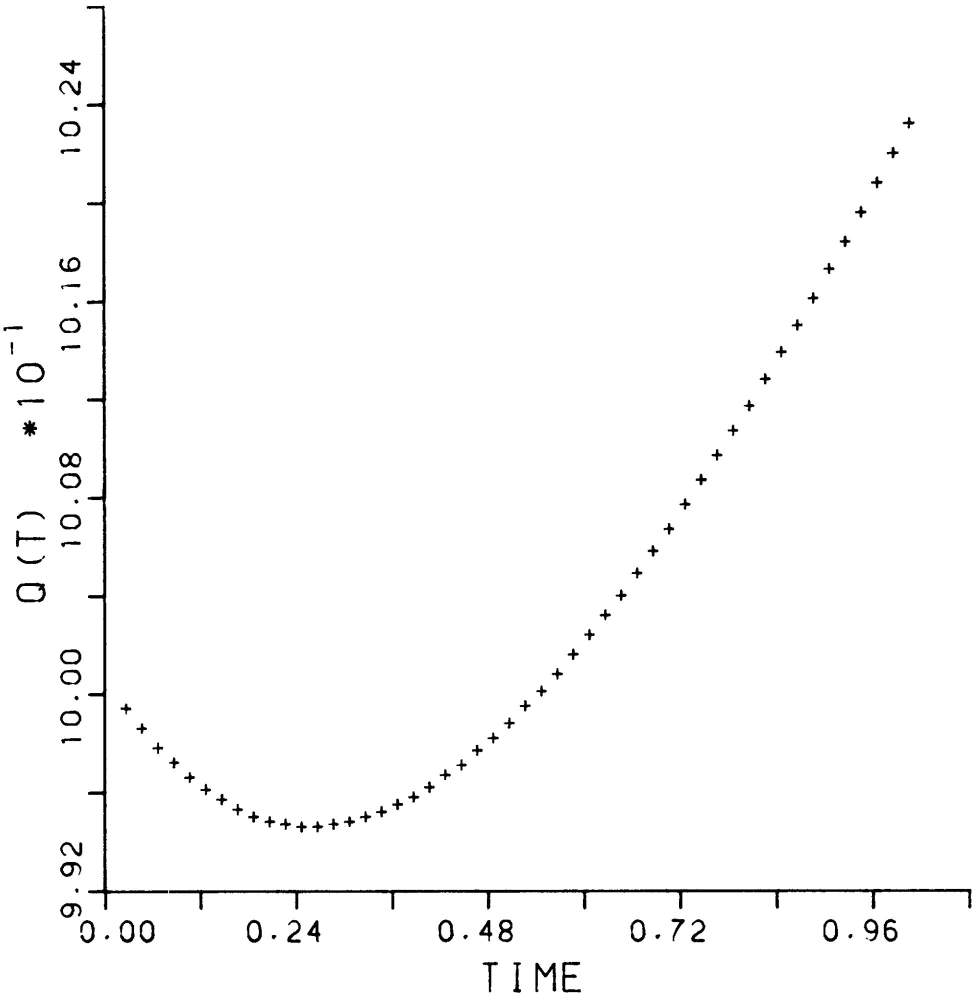


Figure 4.9: The Expected Queue Length Versus Time for an M/M/1 System With $\rho=.9$, $\mu=1$, and $P_2(0)=1$

Q(t) approaches Q(∞) from above in a monotonic manner. The initial convexity of the $\log|Q(\infty) - Q(t)|$ curve implies transient decay which is faster than that of the eventually dominating exponential function. Thus, there exist positive values of the parameters A and τ, such that Q(t) is bounded from above by expression (4.1) with $\hat{\iota}=0$.

Finally, we examine the case in which the system is initially heavily saturated. In Figures 4.4 and 4.8 we illustrate the transient response of this M/M/1 system when 33 customers are in queue at time t=0. Note that Q(t) decreases initially in a linear manner with slope -.25.

This behavior can be explained intuitively by noting that the dominating feature is the presence initially of 34 customers in the system. Clearly, for a time the server will be working continuously to clear the system of these excess customers. On average, μ customers per unit time will leave the system. Concurrently, an average of λ customers per unit time will enter the system. Thus, for small t,

$$Q(t) \overset{\sim}{=} Q(0) - (\mu-\lambda)t, \qquad (4.3)$$

i.e., Q(t) decreases in a linear manner with slope $-(\mu-\lambda)$ (which for this M/M/1 system equals -.25). Once the system recovers from its oversaturated condition, the probabilistic nature of the interarrival and service time distributions will again dominate and Q(t) will no longer be linear but will approach its equilibrium value in an approximately exponential manner.

To check the validity of this intuitive argument, for an M/M/1 system with $\mu=1$ we specify initial conditions of 34 customers in the system, and vary λ. Both the traffic intensity ρ and the steady-state expected queue length Q(∞) increase with λ. Since in all cases the initial number of customers in the system is fixed at 34, if ρ is large the system begins relatively closer to its equilibrium state. Therefore, we expect $Q(t)$ to decay in an approximately exponential manner at an earlier time than a system with a smaller value of ρ. We illustrate this behavior in Figure 4.10 through plots of $Q(t)$ versus t for an M/M/1 system with $\rho = .75, .8, .85$ and $.9$.

Note also that as expected, for small t the slope of $Q(t)$ is $-(\mu-\lambda)$. Figure 4.11 illustrates $\log|Q(\infty) - Q(t)|$ versus t for these cases, confirming that $Q(t)$ decays in an approximately exponential manner for large t.

The preceding results suggest that, with regard to deterministic initial conditions, four categories of transient response can be identified:

(i) If $Q(0) = 0$, $Q(t)$ increases monotonically to $Q(\infty)$.
In addition, $Q(t)$ is bounded from below by (4.1) with $A = -Q(\infty)$ and $\hat{t}=0$.

(ii) If $1 \leq Q(0) \leq \tilde{Q}_1$, for some $\tilde{Q}_1 > Q\ (\infty)$, $Q(t)$ will not be a monotonic function of t but will initially decrease to a global minimum before approaching $Q(\infty)$ in a monotonic manner. If $Q(\infty) \leq Q(0) \leq \tilde{Q}_1$, this initial decrease in $Q(t)$ will overshoot the equilibrium value $Q(\infty)$. For large t, $Q(t)$ can be approximated by expression (4.1) with a negative value of A.

(iii) If $\tilde{Q}_1 < Q(0) \leq \tilde{Q}_2$, for some $\tilde{Q}_2 > \tilde{Q}_1 > Q(\infty)$, $Q(t)$ will decrease monotonically to $Q(\infty)$, bounded from above by expression (4.1) with a positive value of A and $\hat{t}=0$.

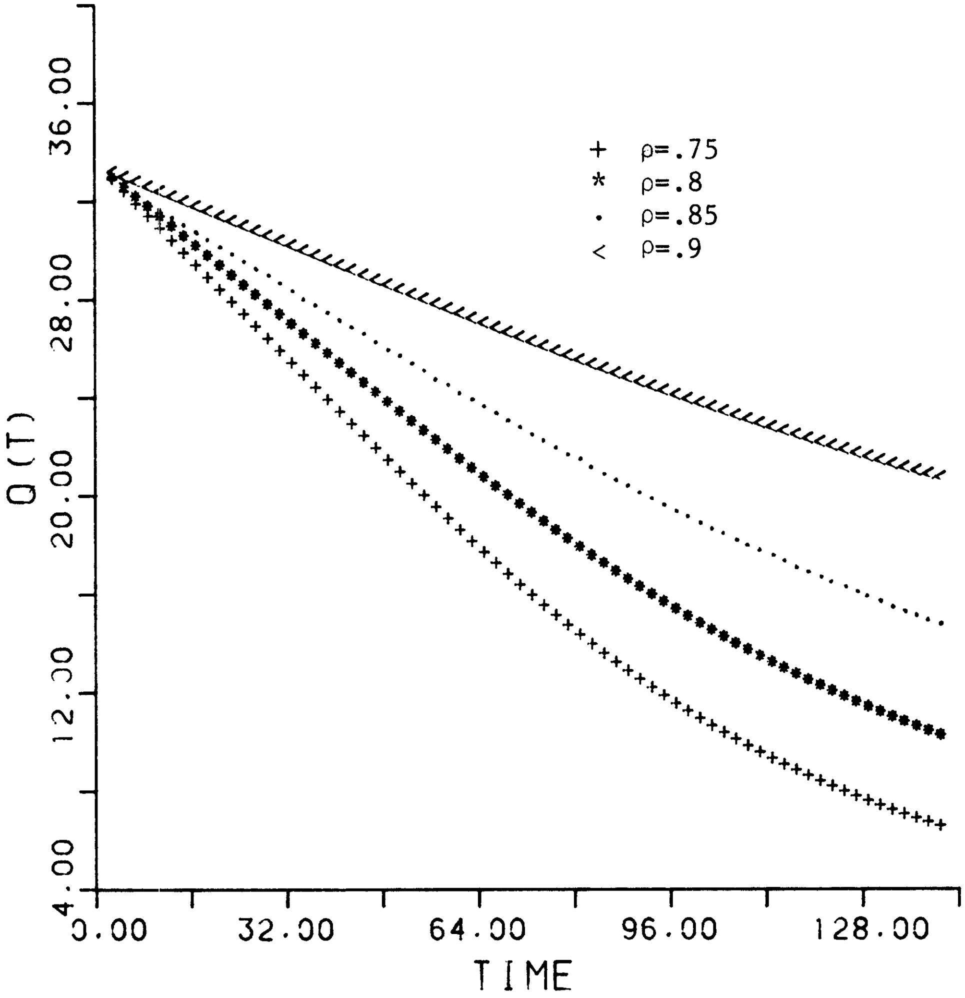


Figure 4.10: The Expected Queue Length Versus Time for an M/M/1 System With μ=1 and $P_{34}(0)$=1; ρ=.75, .8, .85, and .9

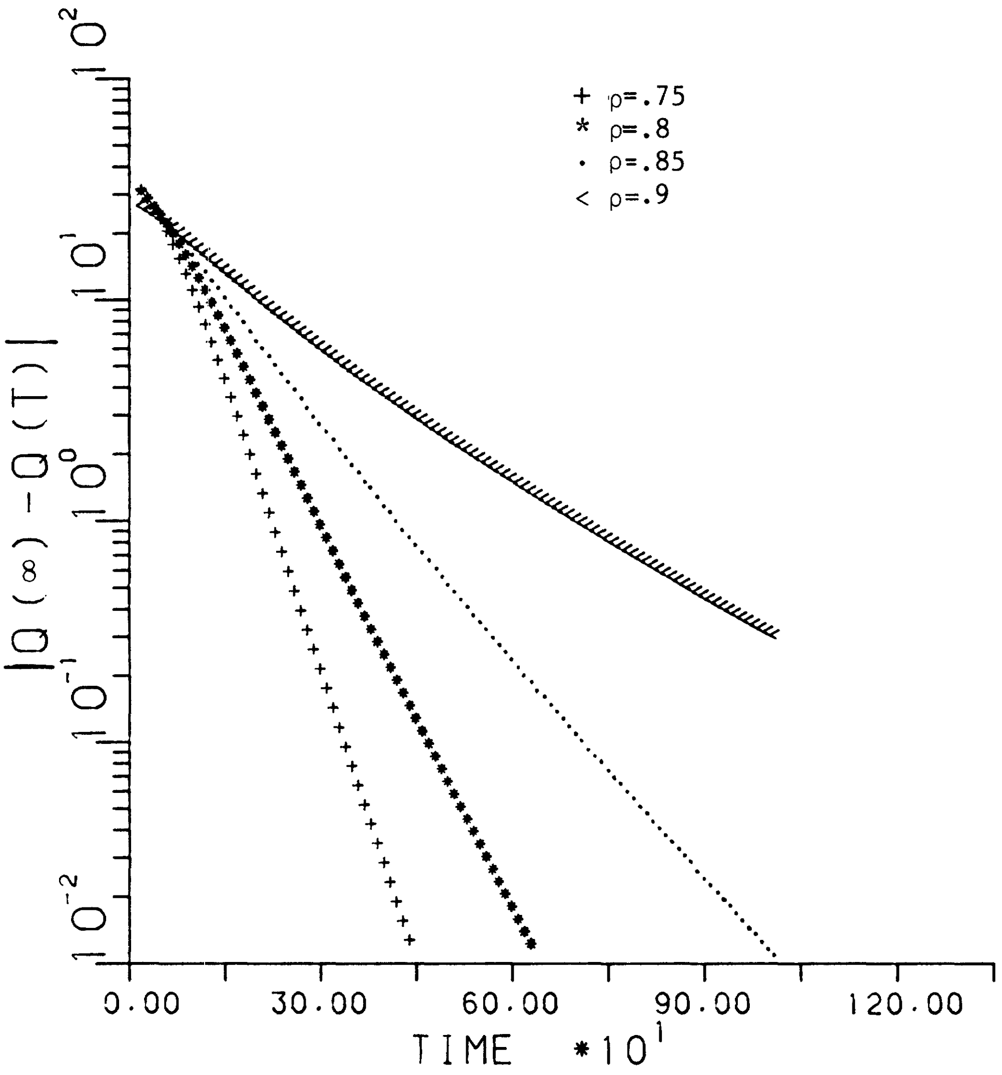


Figure 4.11: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an M/M/1 System With $\mu=1$ and $P_{34}(0)=1$; $\rho=.75$, .8, .85, and .9

(iv) If $\tilde{Q}_2 < Q(0)$, initially Q(t) will decrease linearly, then Q(t) will behave as in category (iii).

The next group of figures illustrate that these results apply also to non-M/M/1 systems, specifically to the more general class of Markovian systems. First, we consider an $M/H_2/1$ system with $\rho=.75$, $\alpha=.2$, $\mu_1=1$, and $\mu_2=2$. This system has $Q(\infty) = 2.5$. In Figures 4.12-4.15, we show examples of $\log|Q(\infty) - Q(t)|$ versus time for each of the four classes of initial conditions confirming that, for large t, Q(t) approaches $Q(\infty)$ in an approximately exponential manner.[1] Figures 4.16 and 4.17 are plots of Q(t) versus t. Figure 4.16 illustrates that if $1 \leq Q(0) \leq \tilde{Q}_1$ (where $\tilde{Q}_1$ is slightly larger than $Q(\infty)$), Q(t) overshoots its equilibrium value once, before monotonically approaching $Q(\infty)$. Figure 4.17 shows that if the system is initially saturated, Q(t) decreases in a linear manner with slope $-(\mu-\lambda)$ for small t.

Figures 4.18 - 4.23 illustrate the same behavior for an $E_3/M/1$ system with $\rho=.75$ and $\mu=1$ ($Q(\infty)=1.3479$). Figures 4.18 - 4.21 are graphs of $\log|Q(\infty)-Q(t)|$ for a range of initial conditions. All curves are approximately linear for large t indicating that after an initial time period Q(t) may be approximated by an exponential function. Figure 4.22 shows the "overshooting" characteristic of systems with initial conditions in category (ii) and Figure 4.23 shows the initial linear decrease (with slope $-(\mu-\lambda)$) of Q(t) for systems in category (iv).

4.1.2 Examination of the Time to Equilibrium

We now discuss the implications of these results with regard to the amount of time required for transient effects to become negligible.

[1] In each case, if there is a customer in service at time t=0, she is assumed to have an expected service time of $1/\mu_1$ time units.

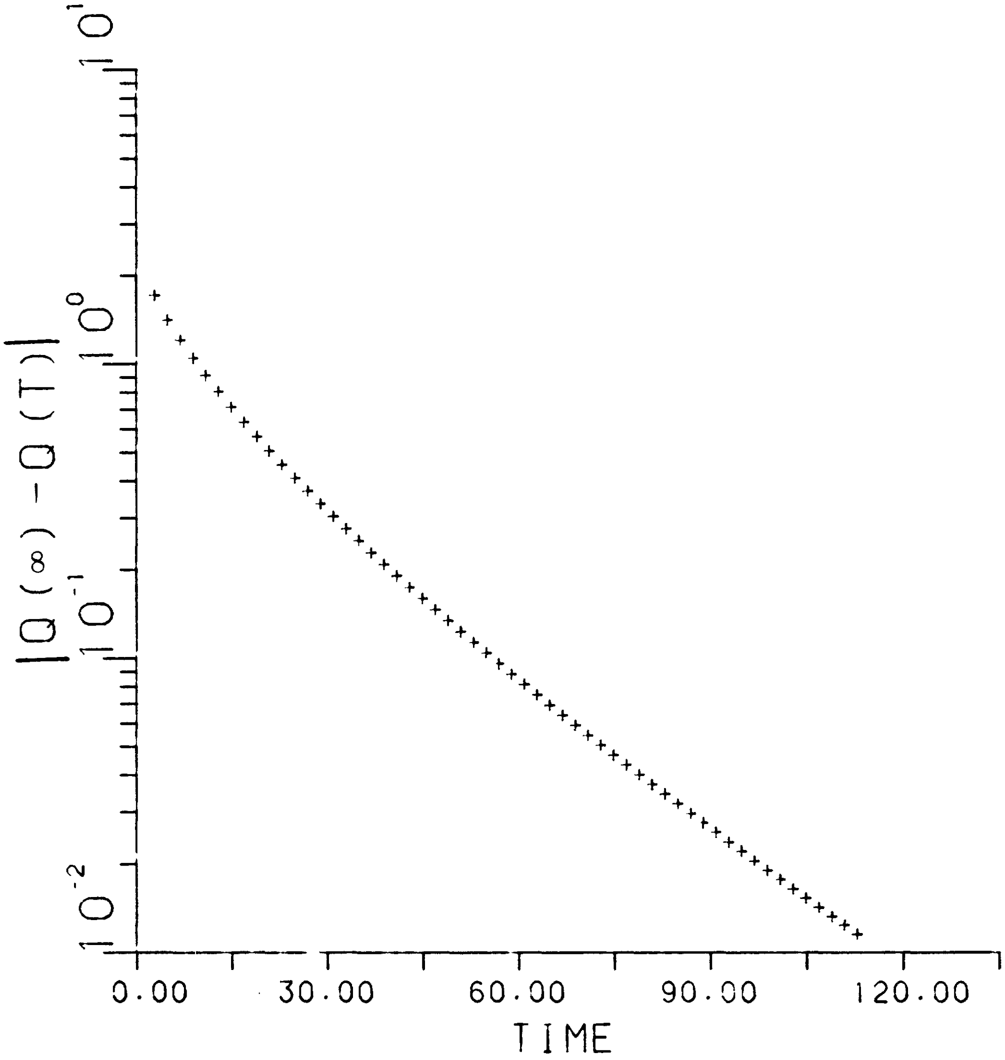


Figure 4.12: A Semilogarithmic Plot of $|Q(\infty) - Q(T)|$ Versus Time for an $M/H_2/1$ System With $\rho=.75$, $\alpha=.2$, $\mu_1=1$, $\mu_2=2$, and $P_1(0)=1$

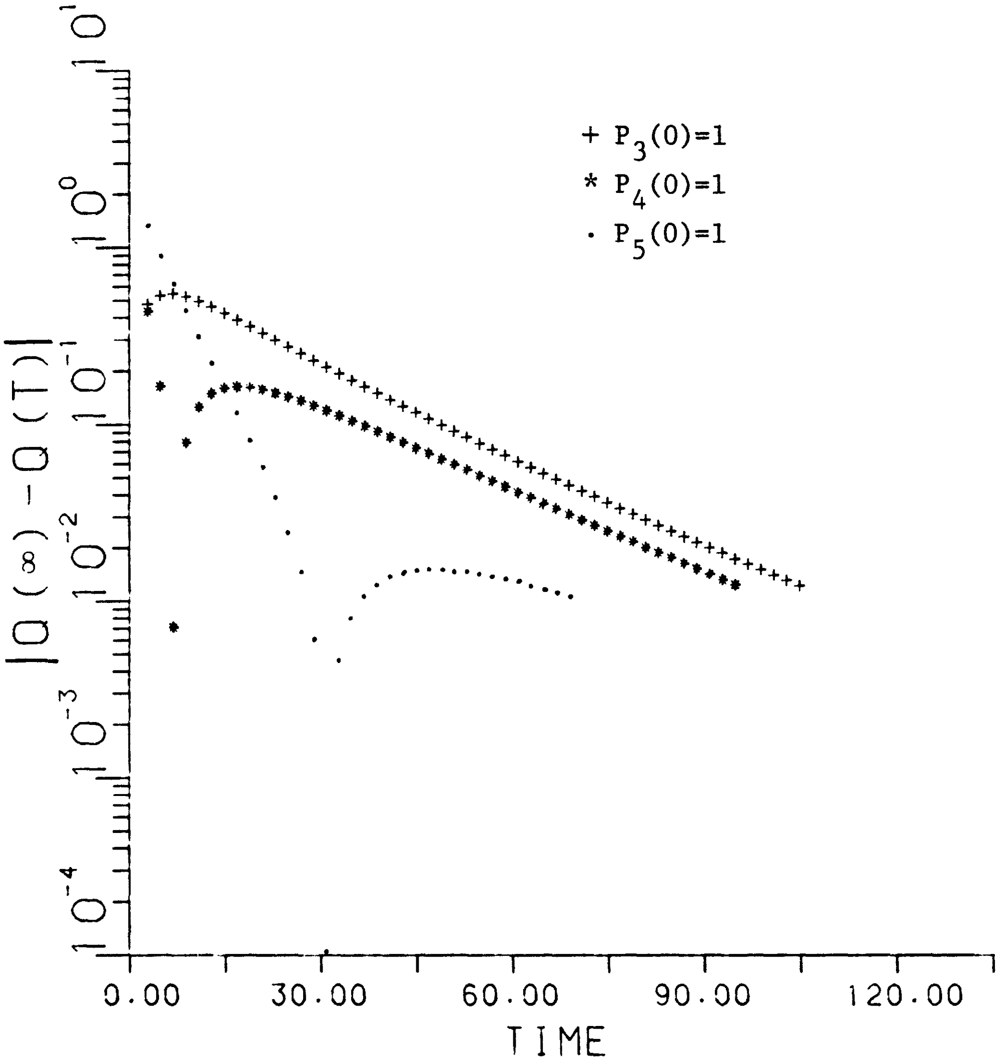


Figure 4.13: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an $M/H_2/1$ System With $\rho=.75$, $\alpha=.2$, $\mu_1=1$, and $\mu_2=2$; $P_3(0)=1$, $P_4(0)=1$, and $P_5(0)=1$

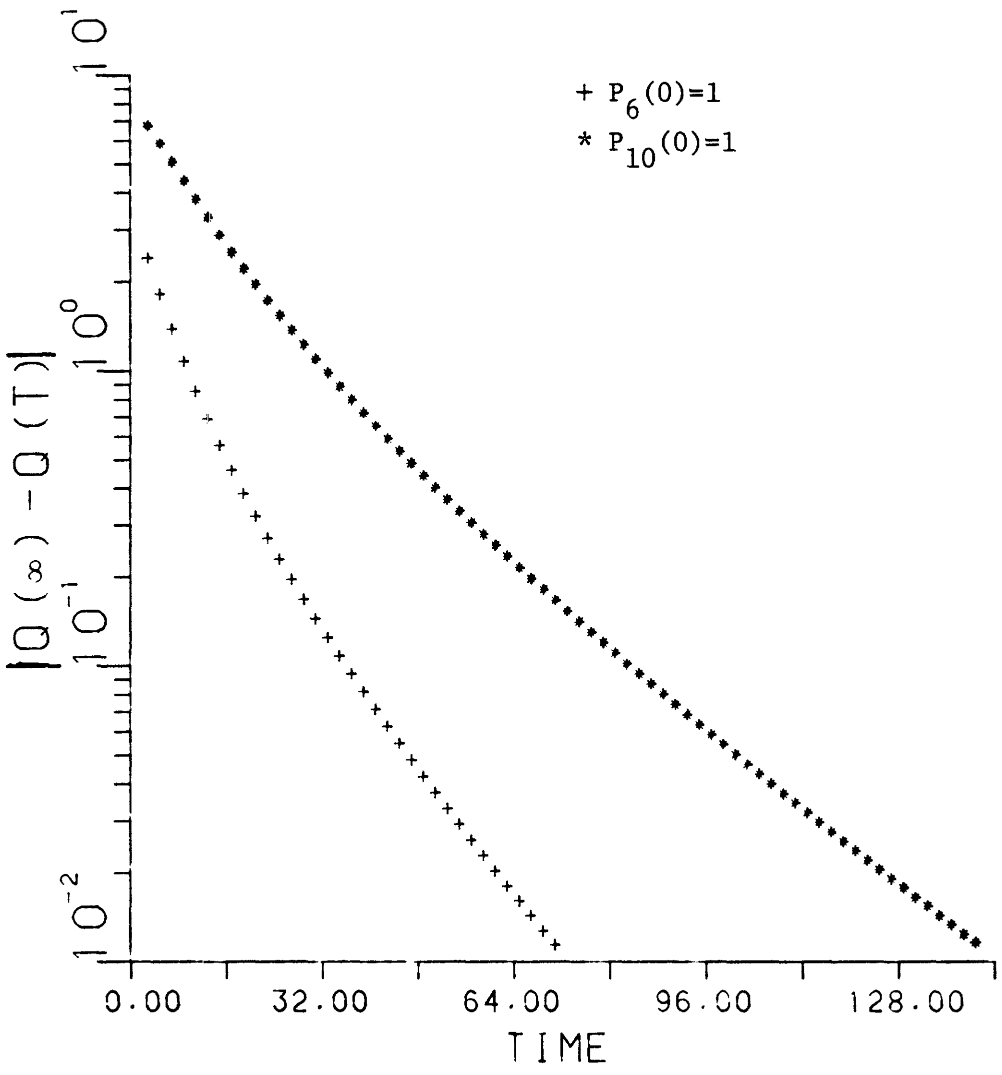


Figure 4.14: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an $M/H_2/1$ System With $\rho=.75$, $\alpha=.2$, $\mu_1=1$, and $\mu_2=2$; $P_6(0)=1$ and $P_{10}(0)=1$.

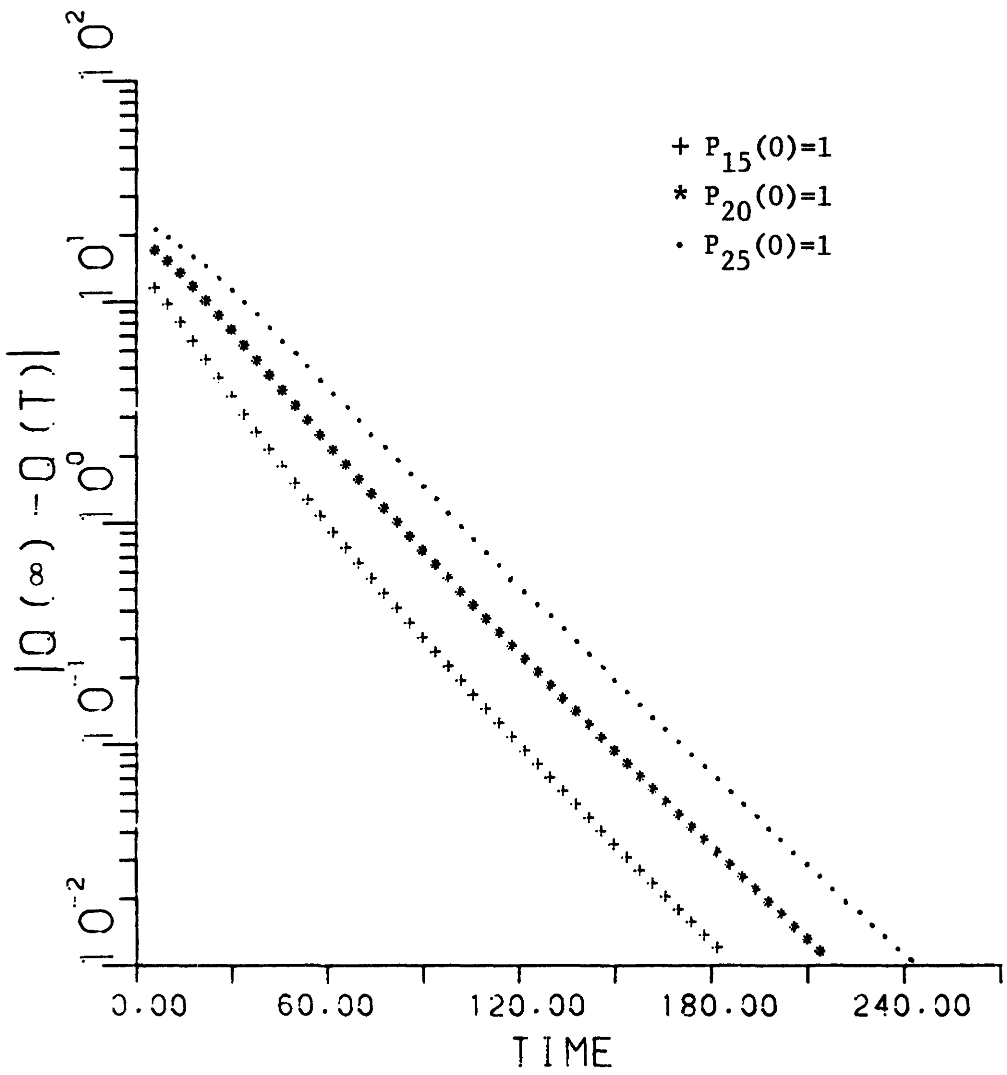


Figure 4.15: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an $M/H_2/1$ System With $\rho=.75$, $\alpha=.2$, $\mu_1=1$, and $\mu_2=2$; $P_{15}(0)=1$, $P_{20}(0)=1$, and $P_{25}(0)=1$

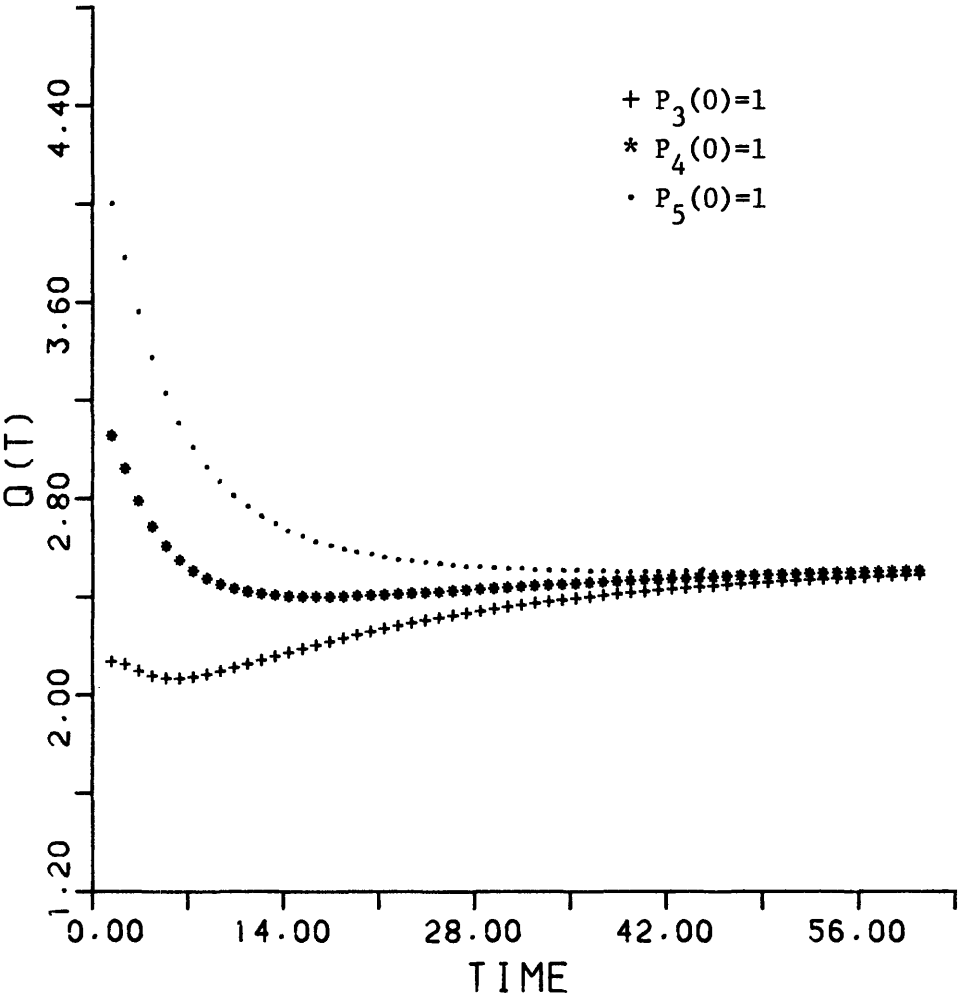


Figure 4.16: The Expected Queue Length Versus Time for an $M/H_2/1$ System With $\rho=.75$, $\alpha=.2$, $\mu_1=1$, and $\mu_2=2$: $P_3(0)=1$, $P_4(0)=1$, and $P_5(0)=1$

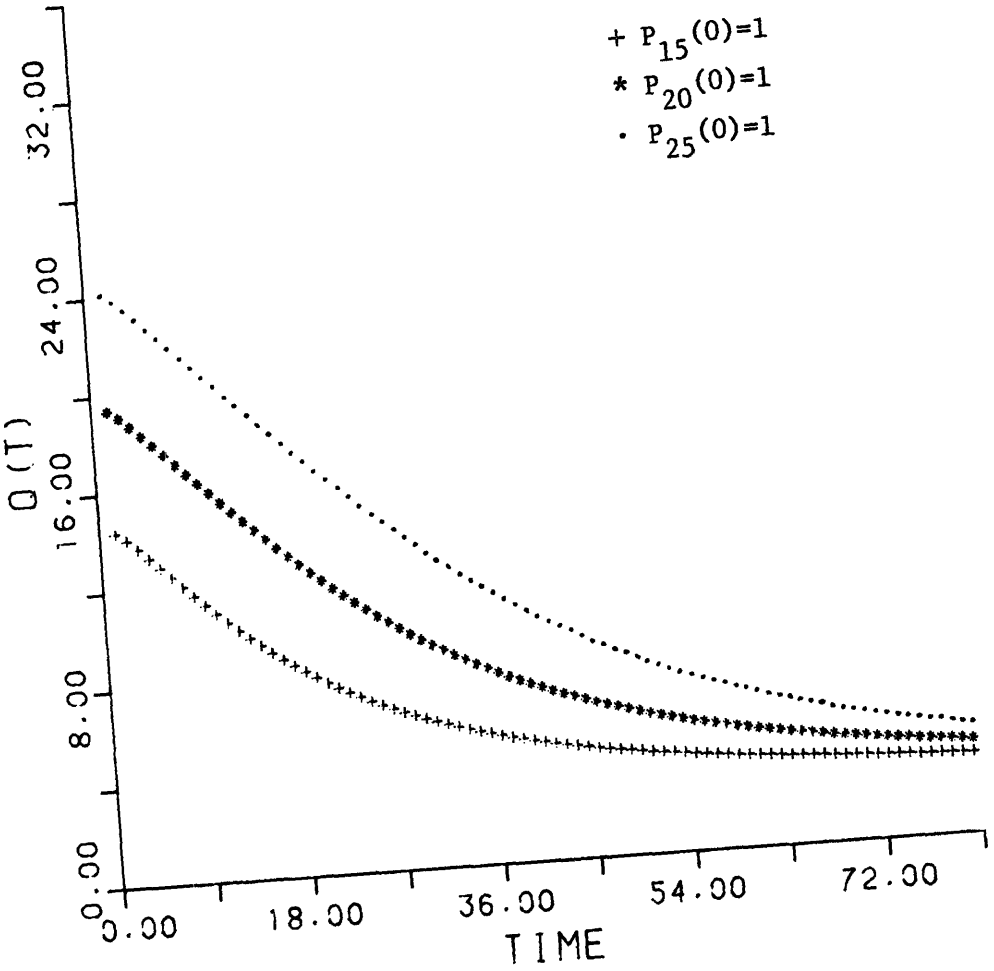


Figure 4.17: The Expected Queue Length Versus Time for an $M/H_2/1$ System With $\rho=.75$, $\alpha=.2$, $\mu_1=1$, and $\mu_2=2$; $P_{15}(0)=1$, $P_{20}(0)=1$, and $P_{25}(0)=1$

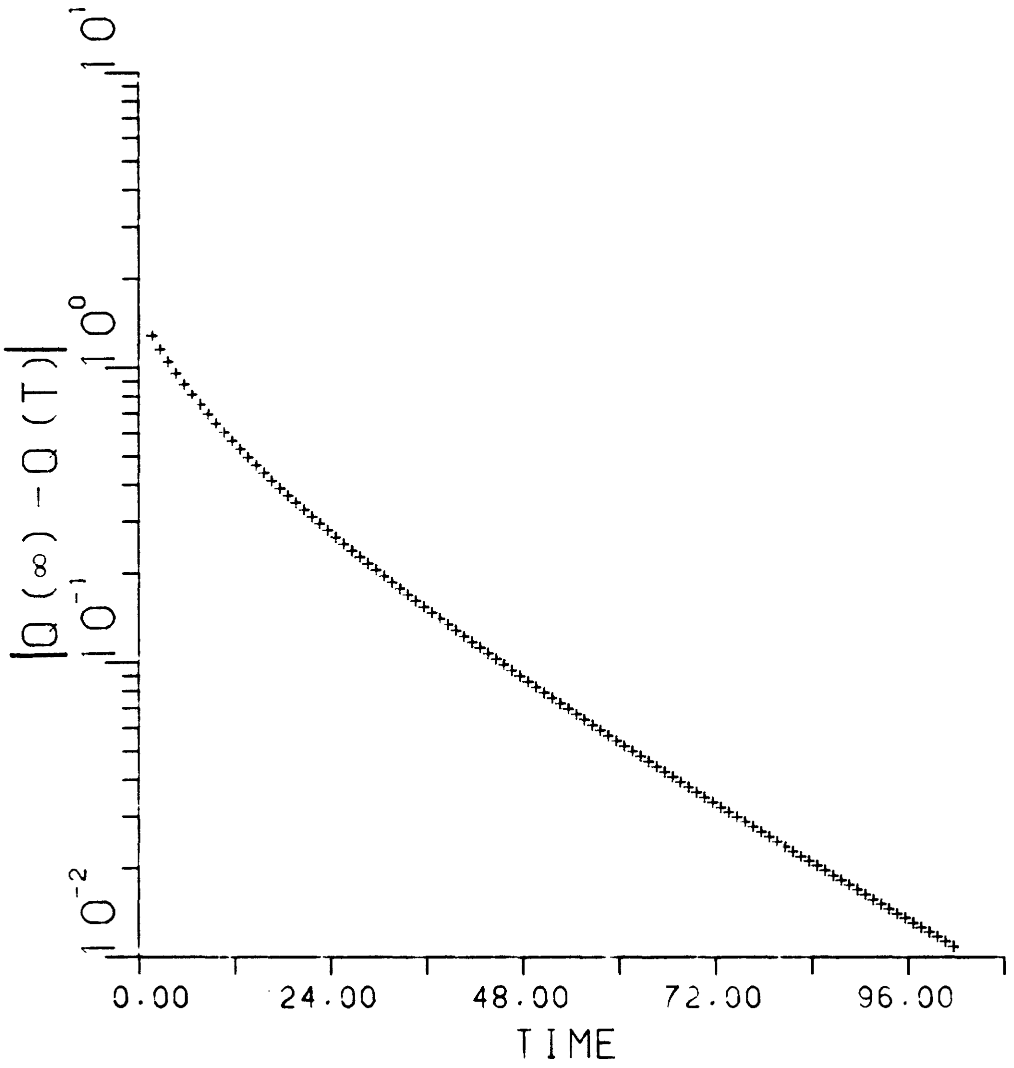


Figure 4.18: A Semilogarithmic Plot of $|Q(\infty) - Q(T)|$ Versus Time for an $E_3/M/1$ System With $\rho=.75$, $\mu=1$, and $P_1(0)=1$

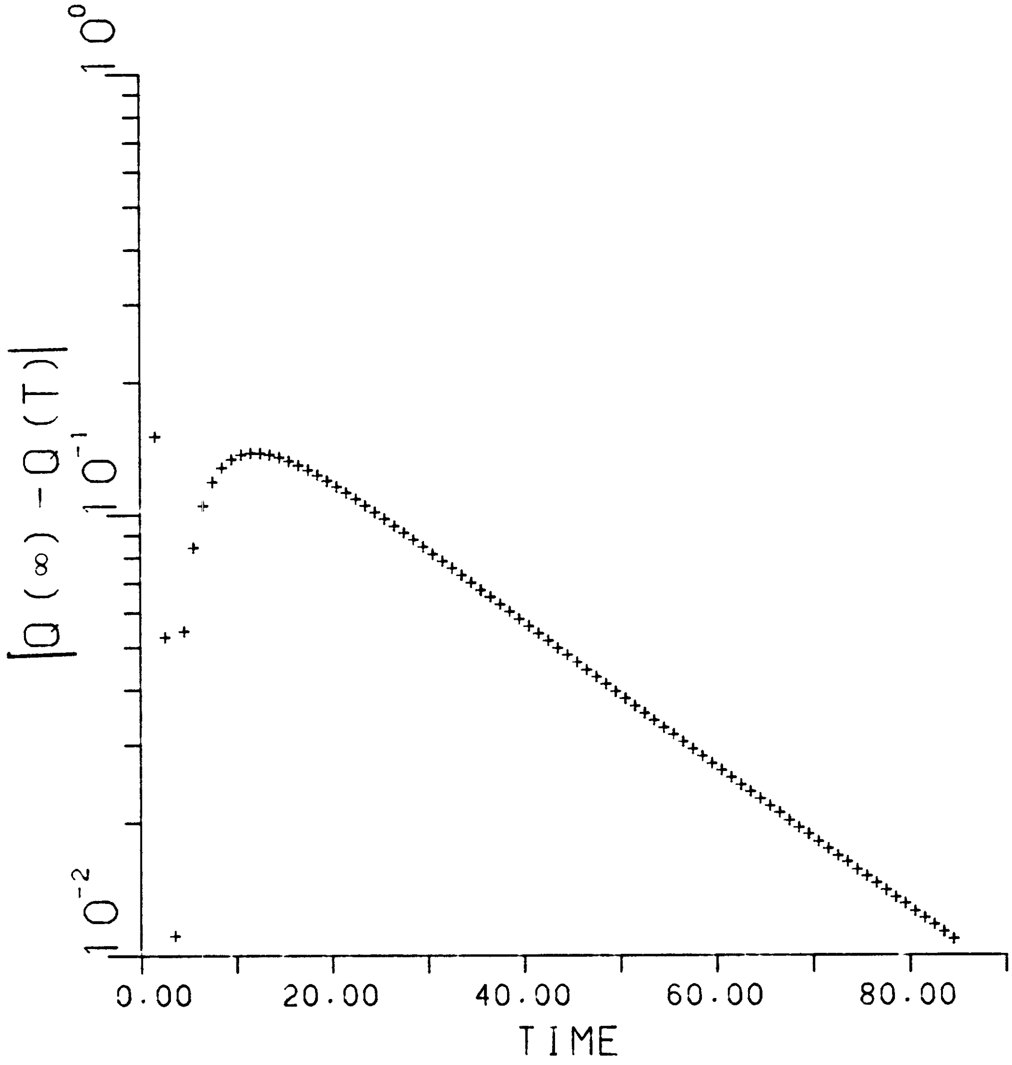


Figure 4.19: A Semilogarithmic Plot of |Q(∞) - Q(T)| Versus Time for an E_3/M/1 System With ρ=.75, μ=1, and $P_3(0)$=1

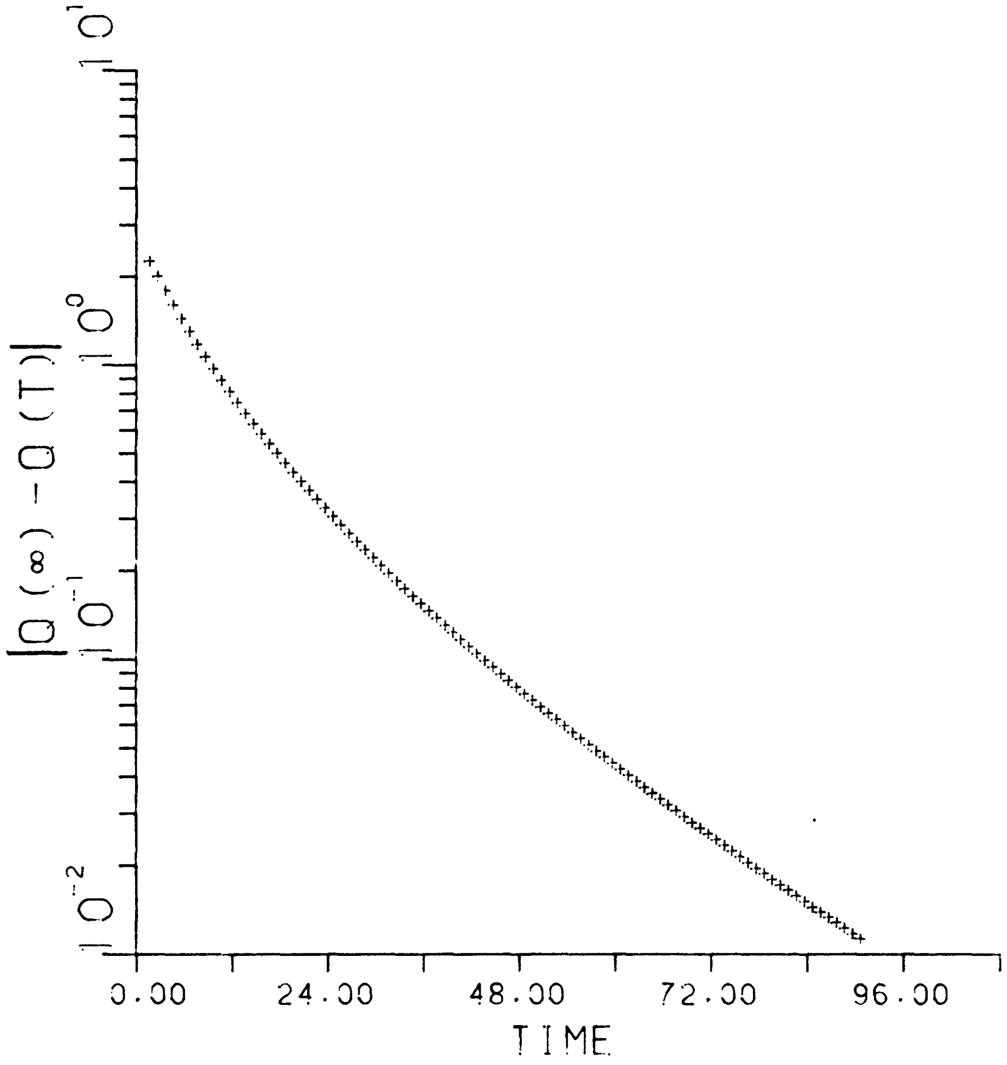


Figure 4.20: A Semilogarithmic Plot of $|Q(\infty) - Q(T)|$ Versus Time for an $E_3/M/1$ System With $\rho=.75$, $\mu=1$, and $P_5(0)=1$

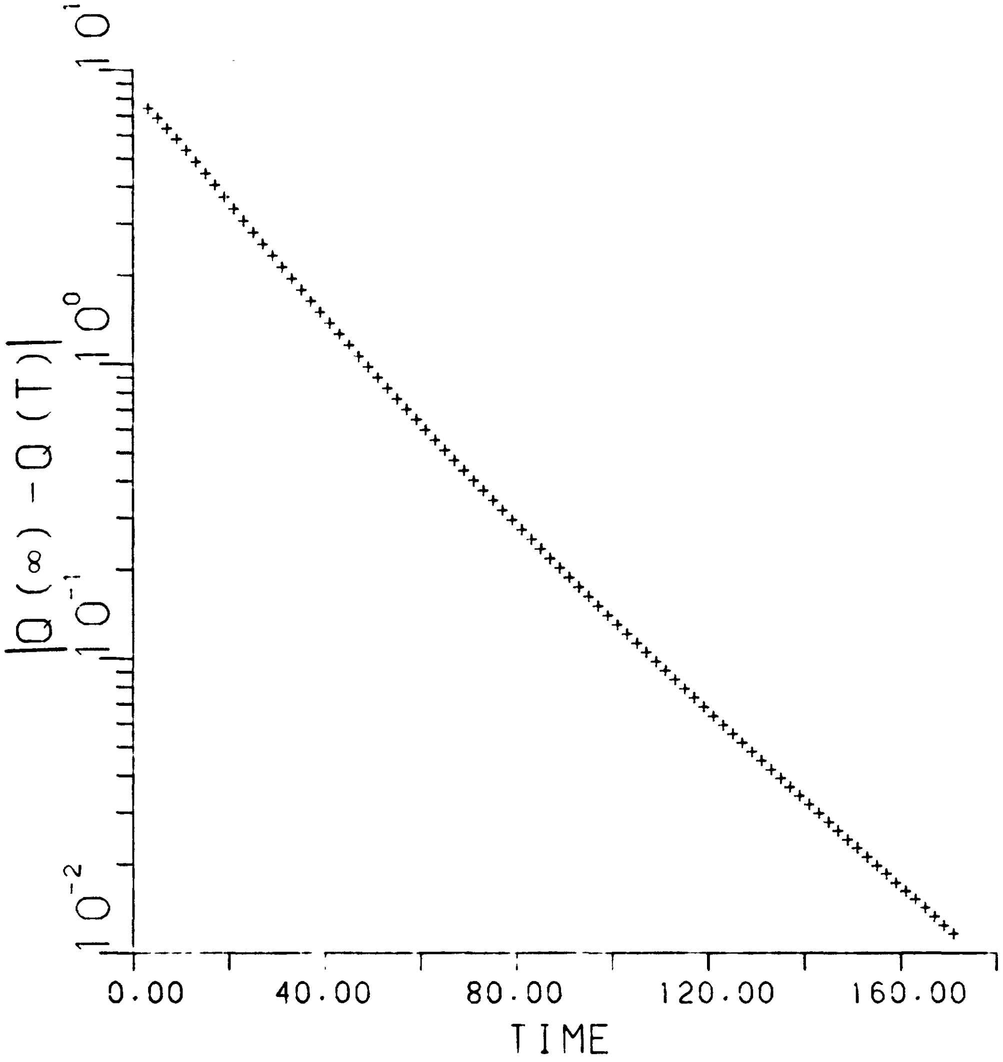


Figure 4.21: A Semilogarithmic Plot of $|Q(\infty) - Q(T)|$ Versus Time for an $E_3/M/1$ System With $\rho=.75$, $\mu=1$, and $P_{10}(0)=1$

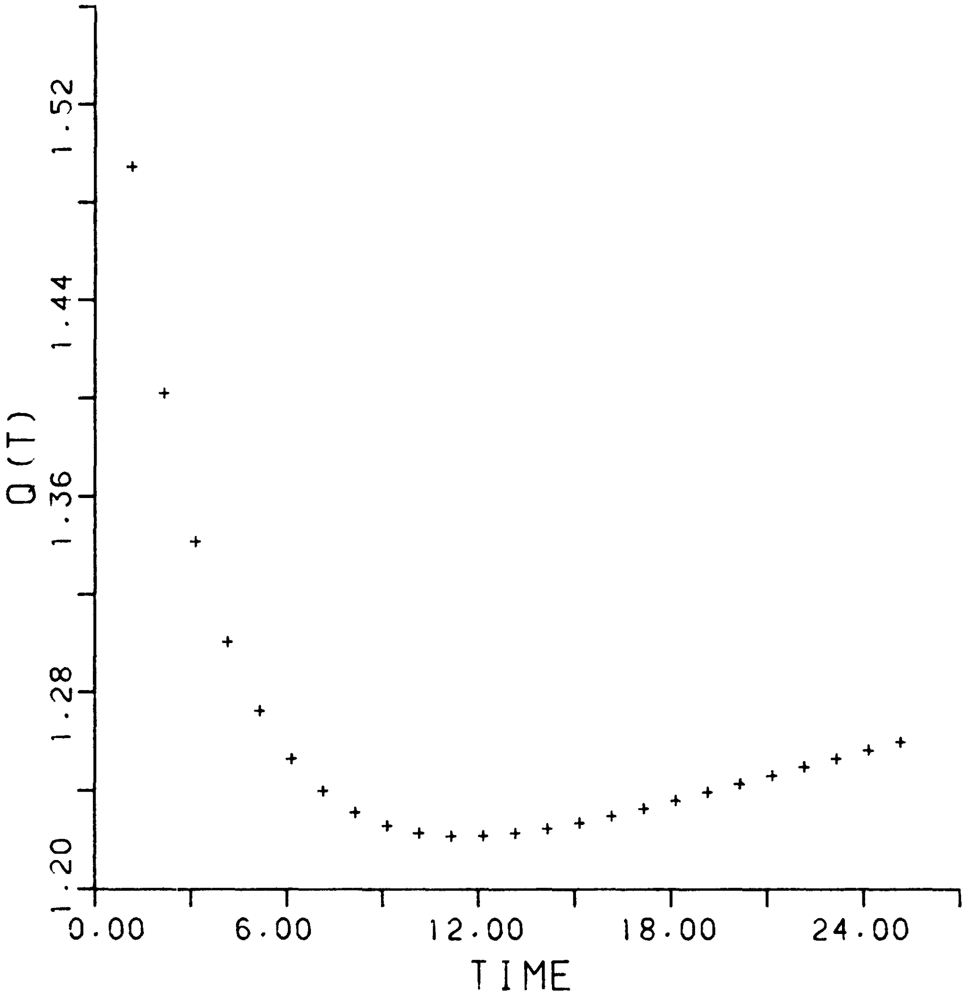


Figure 4.22: The Expected Queue Length Versus Time for an $E_3/M/1$ System With ρ=.75, μ=1, and $P_3(0)=1$

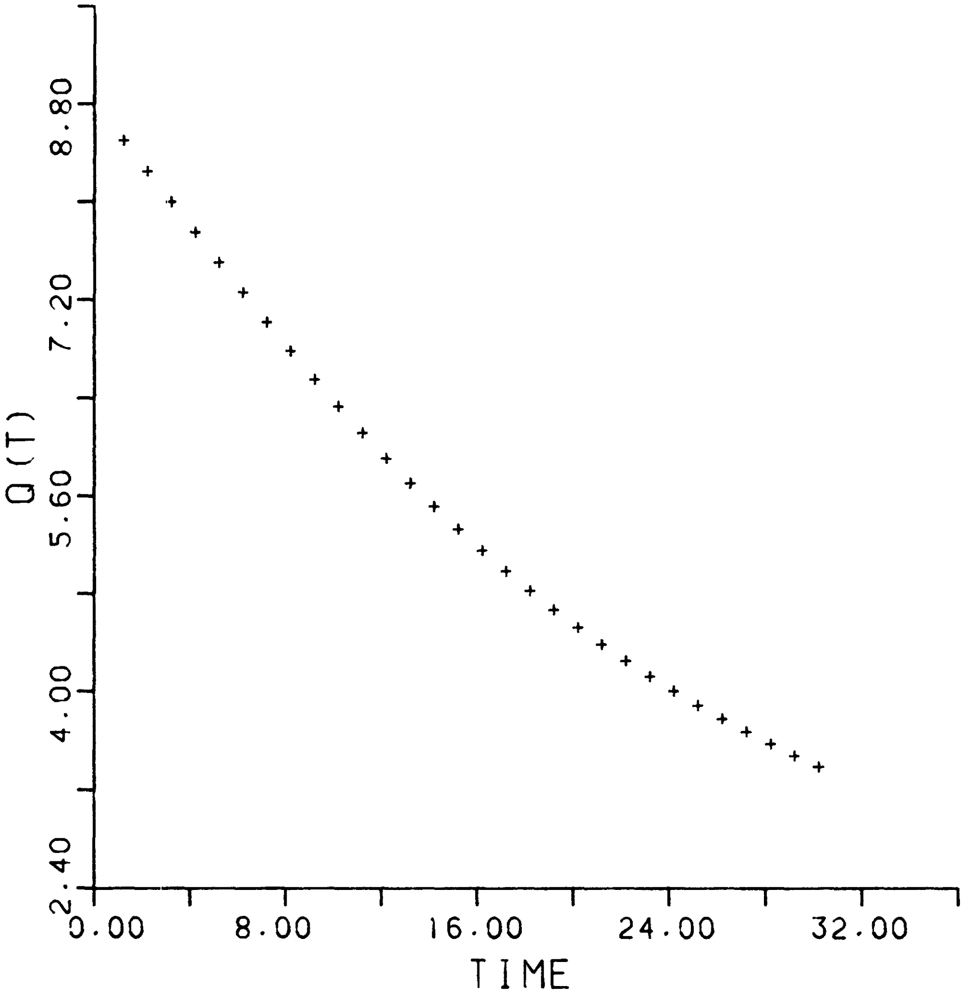


Figure 4.23: The Expected Queue Length Versus Time for an $E_3/M/1$ System With $\rho=.75$, $\mu=1$, and $P_{10}(0)=1$

In all cases, for large t decay has been shown to be approximately exponential. Closer examination of the $\log|Q(\infty) - Q(t)|$ curves shows that the time constant can be approximated by τ_R (expression (3.24)). In Table 4.1, we compare τ_R with τ_{exp}, the experimental time constant, for each of the examples considered previously in this section. It is important to note that as systems which begin close to equilibrium reach equilibrium very quickly, greater accuracy (on the order of 10^{-6}) is needed to graphically measure the experimental time constant.

The expected queue length has been shown empirically to approach its equilibrium value at least as quickly as a decaying exponential function with time constant τ_R for systems in which Q(0) equals 0 and also for systems in which Q(0) is moderately larger than $Q(\infty)$. Thus, for these systems, after $4\tau_R$ time units at most 1.8% of the transient effects will remain.

Systems in category (ii) begin closer to equilibrium than those in category (i). Therefore, as the time constant τ appears to be independent of initial conditions, we expect the time to equilibrium to be bounded above by that of a corresponding system which is initially at rest. By comparing the Q(t) curves in Figures 4.5 and 4.6, we see that this is, in fact, the case.

The expected queue length of systems in which $Q(0) >> Q(\infty)$ has been shown to decay initially in a linear manner with slope $-(\mu-\lambda)$ and then to approximate an exponentially decaying function which has time constant τ_R. Thus, the rate of decay can be estimated for each part. Unfortunately, at this point we do not have the means to determine at what time Q(t) begins to appear exponential. We can determine, however, an upper bound for the time to equilibrium by using a generous estimate for the length of the

Table 4.1: A Comparison of Estimated and Observed Time Constants

System	ρ	α	μ_1	μ_2	Initial Conditions	τ_{exp}	τ_R	τ_{exp}/τ_R
M/M/1	.75	-	1	-	$P_0(0)=1$	40	41	1.0
					$P_1(0)=1$	41	41	1
					$P_3(0)=1$	41	41	1.0
					$P_4(0)=1$	43	41	1.0
					$P_5(0)=1$	44	41	1.1
					$P_6(0)=1$	47	41	1.1
					$P_7(0)=1$	35	41	.9
					$P_8(0)=1$	37	41	.9
					$P_{34}(0)=1$	49	41	1.2
	.8	-	1	-	$P_{34}(0)=1$	76	66	1.2
	.85	-	1	-	$P_{34}(0)=1$	124	122	1.0
	.9	-	1	-	$P_{34}(0)=1$	261	281	.9
$M/H_2/1$	.75	.2	1	2	$P_1(0)=1$	27	28	1.0
					$P_3(0)=1$	28	28	1.0
					$P_4(0)=1$	27	28	1.0
					$P_5(0)=1$	32	28	1.1
					$P_6(0)=1$	31	28	1.1
					$P_{10}(0)=1$	31	28	1.1
					$P_{15}(0)=1$	28	28	1.0
					$P_{20}(0)=1$	30	28	1.1
					$P_{25}(0)=1$	34	28	1.2
$E_3M/1$	.75	-	1	-	$P_1(0) =1$	27	28	1.0
					$P_3(0)=1$	28	28	1.0
					$P_5(0)=1$	25	28	.9
					$P_{10}(0)=1$	31	28	1.1

linear portion of $Q(t)$. Our results indicate that once $Q(t)$ is within 10% of $Q(\infty)$, transient decay is approximately exponential. Assuming linear decay, if a system initially contains i customers with probability 1 ($i >> Q(\infty)$), t_ℓ units of time will be needed for $Q(t)$ to be equal to $(1.1)Q(\infty)$, where

$$(1.1)Q(\infty) = Q(0) - (\mu-\lambda)t_\ell \ , \tag{4.4}$$

or

$$t_\ell = \frac{Q(0) - (1.1)Q(\infty)}{(\mu-\lambda)} \ . \tag{4.5}$$

Thus, the time to equilibrium is bounded above by t_ℓ <u>plus</u> t_e, where t_e is the time required for a function which decays in an exponential manner with parameter τ_R to decrease from $(1.1)Q(\infty)$ to the desired accuracy level.

To summarize the empirical results presented in this section, let T_η be defined as the amount of time required for $(1-\eta)\%$ of the initial transient effects to dissipate, $0<\eta<1$. For the categories defined as in Section 4.1.1, our empirical results indicate the following bounds for T_η:

(i) $T_\eta \leq -\tau_R \log \eta$

(ii) $T_\eta \leq -\tau_R \log \eta$

(iii) $T_\eta \leq -\tau_R \log \eta$

(iv) $T_\eta \leq \dfrac{(1-\eta)\ [Q(0) - Q(\infty)]}{\mu-\lambda}$, if $\eta[Q(0) - Q(\infty)] \geq .1Q(\infty)$

$$T_\eta \leq \frac{Q(0) - 1.1Q(\infty)}{\mu-\lambda}$$

$$-\tau_R \log \frac{\eta[Q(0) - Q(\infty)]}{.1Q(\infty)} \ , \text{ if } \eta[Q(0) - Q(\infty)] < .1Q(\infty)$$

4.2 Systems with Probabilistic Specification of Initial Conditions

All of the cases considered to this point have had a deterministic specification of the number of customers in the system at time t=0. We illustrate the effect of a probabilistic specification of the initial conditions, with several examples of an M/M/1 system with $\rho=.75$, $\mu=1$, $Q(0) = 9$, but different $P_i(0)$ distributions, $i=0,1,\ldots$. Figures 4.24 and 4.25 show plots of $\log|Q(\infty) - Q(t)|$ and Q(t) versus t, respectively, for four such cases.

For large t, Q(t) approaches $Q(\infty)$ in an approximately exponential manner as in the cases in which initial conditions are deterministic. For small t, behavior is heavily influenced by the probabilistic aspect of the initial conditions. Note that as the probability increases for the system to have a large number of customers at t=0, the time to approach steady-state increases as well. This is a result of the initial linear behavior of oversaturated systems which we discussed in the previous section.

We expect Q(t) to behave as in Figures 4.24 and 4.25 for the following reason: the transient response of any system with given initial conditions $P_i(0)$, $i=0,1,\ldots$, can be determined directly from the Q(t) values for systems in which initial conditions are deterministic. Justification is simple and will be illustrated first through a particular example; the M/M/1 system with $\rho=.75$, $\mu=1$, and $P_0(0) = P_{19}(0) = 1/2$. The transient behavior of this system is illustrated in Figures 4.24 and 4.25.

An equivalent interpretation of this system is through use of a Bernoulli trial, i.e., with probability 1/2, Q(t) is the "output" of an M/M/1 system (with $\rho=.75$ and $\mu=1$) which is empty at t=0, and with probability 1/2 it is the "output" of the corresponding system which

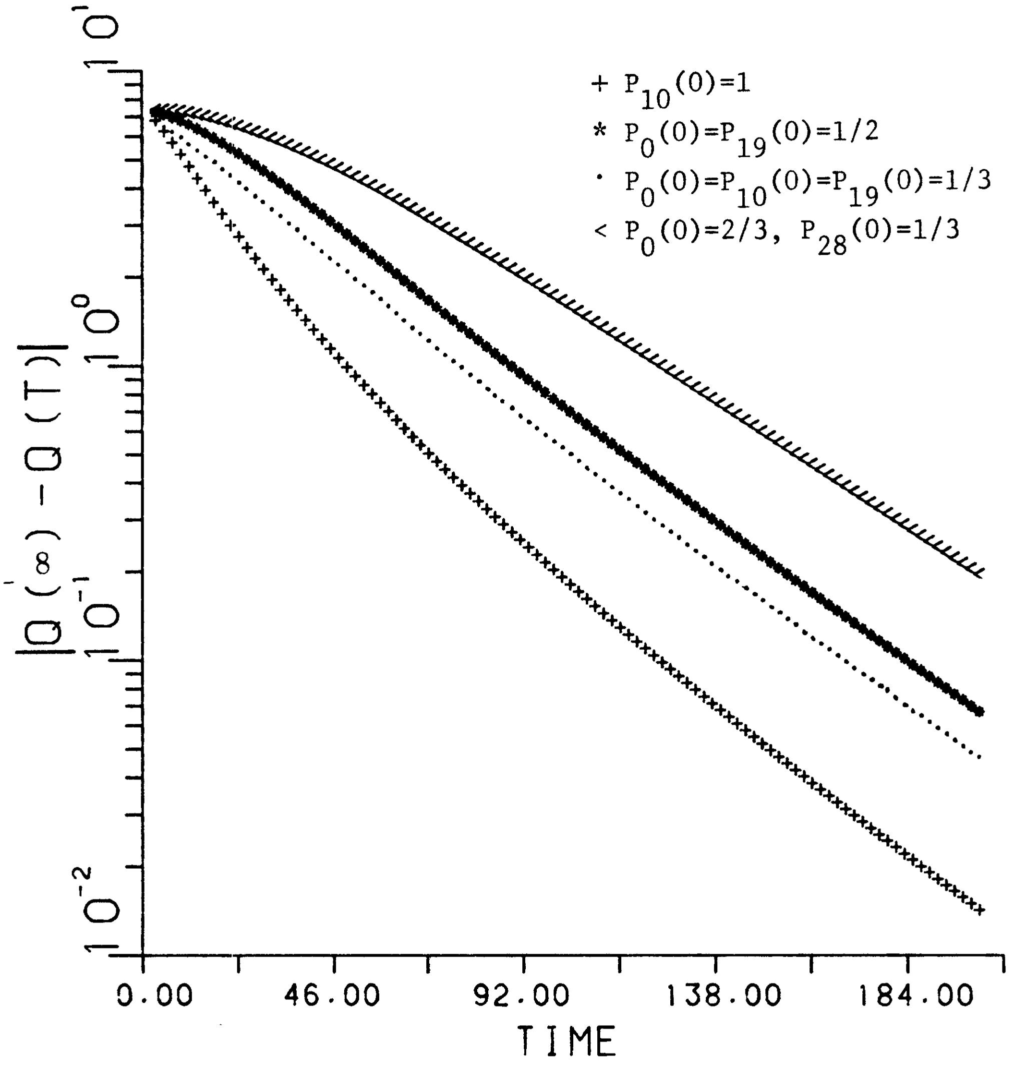


Figure 4.24: **Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an M/M/1 System With $\rho=.75$ and $\mu=1$; Probabilistic Initial Conditions**

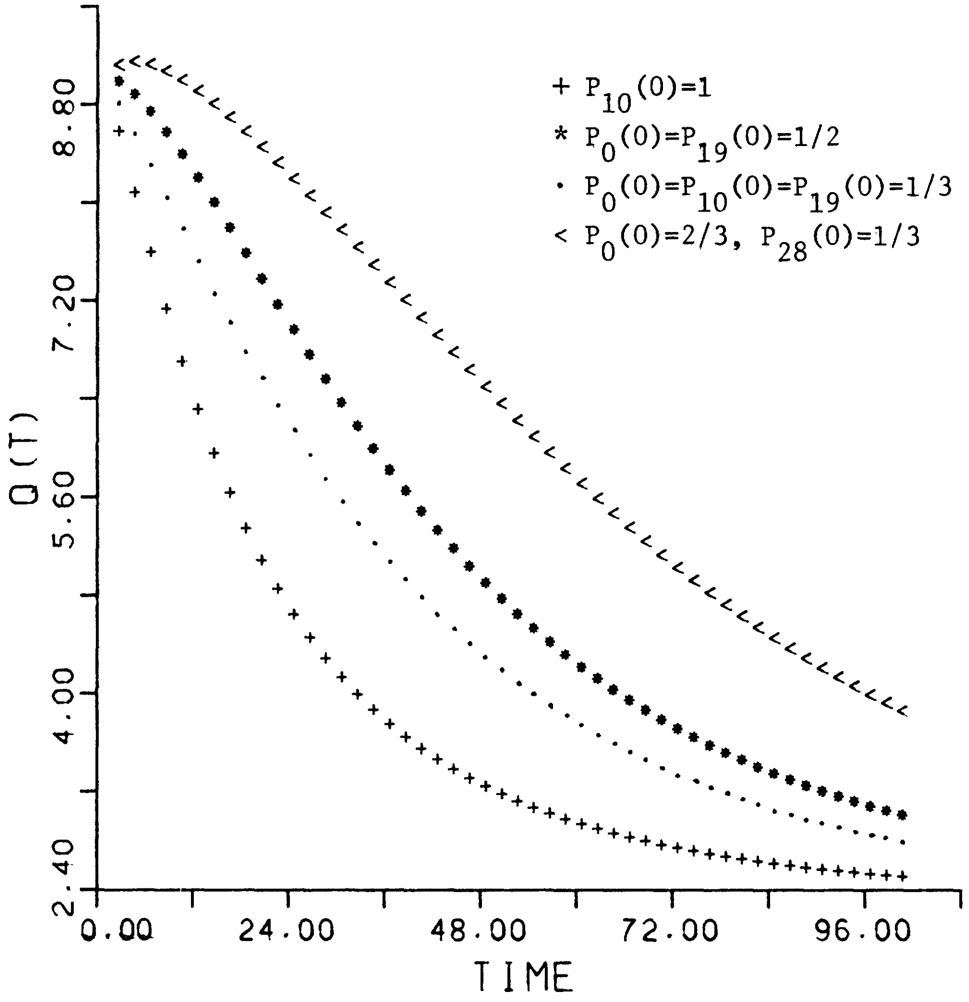


Figure 4.25: The Expected Queue Length Versus Time for an M/M/1 System With $\rho=.75$ and $\mu=1$; Probabilistic Initial Conditions

initially contains 19 customers. Thus, the expected queue length, $Q(t)$, for the system with $P_0(0) = P_{19}(0) = 1/2$ is given by

$$Q(t) = (1/2)\ Q_0(t) + (1/2)\ Q_{19}(t), \quad t \geq 0\ , \tag{4.6}$$

where $Q_i(t)$ is defined as the expected queue length at time t for a system which contains i customers, $i=0,1,\ldots$, (with probability 1) at time $t=0$.

To prove this result in a more formal manner, let "system *" be a queueing system with any specified set of initial conditions $P_i(0)$, $i=0,1,\ldots$. We define random variable $q(t)$ to be the queue length of "system *" at time t and $Q(t) = E(q(t))$. Also, let "system i" be an equivalent queueing system with the deterministic initial conditions $P_i(0)=1$, for some $i=0,1,\ldots$, and $P_j(0) = 0$, for all $j \neq i$. Define random variable $q_i(t)$ as the queue length of system i at time t, $i=0,1,\ldots$, and $Q_i(t) = E(q_i(t))$, $i=0,1,\ldots$. With probability $P_i(0)$, $i=0,1,\ldots$, $q(t) = q_i(t)$ for all t. This implies that

$$\begin{aligned} E(q(t)) &= \sum_{k=0}^{\infty} kP[q(t)=k] \\ &= \sum_{k=0}^{\infty} k \sum_{i=0}^{\infty} P_i(0)\ P[q_i(t) = k] \\ &= \sum_{i=0}^{\infty} P_i(0) \sum_{k=0}^{\infty} kP[q_i(t) = k] \\ &= \sum_{i=0}^{\infty} P_i(0)\ E(q_i(t)). \end{aligned} \tag{4.7}$$

Thus,

$$Q(t) = \sum_{i=0}^{\infty} P_i(0)\ Q_i(t) \tag{4.8}$$

Note that this is is a general result; we made no assumptions on the type of queueing system at hand.

There are three important consequences of this result. First, knowledge of the entire probability mass function $P_i(0)$, $i=0,1,\ldots$, is required to specify the transient behavior of a queueing system; as indicated in Figures 4.24 and 4.25, Q(0) will not suffice. Second, as the expected queue length for a system with a probabilistic specification of the initial conditions is just equal to a weighted sum of the expected queue lengths of a collection of corresponding systems which have deterministic initial conditions (i.e., expression (4.8)), we need only be able to characterize systems with deterministic initial conditions in order to understand the behavior of systems with a probabilistic specification of initial conditions. Third, if a system has initial conditions given by a probabilistic choice, the time required for the transients to become negligible can be bounded above by the maximum of the times required by each of the corresponding deterministic systems. Note that this may be a very conservative bound if the probability is extremely small that the actual system (with probabilistic initial conditions) behaves like the particular component system (with deterministic initial conditions) which requires the longest time to effectively reach equilibrium.

CHAPTER 5

AN APPROXIMATE SOLUTION TECHNIQUE FOR STATIONARY MARKOVIAN AND PARTIALLY DETERMINISTIC SYSTEMS

In this chapter, we propose a heuristic in which exact numerical solutions of M/M/1 systems are modified to yield approximations to the expected queue length, Q(t), for Markovian systems and for partially deterministic systems in which the embedded chain is a first-order Markov process. This heuristic was designed to take advantage of some of the best features of exact numerical solution techniques and also of the approximation (3.25) developed in Chapter 3. Solution techniques such as the one discussed in Section 2.2 can be used to determine exact numerical values for Q(t) as it evolves with time, but each new system must be solved individually leading potentially to high computation costs. Expressions (3.25) were found to be useful in estimating the time required for a system to effectively reach equilibrium with virtually no computation cost. Unfortunately, (3.25) can, at best, represent only some major attributes of the actual functional form of Q(t); this is due primarily to the nonexponential nature of Q(t) for small t. The heuristic developed in this chapter yields what appear to be excellent approximations to Q(t) at relatively low computation cost, even for fairly small values of t.

There are two major reasons for the relatively low computation cost of this heuristic. First, given the numerical solution to a particular M/M/1 system, approximate solutions can be determined for a number of more general systems with only simple arithmetic calculations. Thus, the numerical solution to a single system yields, in effect, solutions to many systems.

Second, the M/M/1 system has a simple state description--only (N+1) states are required to define a system which has a capacity of N customers.

The state descriptions for most other systems are substantially more complex. As an example, consider an $E_{10}/E_{10}/1$ queueing system with capacity of N customers. The number of states required to define this system as a first-order Markov process is 100N + 10. Thus, for large N, numerically solving the M/M/1 system will be significantly less expensive than obtaining a numerical solution to the corresponding $E_{10}/E_{10}/1$ system.

The key idea in our heuristic is this: we have shown empirically that the transient effects of the expected queue length decay in a similar manner for a rather general class of queueing systems, but that the time constants and equilibrium values are system-specific. From this we conjecture that by scaling the numerical solution to one system and changing the time axis, we should be able to determine the approximate transient response for a second system. As before, we choose the expected queue length as our representative measure of system behavior. The analysis holds for the expected delay as well.

The class of systems under consideration is the same as in Chapter 3; ergodic, infinite-capacity, single-queue, single-server systems which have a first-order Markov chain representation or, partially deterministic systems in which the embedded chain is a first-order Markov process. In addition, due to strong dependence of the initial functional form of Q(t) on the initial conditions (as discussed in Chapter 4), we restrict the discussion to systems which begin at rest.

As summarized in Section 3.3, our empirical results suggest that the expected queue length can be approximated by

$$Q(t) \cong Q(\infty) + Ae^{-t/\tau_R} , \quad t \geq \hat{t} , \qquad (5.1a)$$

where

$$\tau_R = \frac{(1+\sqrt{\rho})^2(C_a^2 + C_s^2)}{2.7\ \mu(1-\rho)^2} \quad , \tag{5.1b}$$

A is a parameter dependent on initial conditions and possibly other system characteristics, and $\hat{t}$ is on the order of $2\tau_R$. In Chapter 3 we showed empirically that expressions (5.1) with A = -Q(∞) provide a lower bound for Q(t) for systems which are in the empty state at time t=0.

We now suggest the following heuristic to approximate the expected queue length at time t, $Q_s(t)$,[1] for any infinite-capacity, single-queue, single-server Markovian or partially deterministic system which begins at rest:

1. Numerically solve (for $Q_{M/M/1}(t)$) an M/M/1 system which is initially idle, and has the same traffic intensity, ρ, as the system under consideration,[2]
2. Calculate $Q_{M/M/1}(\infty)$, $Q_s(\infty)$, $\tau_{M/M/1}$, and τ_s,
3. Multiply $Q_{M/M/1}(t)$ by the ratio $Q_s(\infty)/Q_{M/M/1}(\infty)$, and
4. Scale the time axis by $\tau_{M/M/1}/\tau_s$.

The resulting values provide the approximate expected queue length, $\hat{Q}_s(t)$, $t \geq 0$, i.e.,

$$\hat{Q}_s(t) = \left\{\frac{Q_s(\infty)}{Q_{M/M/1}(\infty)}\right\} \times \left\{Q_{M/M/1}\left(\frac{\tau_{M/M/1}}{\tau_s}\ t\right)\right\}, \ t \geq 0. \tag{5.2}$$

[1] In this chapter we will use a modified notation. $Q_s(t)$ is the expected queue length of a specific queueing system at time t. τ_s is the time constant for this system. $\hat{Q}_s(t)$ will refer to the approximate expected queue length at time t, resulting from use of our approximation technique. Similarly, $Q_{M/M/1}(t)$ and $\tau_{M/M/1}$ are the expected queue length and time constant for a specified M/M/1 system.

[2] The requirement that the two systems have the same traffic intensity improves accuracy--this will be discussed in greater detail later.

Note that, due to the manner in which this expression was defined, $\hat{Q}_s(\infty) = Q_s(\infty)$.

To accomplish step 1, we use the numerical solution technique described in Section 2.2. For step 2, $Q_{M/M/1}(\infty)$ can be calculated exactly (expression (3.5)). If, due to the form of the interarrival and/or service time distributions, $Q_s(\infty)$ cannot be evaluated analytically, it may be approximated through use of a numerical technique (such as the one described in Section 2.2) or of simulation (see Section 2.1.2). We use τ_R (expression (5.1b)) to approximate the time constants $\tau_{M/M/1}$ and τ_s.

We illustrate this heuristic through the following example. Consider an $M/E_2/1$ queueing system with $\rho=.75$ and $\mu=1$. We will approximate the expected queue length for this system by modifying the numerical solution of an M/M/1 queueing system with the same values of ρ and μ. (Both systems are assumed to begin at rest.) The steady-state expected queue lengths are different for these two systems, but both can be calculated exactly through use of the Pollaczek-Khintchine formula (3.5). Specifically, $Q_{M/E_2/1}(\infty) = 1.6875$, and $Q_{M/M/1}(\infty) = 2.25$. Also, the approximate time constants are different with $\tau_{R_{M/E_2/1}} \cong 31$ and $\tau_{R_{M/M/1}} \cong 41$. Thus, if our heuristic and expressions (5.1) are valid, we should be able to approximate the expected queue length of the $M/E_2/1$ system, $Q_{M/E_2/1}(t)$, by $\hat{Q}_{M/E_2/1}(t)$, where

$$\hat{Q}_{M/E_2/1}(t) = \left(\frac{Q_{M/E_2/1}(\infty)}{Q_{M/M/1}(\infty)}\right) Q_{M/M/1}\left(\frac{\tau_{R_{M/M/1}}}{\tau_{R_{M/E_2/1}}}\, t\right)$$

$$\cong .75\, Q_{M/M/1}(1.3\, t), \quad t \geq 0 \,. \tag{5.3}$$

In Table 5.1, we list the true value of $Q_{M/E_2/1}(t)$ (obtained numerically), $\hat{Q}_{M/E_2/1}(t)$ (obtained through use of our heuristic), and their ratio, for a range of t. In Figure 5.1, we show graphs of $Q_{M/E_2/1}(t)$ and $\hat{Q}_{M/E_2/1}(t)$ versus time. It is clear that for this example the heuristic yields an approximate solution, $\hat{Q}_{M/E_2/1}(t)$, which is in very close agreement to the exact numerical solution $Q_{M/E_2/1}(t)$.

Tables 5.2 - 5.6 illustrate additional examples of the application of this heuristic to other queueing systems. In each of these cases, the agreement between $Q_s(t)$ and $\hat{Q}_s(t)$ is very good. Note that $Q_s(t)$ and $\hat{Q}_s(t)$ differ most for small t. This is not surprising as our approximate solution technique is based on the hypothesis that Q(t) approaches Q(∞) in an approximately exponential manner and the empirical results of Chapter 3 show that this hypothesis is, in general, valid only for relatively large t. In each of the cases that we studied, $\hat{Q}_s(t)$ was always within 6% of $Q_s(t)$ after at most τ_s units of time. All of these results were obtained by modifying the numerical solution of an M/M/1 system with the same traffic intensity as the system to be approximated. In all cases listed, $Q_{M/M/1}(t)$ was computed to four decimal places and the ratio $\tau_{R_{M/M/1}}/\tau_{R_s}$ to one decimal place. If ρ is small, greater accuracy is required. In particular, we have found that more significant figures should be retained if system parameters are such that $Q_s(\infty)$ is less than 1.

Note that the M/M/1 system on which the approximation is based need not have the same service rate μ as the system to be approximated. This is due to the result shown in Section 3.2 that, given ρ, the <u>only</u> dependence of Q(t) on μ is a scaling of the time axis by a factor of 1/μ. This is completely accounted for in the approximate time constant τ_R.

TABLE 5.1: Estimate of the Expected Queue Length of an $M/E_2/1$ System With $\rho=.75$, $\mu=1$ From an $M/M/1$ System With $\rho=.75$, $\mu=1$

t	$Q_{M/E_2/1}(t)$	$\hat{Q}_{M/E_2/1}(t)$	$Q_{M/E_2/1}(t)/\ Q_{M/E_2/1}(t)$	
1	.1513	.1400	1.08	
2	.3289	.3046	1.08	
3	.4682	.4397	1.06	
4	.5815	.5514	1.05	
5	.6764	.6458	1.05	
6	.7576	.7269	1.04	
7	.8283	.7978	1.04	
8	.8905	.8604	1.03	
9	.9458	.9161	1.03	
10	.9953	.9662	1.03	
20	1.3058	1.2831	1.02	
30	1.4559	1.4386	1.01	
40	1.5397	1.5266	1.01	$\tau_{R_{M/E_2/1}}$
50	1.5901	1.5802	1.01	
60	1.6219	1.6144	1.00	
70	1.6426	1.6368	1.00	
80	1.6563	1.6519	1.00	

$$\hat{Q}_{M/E_2/1}(t) = .75\ Q_{M/M/1}(1.3t),\ t \geq 0$$

$$\tau_{R_{M/M/1}} = 41 \qquad Q_{M/M/1}(\infty) = 2.25$$

$$\tau_{R_{M/E_2/1}} = 31 \qquad Q_{M/E_2/1}(\infty) = 1.6875$$

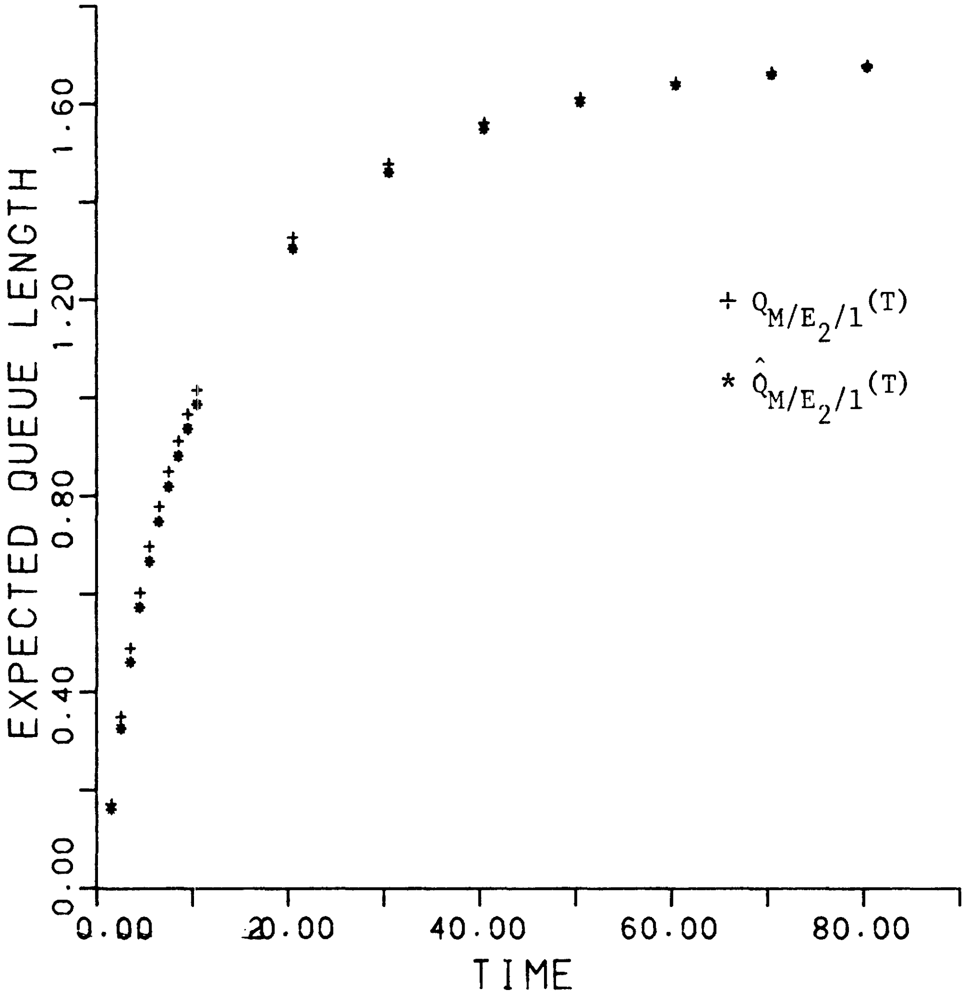


Figure 5.1: Estimate of the Expected Queue Length of an $M/E_2/1$ System With $\rho=.75$, $\mu=1$ from an M/M/1 System With $\rho=.75$ and $\mu=1$

TABLE 5.2: Estimate of the Expected Queue Length of an M/D/1 System With ρ=.85, μ=1 From an M/M/1 System With ρ=.85, μ=1

t	$Q_{M/D/1}(t)$	$\hat{Q}_{M/D/1}(t)$	$Q_{M/D/1}(t)/\hat{Q}_{M/D/1}(t)$	
2	.4654	.3861	1.21	
4	.7304	.6592	1.11	
6	.9210	.8560	1.08	
8	1.0702	1.0101	1.06	
10	1.1924	1.1363	1.05	
20	1.5909	1.5485	1.03	
30	1.8181	1.7844	1.02	
40	1.9661	1.9386	1.01	
50	2.0689	2.0462	1.01	
60	2.1434	2.1245	1.01	
70	2.1989	2.1830	1.01	
80	2.2412	2.2277	1.01	$\tau_{R_{M/D/1}}$
90	2.2738	2.2624	1.01	
100	2.2993	2.2896	1.00	
110	2.3194	2.3112	1.00	
120	2.3355	2.3285	1.00	

$Q_{M/D/1} = .5\ \hat{Q}_{M/M/1}(2t),\ t \geq 0$

$\tau_{R_{M/M/1}} = 122 \qquad Q_{M/M/1}(\infty) = 4.8167$

$\tau_{R_{M/D/1}} = 61 \qquad Q_{M/D/1}(\infty) = 2.4083$

TABLE 5.3: Estimate of the Expected Queue Length of an $E_4/M/1$ System With $\rho=.75$, $\mu=1$ From an M/M/1 System With $\rho = .75$, $\mu=1$

t	$Q_{E_4/M/1}(t)$	$\hat{Q}_{E_4/M/1}(t)$	$Q_{E_4/M/1}(t)/\hat{Q}_{E_4/M/1}(t)$	
1	.0069	.1321	.05	
2	.0941	.2712	.35	
3	.2129	.3802	.56	
4	.3141	.4679	.67	
5	.3995	.5407	.74	
6	.4725	.6025	.78	
7	.5356	.6557	.82	
8	.5908	.7022	.84	
9	.6397	.7431	.86	
10	.6831	.7796	.88	
20	.9482	1.0009	.95	
30	1.0697	1.1015	.97	$\tau_{R_{E_4/M/1}}$
40	1.1346	1.1547	.98	
50	1.1720	1.1851	.99	
60	1.1946	1.2034	.99	
70	1.2087	1.2146	1.00	
80	1.2177	1.2218	1.00	
90	1.2236	1.2264	1.00	

$$\hat{Q}_{E_4/M/1}(t) = .5492\ Q_{M/M/1}(1.6t), \quad t \geq 0$$

$\tau_{R_{M/M/1}} = 41$ $\qquad Q_{M/M/1}(\infty) = 2.25$

$\tau_{R_{E_4/M/1}} = 26$ $\qquad Q_{E_4/M/1}(\infty) = 1.2357$

TABLE 5.4: Estimate of the Expected Queue Length of an $E_2/E_2/1$ System With $\rho=.85$, $\mu=1$ From an M/M/1 System With $\rho=.85$, $\mu=1$

t	$Q_{E_2/E_2/1}(t)$	$\hat{Q}_{E_2/E_2/1}(t)$	$Q_{E_2/E_2/1}(t)/\hat{Q}_{E_2/E_2/1}(t)$	
2	.0468	.2656	.18	
4	.1801	.4534	.40	
6	.3094	.5887	.53	
8	.4228	.6947	.61	
10	.5215	.7815	.67	
20	.8679	1.0650	.81	
30	1.0777	1.2273	.88	
40	1.2181	1.3333	.91	
50	1.3173	1.4074	.94	
60	1.3901	1.4612	.95	
70	1.4448	1.5015	.96	$\tau_{R_{E_2/E_2/1}}$
80	1.4868	1.5322	.97	
90	1.5195	1.5560	.98	
100	1.5452	1.5748	.98	
110	1.5657	1.5896	.98	
120	1.5821	1.6015	.99	
130	1.5953	1.6111	.99	
140	1.6060	1.6189	.99	
150	1.6147	1.6252	.99	
160	1.6219	1.6304	.99	
170	1.6278	1.6347	1.00	
180	1.6326	1.6382	1.00	
190	1.6366	1.6411	1.00	

$$\hat{Q}_{E_2/E_2/1}(t) = .3439\ Q_{M/M/1}(2t),\ t \geq 0$$

$\tau_{R_{M/M/1}} = 122$ $\qquad Q_{M/M/1}(\infty) = 4.8167$

$\tau_{R_{E_2/E_2/1}} = 61$ $\qquad Q_{E_2/E_2/1}(\infty) = 1.6566$

TABLE 5.5: Estimate of the Expected Queue Length of an $M/H_2/1$ System With $\rho=.75$, $\alpha=.2$, $\mu_1=1$, $\mu_2=2$ ($\mu=1.667$) From an M/M/1 System With $\rho=.75$, $\mu=1$

t	$Q_{M/H_2/1}(t)$	$\hat{Q}_{M/H_2/1}(t)$	$Q_{M/H_2/1}(t)/\hat{Q}_{M/H_2/1}(t)$	
2	.5003	.5172	.97	
3	.7072	.7315	.97	
4	.8779	.9055	.97	
5	1.0219	1.0507	.97	
6	1.1454	1.1743	.98	
7	1.2527	1.2813	.98	
8	1.3470	1.3751	.98	
9	1.4307	1.4581	.98	
10	1.5055	1.5321	.98	
20	1.9684	1.9875	.99	
30	2.1863	2.1999	$\overline{.99}$	$\tau_{R_{M/H_2/1}}$
40	2.3050	2.3148	1.00	
50	2.3748	2.3818	1.00	
60	2.4177	2.4228	1.00	

$$\hat{Q}_{M/H_2/1}(t) = 1.1111\ Q_{M/M/1}(1.5t),\ t \geq 0$$

$\tau_{R_{M/M/1}} = 41$ $\qquad Q_{M/M/1}(\infty) = 2.25$

$\tau_{R_{M/H_2/1}} = 28$ $\qquad Q_{M/H_2/1}(\infty) = 2.5$

Table 5.6: Estimate of the Expected Queue Length of an $M/\hat{G}/1^*$ System With $\rho=.8$, $\beta=1/6$, $\mu_1=\mu_2=1$, From an M/M/1 System With $\rho=.8$, $\mu=1$

t	$Q_{M/\hat{G}/1}(t)$	$\hat{Q}_{M/\hat{G}/1}(t)$	$Q_{M/\hat{G}/1}(t)/\hat{Q}_{M/\hat{G}/1}(t)$	
2	.1160	.1083	1.07	
4	.2835	.2544	1.11	
6	.4216	.3839	1.10	
8	.5378	.4962	1.08	
10	.6384	.5946	1.07	
20	1.0033	.9567	1.05	
30	1.2435	1.1979	1.04	
40	1.4186	1.3751	1.03	
50	1.5533	1.5121	1.03	
60	1.6606	1.6218	1.02	
70	1.7480	1.7116	1.02	
80	1.8204	1.7863	1.02	
90	1.8813	1.8493	1.02	
100	1.9330	1.9031	1.02	
110	1.9773	1.9493	1.01	
120	2.0155	1.9893	$\underline{1.01}$	
130	2.0487	2.0242	1.01	$\tau_{R_{M/\hat{G}/1}}$
140	2.0777	2.0548	1.01	
150	2.1031	2.0816	1.01	
160	2.1254	2.1054	1.01	
170	2.1452	2.1264	1.01	

$\hat{Q}_{M/\hat{G}/1}(t) = .7266\ Q_{M/M/1}(.5),\ t \geq 0$

$\tau_{R_{M/M/1}} = 66$ $\qquad Q_{M/M/1}(\infty) = 3.2$

$\tau_{R_{M/\hat{G}/1}} = 129$ $\qquad Q_{M/\hat{G}/1}(\infty) = 2.325$

*where the service time, s, has pdf

$$f_s(s_o) = \beta\mu_1 e^{-\mu_1 s_o} + (1-\beta)\,\frac{\mu_2^3}{2}\, s_o^2\, e^{-\mu_2 s_o}\ , \quad s_o \geq 0.$$

A consequence of this result is that in applications of the heuristic, if the system to be approximated differs from the underlying system only in the value of μ (i.e., the same type of interarrival and service time distributions and the same arrival rate λ), the solution generated through use of the heuristic will, with one possible exception, have exactly the same level of accuracy as the numerical solution on which the approximation is based. The only exception occurs if, in the time scale modification of step 4, the ratio of the time constants is not retained at a high enough level of accuracy.

Use of this observation can lead to substantial savings in computation costs. The time constant varies inversely with μ, thus equilibrium will occur more rapidly in a system with large μ. This implies that if ρ remains fixed, unless a variation in the behavior of the numerical subroutine in response to the change in μ dominates, the cost to solve a system numerically until it effectively reaches steady-state will increase as μ decreases; significant variation in the behavior of the numerical routine is likely to occur only with large changes in μ (greater than an order of magnitude). Therefore, if a solution is desired for a system which has service rate μ, scaling the numerical solution of a system with the same interarrival and service time distributions and traffic intensity but a moderately larger service rate will produce a solution at the same level of accuracy and typically at a lower cost than that obtained through direct numerical solution.

Although any value of μ can be used, the traffic intensity should be the same for the system to be approximated and the M/M/1 system on which the approximation is based. This requirement greatly improves the accuracy of the approximation technique.

To illustrate this last point, Table 5.7 is a comparison of $Q_{M/E_2/1}(t)$ and $\hat{Q}_{M/E_2/1}(t)$ of a system with ρ=.75 and ρ=1 obtained from the numerical solution of an M/M/1 queueing system with ρ=.5 and μ=1. As can be seen through comparison with Table 5.1, the approximation is significantly better for early t if the underlying M/M/1 system has the same traffic intensity. This suggests that the actual variation of the form of Q(t) with ρ is different than the $\frac{(1+\sqrt{\rho})^2}{(1-\rho)^2}$ factor accounted for in τ_R.

Our heuristic will yield similar results if the underlying system is something other than M/M/1. As an example, we use the heuristic to obtain an approximate solution for an $M/E_{10}/1$ system with ρ=.75 and μ=1. Table 5.8 is a comparison of $Q_{M/E_{10}/1}(t)$ for this system with $\hat{Q}_{M/E_{10}/1}(t)$ obtained through modification of the solution for an M/M/1 system with ρ=.75 and μ=1. In Table 5.9 we compare $Q_{M/E_{10}/1}(t)$ with $\hat{Q}_{M/E_{10}/1}(t)$ obtained from an underlying $M/E_5/1$ system with ρ=.75 and μ=1.

In both cases the heuristic generates a good approximate solution for the $M/E_{10}/1$ system. The solution based on an underlying $M/E_5/1$ system is significantly more accurate for early t. This additional accuracy, however, also results in higher computation costs.[3]

We chose to base the heuristic on a modification of an underlying M/M/1 system, as the M/M/1 solution requires the least amount of CPU time among the class of systems for which we can obtain exact numerical solutions (as opposed to approximate solutions, e.g., M/D/1 systems). The results of this chapter provide evidence that, for most applications, accuracy will be quite good with an M/M/1 system. In practice, if accuracy is of greater importance than computation cost, the user might

[3] See Section 2.2.4 for a comparison of computation costs for these two systems.

TABLE 5.7: Estimate of the Expected Queue Length of an $M/E_2/1$ System With $\rho=.75$, $\mu=1$ From an M/M/1 System With $\rho=.5$, $\mu=1$

t	$Q_{M/E_2/1}(t)$	$\hat{Q}_{M/E_2/1}(t)$	$Q_{M/E_2/1}(t)/\hat{Q}_{M/E_2/1}(t)$	
1	.1513	.0300	5.04	
2	.3289	.0962	3.42	
3	.4682	.1775	2.64	
4	.5815	.2629	2.21	
5	.6764	.3473	1.95	
6	.7576	.4280	1.77	
7	.8283	.5039	1.64	
8	.8905	.5748	1.55	
9	.9458	.6406	1.48	
10	.9953	.702	1.42	
20	1.3058	1.1225	1.16	
30	1.4559	1.3426	1.08	
40	1.5397	1.4681	1.05	$\tau_{R_{M/E_2/1}}$
50	1.5901	1.5441	1.03	
60	1.6219	1.5917	1.02	
70	1.6426	1.6227	1.01	
80	1.6563	1.6430	1.01	
90	1.6656	1.6568	1.01	
100	1.6720	1.6659	1.00	
110	1.6764	1.6723	1.00	
120	1.6796	1.6767	1.00	

$\hat{Q}_{M/E_2/1}(t) = 3.375 \quad Q_{M/M/1}(.3t), \ t \geq 0$

$\tau_{R_{M/M/1}} = 8.6 \qquad Q_{M/M/1}(\infty) = .5$

$\tau_{R_{M/E_2/1}} = 31 \qquad Q_{M/E_2/1}(\infty) = 1.6875$

TABLE 5.8: Estimate of the Expected Queue Length of an $M/E_{10}/1$ System With $\rho=.75$, $\mu=1$ From an M/M/1 System With $\rho=.75$, $\mu=1$

t	$Q_{M/E_{10}/1}(t)$	$\hat{Q}_{M/E_{10}/1}(t)$	$Q_{M/E_{10}/1}(t)/\hat{Q}_{M/E_{10}/1}(t)$	
1	.1835	.1516	1.21	
2	.3396	.3013	1.13	
3	.4553	.4157	1.10	
4	.5456	.5067	1.08	
5	.6191	.5813	1.07	
6	.6804	.6440	1.06	
7	.7326	.6977	1.05	
8	.7776	.7442	1.04	
9	.8170	.7850	1.04	
10	.8516	.8211	1.04	
20	1.0540	1.0344	$\underline{1.02}$	
30	1.1397	1.1268	1.01	$\tau_{R_{M/E_{10}/1}}$
40	1.1822	1.1736	1.01	
50	1.2050	1.1993	1.00	
60	1.2179	1.2140	1.00	
70	1.2255	1.2228	1.00	

$$\hat{Q}_{M/E_{10}/1}(t) = .55\ Q_{M/M/1}(1.8t),\ t \geq 0$$

$\tau_{R_{M/M/1}} = 41$ $\qquad Q_{M/M/1}(\infty) = 2.25$

$\tau_{R_{M/E_{10}/1}} = 23$ $\qquad Q_{M/E_{10}/1}(\infty) = 1.2375$

TABLE 5.9: Estimate of the Expected Queue Length of an $M/E_{10}/1$ System With $\rho=.75$, $\mu=1$ From an $M/E_5/1$ System With $\rho=.75$, $\mu=1$

t	$Q_{M/E_{10}/1}(t)$	$\hat{Q}_{M/E_{10}/1}(t)$	$Q_{M/E_{10}/1}(t)/\hat{Q}_{M/E_{10}/1}(t)$
1	.1835	.1760	1.04
2	.3396	.3349	1.01
3	.4553	.4508	1.01
4	.5456	.5416	1.01
5	.6191	.6154	1.01
6	.6804	.6771	1.00
7	.7326	.7295	1.00
8	.7776	.7749	1.00

$Q_{M/E_{10}/1}(t) = .9167\ Q_{M/E_2/1}(1.1t),\ t \geq 0$

$\tau_{R_{M/E_5/1}} = 25 \qquad Q_{M/E_5/1}(\infty) = 1.35$

$\tau_{R_{M/E_{10}/1}} = 23 \qquad Q_{M/E_{10}/1}(\infty) = 1.2375$

want to base the heuristic on a system which has interarrival and/or service time distributions which are more similar to those of the system to be approximated than those of an M/M/1 system.

It is important to recognize that we have examined the accuracy of this solution technique for Markovian systems and for partially deterministic systems in which the embedded chain is a first-order Markov process. Verification for truly general cases is a difficult task. At this time, for more general systems, we are unable to obtain transient solutions which are sufficiently accurate to verify the accuracy of approximate solutions generated by the heuristic. This is certainly one area for further research.

CHAPTER 6

CONCLUSIONS AND TOPICS FOR FURTHER RESEARCH

This dissertation presents a collection of techniques and approximations useful in studying the transient response of stationary queueing systems. The primary usefulness of these methods is in applications--in approximating the behavior of actual queueing systems for which exact solutions are unavailable or intractable. In addition, our empirical results on characteristics of certain queueing systems may eventually be helpful in the development of new theoretical results. In this chapter, we summarize this work and suggest directions in which it might be extended.

In Chapter 2 we described the value of numerical solution techniques in determining transient solutions of queueing systems. To provide a background for the empirical work of the later chapters, we then discussed one such technique, originally developed by Koopman, that entails solving a truncated set of state equations to obtain exact numerical solutions for the transient behavior of stationary Markovian systems with any given set of initial conditions. The technique can also be used to provide approximate solutions for partially deterministic systems in which the embedded chain has a first-order Markov process representation. We developed a set of computer programs for solving many types of queueing systems using this numerical technique.

Application of this numerical solution technique yields the state probabilities of the system at discrete points in time. For Markovian systems and for partially deterministic systems in which the embedded chain is a first-order Markov process, use of the technique is limited only by the size of the system to be modeled, a factor that is becoming less of a constraint with the increasing sophistication of computer software and hardware. A discussion of the computation cost of this technique is included in Section 2.2.4.

The remainder of the thesis focused on ergodic, infinite-capacity single-queue, single-server queueing systems. The analysis was confined to Markovian and partially deterministic systems that can be solved through use of the numerical technique discussed in Chapter 2.

The empirical results presented in Chapter 3 indicate that the expected queue length, Q(t), has a similar form for all systems belonging to these classes, provided that they begin at rest. In particular, except for an initial period, Q(t) seems to approach Q(∞) in a virtually exponential manner. The time constant of the exponential function which was used to approximate this behavior was shown to depend on the traffic intensity ρ, the service rate μ, and the coefficients of variation of the interarrival and service time distributions. A closed-form expression was determined which can be used to estimate the amount of time required for the transient response to become negligible.

Also in Chapter 3, we have shown that for general stationary queueing systems, the transient solutions of any two systems that differ only in the values of their arrival and service rates (but have the same traffic intensity) are identical except for a scaling of the time axis. In addition, the scaling factor was shown to be equal to the ratio of the service rates. This result has important practical implications. As mentioned in Chapter 5, when using an exact numerical solution technique to determine the transient response of a queueing system, application of this result can lead to a substantial cost saving since for any given type of system, only one exact numerical solution is needed for each value of

ρ. Through a simple scaling of the time axis, solutions of exactly the same accuracy can be derived for any value of μ.

In Chapter 4, the empirical results of Chapter 3 were extended to include systems which do not begin at rest. We confirmed empirically that, for large values of t, the rate of decay of the transients is a function only of the arrival and service processes of the system--not of the initial conditions.

For systems with deterministic initial conditions, transient behavior for small values of t was observed to fall into four categories, based on the number of customers in the system at time t=0. In addition, unless a system is initially heavily saturated, i.e., $Q(0) >> Q(\infty)$, it appears that the time to equilibrium is not longer than that indicated by exponential decay. If, however, there are a large number of customers in the system at time t=0, we observed that for an initial period the system behaves as if both its interarrival and service time distributions are deterministic, so that Q(t) decays linearly. After this initial time period, our experimental results suggest that Q(t) once again decays in an approximately exponential manner.

These results were then used in the determination of similar bounds for systems which have a probabilistic specification of initial conditions. To accomplish this, we used the fact, valid for general stationary queueing systems, that systems with a probabilistic specification of initial conditions can be "decomposed" into sets of systems each of which has deterministic initial conditions.

Our empirically determined bounds on the time until transient effects become negligible are perhaps the most important contribution of this dissertation. As indicated in Chapter 1, given a particular application, a bound on the time to equilibrium can be useful in determining a suitable solution approach. If the system to be solved is stationary and transient effects are shown, through use of our bounds, to be insignificant, existing theoretical steady-state results, if available, can be used. If, on the other hand, transient effects are shown to be significant for much of the period of interest, a technique which yields transient solutions (such as the numerical technique described in Chapter 2) should be used.

The empirically derived bounds on the time to equilibrium can also be of use when the system has nonstationary parameters. We illustrate this by returning to the airport runway application mentioned in Chapter 1. The demand profile for a major airport is typically approximated as a piecewise constant function with segments one hour in length. Runway delays in each of these one hour time periods are often modeled by a stationary M/G/1 queueing system [39]. Theoretical steady-state results are then used to calculate the expected queue length (or expected delay) for each time period. The approximate expression developed here for estimating the exponential time constant, τ, can be used to obtain a rough indication of whether it is valid, in fact, to use steady-state results (i.e., whether transients are in fact negligible).

For a typical runway situation the expected service time, $1/\mu$, is on the order of 1.5 minutes, and the coefficient of variation for the service time, C_s, is about equal to 1/4 [42]. At a busy airport it is not unusual to find $\rho \geq .9$ during peak periods (this implies that the arrival rate

$\lambda = \rho\mu \geq .6$ operations/minute). Substituting these values into expression (3.24) yields an approximate time constant $\tau_R \geq 264$ minutes. Thus, if the system is not near equilibrium at the start of a one-hour time period, more than 4 hours will be required for the transients to be reduced by $e^{-1} \overset{\sim}{=} 37\%$. This implies that transient effects caused by significant variation in the demand profile over the course of a time period would require several hours to become negligible. Hence, unless the demand profile exhibits only slight variation over time, the use of steady-state results to measure the behavior of the system would not be justified. Instead, a numerical solution technique or simulation should be used to determine approximate transient behavior.

In Chapter 5, we proposed a new approximate solution technique for ergodic, infinite-capacity, single-queue, single-server systems that begin at rest. As before, systems are restricted to be either Markovian, or else partially deterministic with an embedded chain which is a first-order Markov process. Based on the empirical result of Chapter 3 that the behavior of Q(t) can be approximated by the same general functional form for all of these systems, the heuristic specifies a way to scale the exact numerical transient solution of an M/M/1 queueing system to obtain the corresponding approximate solutions for more complex systems using only simple arithmetic operations.

Comparisons of the approximate values of Q(t) obtained through use of our heuristic with exact numerical values suggest that the accuracy of this approximation is excellent for values of t greater than one time constant from the origin. Therefore, unless early behavior of the system is very important, this heuristic can be used to obtain solutions almost as accurate as those given by an exact numerical technique (e.g., the solution technique discussed in Section 2.2), typically at a significantly reduced cost. Computational savings will be particularly pronounced if the system to be solved has a Markov process representation that requires a complex (multidimensional) state description; also, the numerical solution of a single M/M/1 system can be modified to yield approximate solutions to a large number of other, more complex systems.

There are two primary directions for future work in this area--theoretical and empirical. Theoretical work might focus on a derivation or explanation of our empirical observations. Particularly useful would be a demonstration that the transient effects decay in an approximately exponential manner for large t. Given the complexity of existing exact theoretical transient solutions, it is likely that this work would be extremely difficult.

The more promising direction is the extension of our results through additional empirical work. For example, with numerical techniques it is possible to explore the entire class of stationary Markovian systems, as well as of all stationary partially deterministic systems with large ρ that have an embedded chain that is a first-order Markov process. We show some preliminary results on the transient behavior of infinite-capacity M/M/k systems and finite-capacity, single-server systems in Appendix 3. Other systems to be studied include those with bulk arrival/bulk service processes, multiple-queue systems under various priority schemes, and networks of Markovian or partially deterministic systems. In all cases, the analysis can parallel that of our Chapters 3 and 4.

The closed-form expression for τ_R provides an estimate of the exponential time constant used in our approximate expression for Q(t). We have proven that τ must vary linearly with $1/\mu$, but more work is needed to verify and perhaps modify the conjectured relationships between

τ and other system parameters (e.g., the traffic intensity and the coefficients of variation of the interarrival and service times).

We have presented initial results for the heuristic introduced in Chapter 5. However, more thorough investigation is needed, including confirmation for systems with more varied interarrival and service time distributions. In addition, after a careful examination of the results in Chapter 4, it may be possible to modify this heuristic to provide approximate transient solutions for Markovian and partially deterministic systems which do not begin at rest.

Another potentially useful extension is an empirical investigation of the transient response of performance measures other than the expected queue length. For example, the likelihood of the system containing more than some specified number of customers might be of particular interest.

Finally, perhaps the most useful extension of this empirical work would be the development of methods for obtaining accurate transient solutions for more general queueing systems (e.g., M/G/k, GI/M/k, and GI/G/k systems). We suspect that, for large t, decay of transient effects might be approximately exponential for <u>all</u> ergodic, stationary queueing systems, but at this point we do not have the means to substantiate this hypothesis.

APPENDIX 1

OBSERVATIONS ON APPLYING KOOPMAN'S NUMERICAL SOLUTION TECHNIQUE TO NONSTATIONARY QUEUEING SYSTEMS

The numerical solution technique described in Section 2.2 can be used to solve many types of nonstationary queueing systems. Since the state equations are solved iteratively, different values of the arrival and service rates may be used at each iteration, thus taking into account any variation in time of these quantities. There are, however, two potential sources of error that are unique to nonstationary systems. These will be described here.

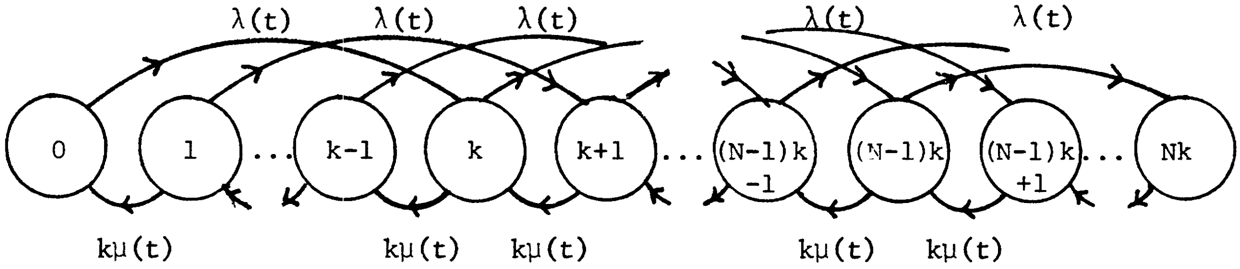
For stationary Markovian systems and for those nonstationary systems in which customer transitions occur according to a first-order Markov process (e.g., nonstationary M/M/k, M^x/M/1, and M/M^x/1 systems), the technique will yield "exact" numerical solutions (subject only to those sources of error described in Section 2.2.3). For many other systems, however, the time dependence of the arrival and/or service rates causes the state equations to represent only approximate system behavior. This error can be illustrated using the analysis of a finite-capacity M/E_k/1 queueing system. Under stationary conditions, we can consider this system to be composed of "stages" rather than customers. A $k^{\underline{th}}$-order Erlang random variable with mean $1/\mu$ is the sum of k independent, identically distributed negative exponential random variables, each with mean $1/k\mu$. Thus, the service of a single customer is equivalent to the service of k independent "stages," where the service time for each stage is given by a negative exponential random variable with mean $1/k\mu$.

For this queueing system we can define states, i = 0, 1, ..., Nk, to be the total number of stages in the system at any time (assuming that the system can have a maximum of N customers). The state transition diagram and Chapman-Kolmogorov equations for this system are shown in Figure A1.1.

If the service rate is independent of time, numerically solving the Chapman-Kolmogorov equations is equivalent to solving an exact model of the $M/E_k/1$ system. For nonstationary $\mu(t)$, the above model is only approximate. An Erlang random variable is a sum of independent identical exponential random variables. Thus, each of the stages associated with a particular customer should have the same parameter. In particular, each of the k stages of any one customer's service should have the same parameter $k\,\mu(\hat{t})$, where $\hat{t}$ is the time at which the first stage began service. In numerically solving the system, however, the parameter used at time t is $k\,\mu(t)$; the system must be memoryless so we cannot retain the original service rate $\mu(\hat{t})$. Therefore, in the numerical solution each stage may have a different parameter. For many applications this error will be negligible as $\mu(t)$ frequently varies slowly with respect to the average service time $1/\mu(t)$.

A second potential source of error is in the calculation of the expected delay. If the service rate is a function of time, the expression for the expected delay (e.g., (2.3) with μ replaced by $\mu(\hat{t})$) is only approximate; if the waiting time is long, the actual service rate for a customer arriving at time $\hat{t}$ may vary significantly from $\mu(\hat{t})$, the service rate when the customer entered the system.

State Transition Diagram



state i: i stages in the system

Chapman-Kolmogorov Equations

$$\dot{P}_0(t) = -\lambda(t)P_0(t) + k\mu(t)P_1(t)$$

$$\dot{P}_i(t) = -(\lambda(t) + k\mu(t))P_i(t) + k\mu(t)P_{i+1}(t) \qquad i = 1,2,\ldots,k-1$$

$$\dot{P}_i(t) = -(\lambda(t) + k\mu(t))P_i(t) + \lambda(t)P_{i-k}(t) + k\mu(t)P_{i+1}(t)$$

$$i = k,k-1,\ldots,Nk-1$$

$$\dot{P}_{Nk}(t) = -k\mu(t)P_{Nk}(t) + \lambda(t)P_{(N-1)k}(t)$$

Figure A1.1: State Transition Diagram and Chapman-Kolmogorov Equations for a Finite-Capacity $M/E_k/1$ Queueing System

APPENDIX 2

THE COMPUTER PROGRAMS

```
C    NUMERICAL SOLUTION OF INFINITE-CAPACITY M/M/K,
C    E(L)/M/1 AND M/E(R)/1 SYSTEMS
      IMPLICIT DOUBLE PRECISION(A-H,L,O-Z)
      INTEGER QM,QMS,STORE,CASES,L,R
      DOUBLE PRECISION MU,INC
      EXTERNAL QUEUE
      DIMENSION Y(351),YPRIME(351),W(351,9),C(24)
     -  ,PM(350),DP(350)
      EQUIVALENCE(PMO,Y(1)),(DPO,YPRIME(1)),(NM,MMAX)
     -  ,(PM(1),Y(2)),(DP(1),YPRIME(2))
      COMMON MU,ARR(145),UK,PMATO(15),
     -  PMAT(65,15),INT,K,MMAX,L,R
      OPEN (UNIT=20,ACCESS='SEQOUT',FILE='PLTQM6.DAT')
      OPEN(UNIT=21,ACCESS='SEQOUT',FILE='PLTM13.DAT')
 7701 FORMAT (////' SUM OF STATE PROBABILITIES')
 7702 FORMAT ('OEXPECTED NO. IN QUEUE')
 7703 FORMAT ('OEXPECTED DELAY')
 7705 FORMAT ('OELAPSED TIME IN MINUTES')
 7706 FORMAT ('OMAXIMUM QUEUE LENGTH')
 7707 FORMAT ('OARRIVAL RATE CUSTOMERS/MIN')
 7708 FORMAT('OEXPECTED NUMBER OF REJECTED TRAFFIC')
 7709 FORMAT('OEXPECTED NUMBER IN SYSTEM')
 8800 FORMAT ('1',F7.3,14F8.3)
 8801 FORMAT (15F8.4)
 8802 FORMAT (15F8.3)
 8805 FORMAT(15F8.4)
 9900 FORMAT (3I)
 9901 FORMAT (2F)
 9903 FORMAT (1I,1F,2I)
 9907 FORMAT(2F,2I)
 3301 FORMAT(3I5,D10.3)
 3302 FORMAT(12D10.2,/,12D10.2)
C     NC=NUMBER OF CASES
C     QM=MAXIMUM NUMBER OF EQUATIONS
C     INT=TIME INTERVAL BETWEEN SPECIFIED LAMBDA VALUES
      READ(5,9900)NC,QM,INT
      DO 1000 CASES=1,NC
C     Q=EQUILIBRIUM EXPECTED QUEUE LENGTH
C     D=EQUILIBRIUM EXPECTED DELAY
      READ(5,9907) Q,D,L,R
C     MU IN CUSTOMERS/HOUR
C     STORE=TIME INTERVAL BETWEEN PRINTED OUTPUT
C     MINS=TOTAL RUN LENGTH
      READ (5,9903) K,MU,STORE ,MINS
      ILAM=MAXO(2,MINS/INT+1)
C     ARR(I)=SPECIFIED LAMBDA VALUES(CUSTOMERS/HOUR)
      READ (5,9901) (ARR(I), I=1,ILAM)
      DO 200 I=1,ILAM
  200 ARR(I)=ARR(I)/60.DO
```

```
      KPLUS=K+1
      QMS=20
      PER=STORE
      UK=MU/60.D0
      MU=UK/K
      X=0.D0
      TOL=1.D-6
      ICOL=0
      MMAXM=0
      MMAX=50
      PM0=0.D0
      DO 210 I=1,MMAX
  210 PM(I)=0.D0
C     INIT=INITIAL STATE(DETERMINISTIC)
      READ(5,222)INIT
  222 FORMAT(1I)
      Y(INIT)=1.D0
      IND=1
      DO 2000 ITIM=1,MINS
      XEND=DFLOAT(ITIM)
2     N=NM+1
      CALL DVERK(N,QUEUE,X,Y,XEND,TOL,IND,C,QM,W,IER)
3     IF(IND .LT. 0 .OR. IER .GT. 0) GO TO 3000
      IF(DFLOAT(ITIM) .LT. PER)GO TO 2000
      XP=ITIM-.0001
      IHR=INT(XP/INT)+1
      LAM=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
      XP=ITIM-STORE/2.
      IHR=INT(XP/INT)+1
      LAV=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
      ICOL=ICOL+1
      PER=PER+STORE
  202 PMAT0(ICOL)=PM0
      PM65=0.D0
      PM68=PM0
      PM66=0.D0
      IF(K .GT. 1) GO TO 400
      IF(R .GT. 1) GO TO 320
      DO 314 I=1,L
      PMAT(I,ICOL)=PM(I)
      PM66=PM66+PM(L+I-1)
 314  PM68=PM68+PM(I)
      DO 310 I=1,L
      KPLUS=L+I
      DO 315 J=KPLUS,MMAX,L
      PM68=PM68+PM(J)
      INW=J/L-1
      IF(J .LE. QMS) PMAT(J,ICOL)=PM(J)
      PM66=PM66+PM(J)
  315 PM65=PM65+DFLOAT(INW)*PM(J)
 310  CONTINUE
      PMAT(QMS+3,ICOL)=(PM66+PM65)/UK
      GO TO 350
 400  PM66=PM(K)
```

```
C     CALCULATION OF Q AND D FOR M/M/K SYSTEM
      PEMPTY=1.D0
      IF (K .EQ. 1) GO TO 60
      DO 50 I=2,K
      FAC=1.D0
      DO 55 J=2,I
   55 FAC=1.D0/(DFLOAT(J)-1.D0)*FAC
   50 PEMPTY=PEMPTY+FAC*(LAM/MU)**(I-1)
   60 FAC=FAC*(1.D0/DFLOAT(K))
      PEMPTY=PEMPTY+FAC*((LAM/MU)**K)*(UK/(UK-LAM))
      PEMPTY=1.D0/PEMPTY
      Q=PEMPTY*FAC*(K**K)*((LAM/UK)**(K+1))/((1.D0-(LAM/UK))**2)
      D=Q/LAM
      SYSNUM=Q+LAM/MU
      PM67=0.D0
      DO 209 I=1,K
      PM67=I*PM(I)+PM67
      PMAT(I,ICOL)=PM(I)
  209 PM68=PM68+PM(I)
      DO 250 J=KPLUS,MMAX
      PM67=PM67+J*PM(J)
      PVJ=PM(J)
      PM65 =PM65+(J-K)*PVJ
      PM66=PM66+PVJ
      PM68=PM68+PVJ
  250 IF (J .LE. QMS) PMAT(J,ICOL)=PVJ
      PMAT(QMS+3,ICOL)=(PM66+PM65)/UK
      PMAT(QMS+9,ICOL)=PM67
      PMAT(QMS+10,ICOL)=SYSNUM-PM67
      GO TO 350
  320 DO 325 I=1,R
      PMAT(I,ICOL)=PM(I)
      PM66=PM66+I*PM(I)
  325 PM68=PM68+PM(I)
      DO 330 I=1,R
      KPLUS=R+I
      DO 340 J=KPLUS,MMAX,R
      PM68=PM68+PM(J)
      INW=(J-1)/R
      IF(J .LE. QMS) PMAT(J,ICOL)=PM(J)
      PM66=PM66+J*PM(J)
  340 PM65=PM65+DFLOAT(INW)*PM(J)
  330 CONTINUE
      PMAT(QMS+3,ICOL)=PM66/(UK*R)
  350 PMAT(QMS+1,ICOL)=PM68
      PMAT(QMS+2,ICOL)=PM65
      PMAT(QMS+8,ICOL)=Q-PMAT(QMS+2,ICOL)
      PMAT(QMS+4,ICOL)=D-PMAT(QMS+3,ICOL)
      PMAT(QMS+5,ICOL)=DFLOAT(ITIM)
      PMAT(QMS+6,ICOL)=DFLOAT(MMAX)
      PMAT(QMS+7,ICOL)=LAM
      MMAXM=MAX0(MMAXM,MIN0(MMAX,QMS))
      IF (ICOL .LT. 15 .AND. ITIM .NE. MINS) GO TO 201
```

```
 2261 WRITE (20,7701)
 1    WRITE (20,8801) (PMAT(QMS+1,J), J=1,ICOL)
      WRITE (20,7702)
      WRITE (20,8805) (PMAT(QMS+2,J), J=1,ICOL)
      WRITE (20,8805) (PMAT(QMS+8,J), J=1,ICOL)
      WRITE (20,7703)
      WRITE (20,8805) (PMAT(QMS+3,J), J=I,ICOL)
      WRITE (20,8805) (PMAT(QMS+4,J), J=1,ICOL)
      WRITE (20,7705)
      WRITE (20,8802) (PMAT(QMS+5,J), J=1,ICOL)
      WRITE (20,7706)
      WRITE (20,8802) (PMAT(QMS+6,J), J=1,ICOL)
      WRITE (20,7707)
      WRITE (20,8801) (PMAT(QMS+7,J), J=1,ICOL)
      WRITE(20,7709)
      WRITE(20,8805)(PMAT(QMS+9,J),J=1,ICOL)
      WRITE(20,8805)(PMAT(QMS+10,J),J=1,ICOL)
      PROB=Y(N)
      WRITE (20,100) PROB
 100  FORMAT(F8.4)
      ICOL=0
      MMAXM=0
 201  IF (PM(MMAX) .GT. 1.D-8) GO TO 207
  205 IF (MMAX .LE.10 .OR. PM(MMAX-5) .GT. 1.D-8) GO TO 2000
      MMAX=MMAX-5
  206 GO TO 2000
  207 IF (MMAX .GE. QM) GO TO 2000
      IF((MMAX+5).LE.QM) GO TO 2072
      MMAX=MMAX+1
      DO 2071 M=MMAX,QM
2071  PM(M)=0.D0
      MMAX=QM
      GO TO 2000
2072  PM(MMAX+1)=0.D0
      PM(MMAX+2)=0.D0
      PM(MMAX+3)=0.D0
      PM(MMAX+4)=0.D0
      PM(MMAX+5)=0.D0
      MMAX=MMAX+5
 2000 CONTINUE
 1000 CONTINUE
      STOP
 3000 WRITE (20,3301)IND,IER,MMAX,TOL
      WRITE (20,3302)(C(I),I=1,24)
      STOP
      END
      SUBROUTINE QUEUE(N,X,Y,YPRIME)
      IMPLICIT DOUBLE PRECISION(A-H,L,O-Z)
      INTEGER L,LM,LP,R
      DIMENSION Y(351),YPRIME(351)
     -, P(350),DP(350)
      COMMON U,                          ARR(145),UK,PMATO(15),
     -  PMAT(65,15),INT,K,MMAX,L,R
      MMAX=N-1
      P0=Y(1)
      DO 101 I=1,MMAX
  101 P(I)=Y(I+1)
      XP=X-.0001D0
      IHR=INT(XP/INT)+1
      Z=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
      MMAXM=MMAX-1
      IF(L .GT. 1)GO TO 400
      IF(R .GT. 1)GO TO 99
      GO TO 500
C     THESE ARE THE EQUATIONS FOR M/E(R)/1
  99  U=U*R
      DP0 = (-Z*P0) + U*P(1)
       DP(1)=-(Z+U)*P(1)+U*P(2)
       IF(R .LE. 2)GO TO 105
       LM=R-1
       DO 100 I=2,LM
  100 DP(I)=-(Z+U)*P(I)+U*P(I+1)
  105 NL=MMAX-R
       LP=R+1
       DP(R)=-(Z+U)*P(R)+Z*P0+U*P(R+1)
      DO 110 I=LP,NL
  110 DP(I)=-(Z+U)*P(I)+Z*P(I-R)+U*P(I+1)
      IF(R .EQ. 1) GO TO 125
       NLP=NL+1
       DO 120 I=NLP,MMAXM
  120 DP(I)=-U*P(I)+U*P(I+1)+Z*P(I-R)
 125   DP(MMAX)=-U*P(MMAX)+Z*P(MMAX-R)
      U=U/R
      GO TO 300
C     THESE ARE THE EQUATIONS FOR E(L)/M/1
 400  Z=L*Z
      DP0 = (-Z*P0) + U*P(L)
      DP(1)=-(Z+U)*P(1)+Z*P0+U*P(L+1)
      IF(L .EQ. 1) GO TO 205
      DP(1)=DP(1)+U*P(1)
      IF(L .EQ. 2) GO TO 205
       LM=L-1
       DO 200 I=2,LM
  200 DP(I)=-Z*P(I)+Z*P(I-1)+U*P(I+L)
  205 NL=MMAX-L
      ML=MAX0(L,2)
      DO 221 I=ML,NL
  221 DP(I)=-(Z+U)*P(I)+Z*P(I-1)+U*P(I+L)
      IF(L .EQ. 1) GO TO 230
       NLP=NL+1
       DO 220 I=NLP,MMAXM
  220 DP(I)=-(Z+U)*P(I)+Z*P(I-1)
  230  DP(MMAX)=-U*P(MMAX)+Z*P(MMAXM)
      Z=Z/L
      GO TO 300
```

```
C     THESE ARE THE EQUATIONS FOR M/M/K
  500 DP0 = (-Z*P0) + U*P(1)
      IF (K .GE. 2) GO TO 301
      KK = 2
      DP(1) = Z*P0 - (Z+U)*P(1) + U*P(2)
      GO TO 330
  301 DP(1) = Z*P0 -(Z+U)*P(1) + 2*U*P(2)
      DO 310 I = 2,K,1
  310 DP(I) = Z*P(I-1) - (Z+I*U)*P(I) + (I+1)*U*P(I+1)
      KK = K
  330 DO 320 I = KK,MMAXM
  320 DP(I) = Z*P(I-1) - (Z+UK)*P(I) + UK*P(I+1)
      DP(MMAX)=Z*P(MMAXM)-UK*P(MMAX)
 300  YPRIME(1)=DP0
      DO 20 I=1,MMAX
  20  YPRIME(I+1)=DP(I)
      RETURN
      END
```

```
C     NUMERICAL SOLUTION OF FINITE-CAPACITY M/H(2)/1 SYSTEM
      IMPLICIT DOUBLE PRECISION(A-H,L,O-Z)
      INTEGER QM,QMS,STORE,CASES,L,R
      DOUBLE PRECISION MU1,MU2,INC
      EXTERNAL QUEUE
      DIMENSION Y( 351),YPRIME( 351),W( 351,9),C(24)
     -  ,P(175),Q(175)
      COMMON U1,U2,ARR(145),PMATO(15),PMAT(65,15),INT,MMAX,A
      EQUIVALENCE (R,Y(1)),(P(1),Y(2)),(Y(177 ),Q(1)),(NM,MMAX)
      OPEN(UNIT=21,ACCESS='SEQOUT',FILE='MH1.DAT')
      OPEN(UNIT=22,ACCESS='SEQOUT',FILE='PLTMH5.DAT')
 7701 FORMAT (////' SUM OF STATE PROBABILITIES')
 7702 FORMAT ('OEXPECTED NO. A/C IN QUEUE')
C     QM=NUMBER OF STATES
 7705 FORMAT ('OELAPSED TIME IN MINUTES')
 7706 FORMAT ('OMAXIMUM QUEUE LENGTH')
 7707 FORMAT ('OARRIVAL RATE OPS/MIN')
 7708 FORMAT('OEXPECTED NUMBER OF REJECTED TRAFFIC')
 8800 FORMAT ('1',F7.3,14F8.3)
 8801 FORMAT (15F8.3)
 8802 FORMAT (15F8.3)
 8805 FORMAT(15F8.4)
 3301 FORMAT(3I5,D10.3)
 3302 FORMAT(12D10.2,/,12D10.2)
C     NC=NUMBER OF CASES
      READ (5,9900) NC,QM
 9900 FORMAT(2I)
      INT=20000
      DO 1000 CASES=1,NC
C     Q1=EQUILIBRIUM EXPECTED QUEUE LENGTH
      READ (5,9907) Q1
 9907 FORMAT(1F)
C     MU1,MU2 IN CUSTOMERS/HOUR
C     A=PROB TYPE 1 CUSTOMER
C     STORE=TIME INTERVAL BETWEEN PRINTED OUTPUT
C     MINS=TOTAL RUN LENGTH
      READ (5,9903) MU1,MU2,A,STORE,MINS
 9903 FORMAT(3F,2I)
      QMS=20
      ILAM=MAX0(2,MINS/INT+1)
      READ (5,9901) (ARR(I), I=1,ILAM)
C     ARR(I)=LAMBDA(I) IN CUSTOMERS/HOUR
 9901 FORMAT(2F)
C     NM=MAX NUMBER OF CUSTOMERS IN SYSTEM
      READ (5,9910) NM
 9910 FORMAT(1I)
      DO 200 I=1,ILAM
  200 ARR(I)=ARR(I)/60.
      PER=STORE
      U1=MU1/60.
      U2=MU2/60.
      X=0.D0
      TOL=1.D-6
      ICOL=0
      N=2*NM+1
      DO 210 I=2,N
  210 Y(I)=0.
      Y(1)=0.
C     INIT=INITIAL STATE(DETERMINISTIC)
      READ(5,9911)INIT
      Y(INIT)=1.
 9911 FORMAT(1I)
      IND=1
      DO 2000 ITIM=1,MINS
      XEND=DFLOAT(ITIM)
      CALL DVERK(N,QUEUE,X,Y,XEND,TOL,IND,C,QM,W,IER)
      IF(IND .LT. 0 .OR. IER .GT. 0) GO TO 3000
      IF (DFLOAT(ITIM) .LT. PER) GO TO 201
      XP=ITIM-.0001
      IHR=IDINT(XP/INT)+1
      LAM=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
      XP=ITIM-STORE/2.
      IHR=IDINT(XP/INT)+1
      LAV=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
      ICOL=ICOL+1
      PER=PER+STORE
      PM65=0.
      PM68=0.
      DO 33 I=1,N
  33  PM68=PM68+Y(I)
      PM66=1./U1*P(1)+1./U2*Q(1)
      DO 314 I=2,MMAX
      PM65=PM65+(I-1)*(P(I)+Q(I))
 314  PM66=PM66+(A/U1*(I-2)+AC/U2*(I-2))*(P(I)+Q(I))
     -  +1./U1*P(I)+1./U2*Q(I)
  350 PMAT(QMS+1,ICOL)=PM68
      PMAT(QMS+2,ICOL)=PM65
      PMAT(QMS+3,ICOL)=PM66
      PMAT(QMS+8,ICOL)=Q1-PMAT(QMS+2,ICOL)
      PMAT(QMS+5,ICOL)=DFLOAT(ITIM)
      PMAT(QMS+6,ICOL)=DFLOAT(MMAX)
      PMAT(QMS+7,ICOL)=LAM
      IF (ICOL .LT. 15 .AND. ITIM .NE. MINS) GO TO 201
 2261 WRITE(21,7701)
      WRITE(21,8801) (PMAT(QMS+1,J), J=1,ICOL)
      WRITE(21,7702)
      WRITE(21,8805) (PMAT(QMS+2,J), J=1,ICOL)
      WRITE(22,8805) (PMAT(QMS+8,J), J=1,ICOL)
      WRITE(21,7705)
      WRITE(21,8802) (PMAT(QMS+5,J), J=1,ICOL)
      WRITE(21,7706)
       WRITE(21,8802) (PMAT(QMS+6,J), J=1,ICOL)
      WRITE(21,7707)
      WRITE(21,8801) (PMAT(QMS+7,J), J=1,ICOL)
      ICOL=0
      MMAXM=0
  201  CONTINUE
 2000 CONTINUE
 1000 CONTINUE
      STOP
 3000 WRITE(21,3301)IND,IER,MMAX,TOL
      WRITE(21,3302)(C(I),I=1,24)
      STOP
      END
      SUBROUTINE QUEUE(N,X,Y,YPRIME)
      IMPLICIT DOUBLE PRECISION(A-H,L,O-Z)
      DIMENSION Y(351 ),YPRIME(351),P(175),Q(175),DP(175),DQ(175)
      COMMON U1,U2,ARR(145),PMATO(15),PMAT(65,15),INT,MMAX,A
      R=Y(1)
      DO 101 I=1,MMAX
      P(I)=Y(I+1)
  101 Q(I)=Y(MMAX+I+1)
      XP=X-.0001
      IHR=IDINT(XP/INT)+1
      Z=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
      MMAXM=MMAX-1
C
      AC=1.-A
      DR=-Z*R+U1*P(1)+U2*Q(1)
      DP(1)=-(U1+Z)*P(1)+A*Z*R+A*U1*P(2)+A*U2*Q(2)
      DO 300 I=2,MMAXM
  300 DP(I)=-(U1+Z)*P(I)+Z*P(I-1)+A*U1*P(I+1)+A*U2*Q(I+1)
      DP(MMAX)=-U1*P(MMAX)+Z*P(MMAXM)
      DQ(1)=-(U2+Z)*Q(1)+AC*Z*R+AC*U2*Q(2)+AC*U1*P(2)
      DO 310 I=2,MMAXM
  310 DQ(I)=-(U2+Z)*Q(I)+Z*Q(I-1)+AC*U2*Q(I+1)+AC*U1*P(I+1)
      DQ(MMAX)=-U2*Q(MMAX)+Z*Q(MMAXM)
      YPRIME(1)=DR
      DO 20 I=1,MMAX
      YPRIME(I+1)=DP(I)
  20  YPRIME(MMAX+I+1)=DQ(I)
      RETURN
      END
```

```
C     NUMERICAL SOLUTION OF M/PH/1 SYSTEM--SERVICE TIME
C     IS WEIGHTED SUM OF EXPONENTIAL AND THIRD-ORDER
C     ERLANG RANDOM VARIABLES
      IMPLICIT DOUBLE PRECISION(A-H,L,O-Z)
      INTEGER QM,QMS,STORE,CASES
      DOUBLE PRECISION INC
      EXTERNAL QUEUE
      DIMENSION Y(141),YPRIME(141),W(141,9),C(24),P(35),
     *  Q(35),S(35),V(35)
 7701 FORMAT(////' SUM OF STATE PROBABILITIES')
 7702 FORMAT('OEXPECTED QUEUE LENGTH')
 7705 FORMAT('OELAPSED TIME')
 7706 FORMAT('OMAXIMUM QUEUE LENGTH')
 7707 FORMAT('OARRIVAL RATE ')
 7708 FORMAT('OEXPECTED NUMBER REJECTED')
 8801 FORMAT(15F8.3)
 8802 FORMAT(15F8.3)
 8805 FORMAT(15F8.4)
 9900 FORMAT(3I)
 9901 FORMAT(2F)
 9903 FORMAT(2I)
 9907 FORMAT(4F)
 9910 FORMAT(1I)
 3301 FORMAT(3I5,D10.3)
 3302 FORMAT(12D10.2,/,12D10.2)
      OPEN(UNIT=21,ACCESS='SEQOUT',FILE='MP1.DAT')
      OPEN(UNIT=22,ACCESS='SEQOUT',FILE='PLTMP1.DAT')
      EQUIVALENCE(Y(2),P(1)),(Y(37),Q(1)),(Y(72),S(1)),(Y(107),V(1))
      COMMON ARR(145),PMAT0(15),PMAT(65,15),INT,NM,A,U1,U2
C     NC=NUMBER OF CASES
C     QM=NUMBER OF STATES
      READ(5,9900)NC,QM
      INT=20000
      DO 1000 CASES=1,NC
C     Q1=EQUILIBRIUM EXPECTED QUEUE LENGTH
C     A=PROB EXPONENTIAL SERVICE TIME
C     U1=MU OF EXPONENTIAL(CUST/HOUR)
C     U2=MU OF THIRD-ORDER ERLANG(CUST/HOUR)
      READ(5,9907)Q1,A,U1,U2
      U1=U1/60.D0
      U2=U2/60.D0
      READ(5,9903)STORE,MINS
      QMS=20
      ILAM=MAX0(2,MINS/INT+1)
C     ARR(I)=LAMBDA(I) IN CUSTOMERS/HOUR
      READ(5,9901)(ARR(I),I=1,ILAM)
C     NM=MAXIMUM NUMBER OF CUSTOMERS IN SYSTEM
      READ(5,9910)NM
      DO 200 I=1,ILAM
  200 ARR(I)=ARR(I)/60.
      K=1
      KPLUS=K+1
      PER=STORE
      X=0.D0
      TOL=1.D-6
      ICOL=0
      MMAX=NM
      MMAXM=0
      N=4*NM+1
      Y(1)=1.D0
      DO 210 I=2,N
  210 Y(I)=0.D0
      IND=1
      DO 2000 ITIM=1,MINS
      XEND=DFLOAT(ITIM)
      CALL DVERK(N,QUEUE,X,Y,XEND,TOL,IND,C,QM,W,IER)
      IF(IND .LT. 0 .OR. IER .GT. 0)GO TO 3000
      IF(DFLOAT(ITIM) .LT. PER)GO TO 201
      XP=ITIM-.0001D0
      IHR=IDINT(XP/INT)+1
      LAM=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
      XP=ITIM-STORE/2.D0
      IHR=IDINT(XP/INT)+1
      LAV=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
      ICOL=ICOL+1
      PER=PER+STORE
      PM65=0.D0
      PM68=Y(1)
      DO 900 I=1,NM
      PM68=P(I)+Q(I)+S(I)+V(I)+PM68
  900 PM65=PM65+(I-1)*(P(I)+Q(I)+S(I)+V(I))
      PMAT(QMS+3,ICOL)=0.D0
      PMAT(QMS+1,ICOL)=PM68
      PMAT(QMS+2,ICOL)=PM65
      PMAT(QMS+8,ICOL)=Q1-PMAT(QMS+2,ICOL)
      PMAT(QMS+5,ICOL)=DFLOAT(ITIM)
      PMAT(QMS+6,ICOL)=DFLOAT(MMAX)
      PMAT(QMS+7,ICOL)=LAM
      IF(ICOL .LT. 15 .AND. ITIM .NE. MINS)GO TO 201
      WRITE(21,7701)
      WRITE(21,8801)(PMAT(QMS+1,J),J=1,ICOL)
      WRITE(21,7702)
      WRITE(21,8805)(PMAT(QMS+2,J),J=1,ICOL)
      WRITE(22,8805)(PMAT(QMS+8,J),J=1,ICOL)
      WRITE(21,7705)
      WRITE(21,8802)(PMAT(QMS+5,J),J=1,ICOL)
      WRITE(21,7706)
      WRITE(21,8802)(PMAT(QMS+6,J),J=1,ICOL)
      WRITE(21,7707)
      WRITE(21,8801)(PMAT(QMS+7,J),J=1,ICOL)
      PMAT(QMS+9,ICOL)=P(NM)+Q(NM)+S(NM)+V(NM)
      WRITE(21,8805)(PMAT(QMS+9,J),J=1,ICOL)
      ICOL=0
  201 CONTINUE
 2000 CONTINUE
 1000 CONTINUE
      STOP
 3000 WRITE(21,3301)IND,IER,MMAX,TOL
      WRITE(21,3302)(C(I),I=1,24)
      STOP
      END
      SUBROUTINE QUEUE(N,X,Y,YPRIME)
      IMPLICIT DOUBLE PRECISION(A-H,L,O-Z)
      DIMENSION Y(141),YPRIME(141),P(35),Q(35),S(35),V(35),
     *  DP(35),DQ(35),DS(35),DV(35)
      COMMON ARR(145),PMATO(15),PMAT(65,15),INT,NM,A,U1,U2
      R=Y(1)
      DO 5 I=1,NM
      P(I)=Y(I+1)
      Q(I)=Y(NM+I+1)
      S(I)=Y(2*NM+I+1)
    5 V(I)=Y(3*NM+I+1)
      XP=X-.0001D0
      IHR=IDINT(XP/INT)+1
      Z=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
C
C       THESE ARE THE EQUATIONS FOR AN M/PH/1 SYSTEM
C
      AB=1.D0-A
      DR=-Z*R+U1*P(1)+U2*Q(1)
      DP(1)=-(Z+U1)*P(1)+A*Z*R+A*U1*P(2)+A*U2*Q(2)
      DP(NM)=-U1*P(NM)+Z*P(NM-1)
      DQ(1)=-(Z+U2)*Q(1)+U2*S(1)
      DQ(NM)=-U2*Q(NM)+Z*Q(NM-1)+U2*S(NM)
      DS(1)=-(Z+U2)*S(1)+U2*V(1)
      DS(NM)=-U2*S(NM)+Z*S(NM-1)+U2*V(NM)
      DV(1)=-(Z+U2)*V(1)+AB*Z*R+AB*U2*Q(2)+AB*U1*P(2)
      DV(NM)=-U2*V(NM)+Z*V(NM-1)
      NMM=NM-1
      DO 10 I=2,NMM
      DP(I)=-(Z+U1)*P(I)+Z*P(I-1)+A*U1*P(I+1)+A*U2*Q(I+1)
      DQ(I)=-(Z+U2)*Q(I)+Z*Q(I-1)+U2*S(I)
      DS(I)=-(Z+U2)*S(I)+Z*S(I-1)+U2*V(I)
   10 DV(I)=-(Z+U2)*V(I)+Z*V(I-1)+AB*U2*Q(I+1)+AB*U1*P(I+1)
      YPRIME(1)=DR
      DO 50 I=1,NM
      YPRIME(NM+I+1)=DQ(I)
      YPRIME(2*NM+I+1)=DS(I)
      YPRIME(3*NM+I+1)=DV(I)
   50 YPRIME(I+1)=DP(I)
      RETURN
      END
```

```
C     NUMERICAL SOLUTION OF FINITE-CAPACITY E(L)/E(R)/1 SYSTEM
      IMPLICIT DOUBLE PRECISION(A-H,L,O-Z)
      INTEGER QM,QMS,STORE,CASES,L,R
      DOUBLE PRECISION MU,INC
      EXTERNAL QUEUE
      DIMENSION Y(122),YPRIME(122),C(24),P(2,61),W(122,9)
      EQUIVALENCE (Y(1),P(1,1))
      COMMON MU,                          ARR(145),UK,
     -  PMAT(65,15),INT,K,      L,R,NM
      OPEN(UNIT=21,ACCESS='SEQOUT',FILE='EE1.DAT')
      OPEN(UNIT=22,ACCESS='SEQOUT',FILE='PLTEE1.DAT')
 7701 FORMAT (////' SUM OF STATE PROBABILITIES')
 7702 FORMAT ('OEXPECTED NO. CUSTOMERS IN QUEUE')
 7705 FORMAT ('OELAPSED TIME IN MINUTES')
 7706 FORMAT ('OMAXIMUM QUEUE LENGTH')
 7707 FORMAT ('OARRIVAL RATE CUSTOMERS/MIN')
 7708 FORMAT('OEXPECTED NUMBER OF REJECTED TRAFFIC')
 7709 FORMAT('OPROBABILITY OF FULL SYSTEM')
 7710 FORMAT('OPROBABILITY NM CUSTOMERS IN SYSTEM')
 8800 FORMAT ('1',F7.3,14F8.3)
 8801 FORMAT (15F8.3)
 8802 FORMAT (15F8.3)
 8805 FORMAT(15F8.4)
 3301 FORMAT(3I5,D10.3)
 3302 FORMAT(12D10.2,/,12D10.2)
C     NC=NUMBER OF CASES
C     QM=NUMBER OF STATES
      READ (5,*) NC,QM
      INT=20000
      DO 1000 CASES=1,NC
C     Q=EQUILIBRIUM EXPECTED QUEUE LENGTH
      READ(5,*) Q,L,R
C     STORE=TIME INTERVAL BETWEEN PRINTED OUTPUT
C     MINS=TOTAL RUN LENGTH
      READ (5,*) K,MU,STORE ,MINS
      QMS=20
      ILAM=MAX0(2,MINS/INT+1)
C     ARR(I)=LAMBDA(I) IN CUSTOMERS/HOUR
      READ (5,*) (ARR(I), I=1,ILAM)
C     NM=MAXIMUM NUMBER OF CUSTOMERS IN SYSTEM
      READ (5,*) NM
      DO 200 I=1,ILAM
  200 ARR(I)=ARR(I)/60.
      KPLUS=K+1
      PER=STORE
      UK=MU/60.
      MU=UK/K
      X=0.D0
      TOL=1.D-6
      ICOL=0
      N=(NM*R+1)*L
      QM=N
      Y(1)=1.
      DO 210 I=2,N
 210 Y(I)=0.
     IND=1
     DO 2000 ITIM=1,MINS
     XEND=DFLOAT(ITIM)
     CALL DVERK(N,QUEUE,X,Y,XEND,TOL,IND,C,QM,W,IER)
     IF(IND .LT. 0 .OR. IER .GT. 0) GO TO 3000
     IF (DFLOAT(ITIM) .LT. PER) GO TO 201
     XP=ITIM-.0001
     IHR=IDINT(XP/INT)+1
     LAM=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
     XP=ITIM-STORE/2.
     IHR=IDINT(XP/INT)+1
     LAV=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
     ICOL=ICOL+1
     PER=PER+STORE
     NRP=NM*R+1
     PM68=0.
     PM66=0.
     PM65=0.
     NT=2*R
     DO 40 I=1,L
     DO 50 J=NT,NRP
     M=J/R-2
     PM65=PM65+M*P(I,J)
 50  PM66=PM66+P(I,J)*(J-1)
 40  CONTINUE
     DO 41 I=1,N
  41 PM68=PM68+Y(I)
 350 PMAT(QMS+1,ICOL)=PM68
     PMAT(QMS+2,ICOL)=PM65
     PMAT(QMS+8,ICOL)=Q-PMAT(QMS+2,ICOL)
     PMAT(QMS+5,ICOL)=DFLOAT(ITIM)
     PMAT(QMS+6,ICOL)=DFLOAT(NM)
     PMAT(QMS+7,ICOL)=LAM
     PMAT(QMS+9,ICOL)=P(1,NRP)
     NMRP=(NM-1)*R+2
     PMAT(QMS+10,ICOL)=0.
     DO 60 J=NMRP,NRP
 60  PMAT(QMS+10,ICOL)=PMAT(QMS+10,ICOL)+P(1,J)
     IF (ICOL .LT. 15 .AND. ITIM .NE. MINS) GO TO 201
2261 WRITE(21,7701)
     WRITE(21,8801) (PMAT(QMS+1,J), J=1,ICOL)
     WRITE(21,7702)
     WRITE(21,8805) (PMAT(QMS+2,J), J=1,ICOL)
     WRITE(22,8805) (PMAT(QMS+8,J), J=1,ICOL)
     WRITE(21,7705)
     WRITE(21,8802) (PMAT(QMS+5,J), J=1,ICOL)
     WRITE(21,7706)
     WRITE(21,8802) (PMAT(QMS+6,J), J=1,ICOL)
     WRITE(21,7707)
     WRITE(21,8801) (PMAT(QMS+7,J), J=1,ICOL)
     WRITE(21,7709)
     WRITE(21,8805)(PMAT(QMS+9,J),J=1,ICOL)
     WRITE(21,7710)
```

```
      WRITE(21,8805)(PMAT(QMS+10,J),J=1,ICOL)
      ICOL=0
  201  CONTINUE
 2000 CONTINUE
 1000 CONTINUE
      STOP
 3000 WRITE(21,3301)IND,IER,MMAX,TOL
      WRITE(21,3302)(C(I),I=1,24)
      STOP
      END
      SUBROUTINE QUEUE(N,X,Y,YPRIME)
      IMPLICIT DOUBLE PRECISION(A-H,L,O-Z)
      INTEGER L,LM,LP,R
      DIMENSION Y(122),YPRIME(122),P(2,61),DP(2,61)
      COMMON U,                     ARR(145),UK,
     -  PMAT(65,15),INT,K,     L,R,NM
      NRP=NM*R+1
      DO 500 I=1,L
      DO 510 J=1,NRP
 510  P(I,J)=Y((J-1)*L+I)
 500  CONTINUE
      XP=X-.0001
      IHR=IDINT(XP/INT)+1
      Z=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
C     THESE ARE THE EQUATIONS FOR E(L)/E(R)/1
      ZL=L*Z
      UR=R*U
      DP(1,1)=-ZL*P(1,1)+UR*P(1,2)
      IF (L .EQ. 1) GO TO 11
      DO 10 I=2,L
  10  DP(I,1)=-ZL*P(I,1)+ZL*P(I-1,1)+UR*P(I,2)
 11   NR=NM*R
      NMR=(NM-1)*R+1
      DO 20 J=2,NMR
  20  DP(1,J)=-(ZL+UR)*P(1,J)+UR*P(1,J+1)
      NMRP=NMR+1
      IF (R .EQ. 1) GO TO 41
      DO 40 J=NMRP,NR
  40  DP(1,J)=-UR*P(1,J)+UR*P(1,J+1)
 41   DP(1,NR+1)=-UR*P(1,NR+1)+ZL*P(L,NMR)
      NR=NR-1
      DO 30 J=R,NR
  30  DP(1,J+1)=DP(1,J+1)+ZL*P(L,J+1-R)
      NRP=NR+2
      IF (L .EQ. 1) GO TO 51
      NMRM=NMR-1
      DO 50 I=2,L
      DO 60 J=2,NMRM
  60  DP(I,J)=-(ZL+UR)*P(I,J)+ZL*P(I-1,J)+UR*P(I,J+1)
  50  DP(I,NMR)=-(ZL+UR)*P(I,NMR)+ZL*P(I-1,NMR)
      NRP=NM*R+1
 51 DO 100 I=1,L
    DO 110 J=1,NRP
110 YPRIME((J-1)*L+I)=DP(I,J)
100 CONTINUE
    RETURN
    END
```

```
C     NUMERICAL SOLUTION OF FINITE-CAPACITY M/M/K,
C     E(L)/M/1 AND M/E(R)/1 SYSTEMS
      IMPLICIT DOUBLE PRECISION(A-H,L,O-Z)
      INTEGER QM,QMS,STORE,CASES,L,R
      DOUBLE PRECISION MU,INC
      EXTERNAL QUEUE
      DIMENSION Y(351),YPRIME(351),W(351,9),C(24)
     -  ,PM(350),DP(350)
      EQUIVALENCE(PMO,Y(1)),(DPO,YPRIME(1))
     -  ,(PM(1),Y(2)),(DP(1),YPRIME(2)),(MMAX,NM)
      COMMON PMAT(65,15),ARR(145),PMATO(15),MU,UK,INT,K,MMAX,L,R
 7701 FORMAT (////' SUM OF STATE PROBABILITIES')
 7702 FORMAT ('OEXPECTED NO. A/C IN QUEUE')
 7703 FORMAT ('OAVERAGE DELAY PER A/C')
 7705 FORMAT ('OELAPSED TIME IN MINUTES')
 7706 FORMAT ('OMAXIMUM QUEUE LENGTH')
 7707 FORMAT ('OARRIVAL RATE OPS/MIN')
 7708 FORMAT('OEXPECTED NUMBER OF REJECTED TRAFFIC')
 8800 FORMAT ('1',F7.3,14F8.3)
 8801 FORMAT (15F8.3)
 8802 FORMAT (15F8.2)
 8805 FORMAT(15F8.4)
 3301 FORMAT(3I5,F10.3)
 3302 FORMAT(12F10.2,/,12F10.2)
      OPEN (UNIT=21,ACCESS='SEQOUT',FILE='FIN.DAT')
      OPEN(UNIT=22,ACCESS='SEQOUT',FILE='PLTFN3.DAT')
C     NC=NUMBER OF CASES
C     QM=NUMBER OF STATES
      READ (5,9900) NC,QM
 9900 FORMAT(2I)
      INT=20000
      DO 1000 CASES=1,NC
C     Q1=EQUILIBRIUM EXPECTED QUEUE LENGTH
C     D=EQUILIBRIUM EXPECTED DELAY
      READ(5,9907)Q1,D,L,R
 9907 FORMAT(2F,2I)
C     MU IN CUSTOMERS/HOUR
C     STORE=TIME INTERVAL BETWEEN PRINTED OUTPUT
C     MINS=TOTAL RUN LENGTH
      READ (5,9903) K,MU,STORE ,MINS
      ILAM=MAXO(2,MINS/INT+1)
 9903 FORMAT(1I,1F,2I)
C     ARR(I)=LAMBDA(I) IN CUSTOMERS/HOUR
      READ (5,9901) (ARR(I), I=1,ILAM)
 9901 FORMAT(2F)
C     NM=MAXIMUM NUMBER OF CUSTOMERS IN SYSTEM
      READ (5,9910) NM
 9910 FORMAT(1I)
      DO 200 I=1,ILAM
  200 ARR(I)=ARR(I)/60.DO
      KPLUS=K+1
      PER=STORE
      UK=MU/60.DO
      MU=UK/K
      X=0.D0
      TOL=1.D-6
      ICOL=0
      N=NM+1
      MMAX=N-1
      MMAXM=MMAX-1
      PM0=1.D0
      DO 210 I=1,MMAX
  210 PM(I)=0.D0
      IND=1
      DO 2000 ITIM=1,MINS
      XEND=DFLOAT(ITIM)
      CALL DVERK(N,QUEUE,X,Y,XEND,TOL,IND,C,QM,W,IER)
      IF(IND .LT. 0 .OR. IER .GT. 0) GO TO 3000
      IF (DFLOAT(ITIM) .LT. PER) GO TO 201
      XP=ITIM-.0001
      IHR=IDINT(XP/INT)+1
      LAM=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
      XP=ITIM-STORE/2.
      IHR=IDINT(XP/INT)+1
      LAV=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
      ICOL=ICOL+1
      PER=PER+STORE
  202 PMAT0(ICOL)=PM0
      PM65=0.D0
      PM68=PM0
      PM66=0.D0
      IF(K .GT. 1) GO TO 400
      IF(R .GT. 1) GO TO 320
      DO 314 I=1,L
      PMAT(I,ICOL)=PM(I)
      PM66=PM66+PM(L+I-1)
 314  PM68=PM68+PM(I)
      DO 310 I=1,L
      KPLUS=L+I
      DO 315 J=KPLUS,MMAX,L
      PM68=PM68+PM(J)
      INW=J/L-1
      IF(J .LE. QMS) PMAT(J,ICOL)=PM(J)
      PM66=PM66+PM(J)
  315 PM65=PM65+DFLOAT(INW)*PM(J)
 310  CONTINUE
      PMAT(QMS+3,ICOL)=(PM66+PM65)/UK
      GO TO 350
 400  PM66=PM(K)
      DO 209 I=1,K
      PMAT(I,ICOL)=PM(I)
  209 PM68=PM68+PM(I)
      DO 250 J=KPLUS,MMAX
      PVJ=PM(J)
      PM65 =PM65+(J-K)*PVJ
      PM66=PM66+PVJ
      PM68=PM68+PVJ
 250 IF (J .LE. QMS) PMAT(J,ICOL)=PVJ
     PMAT(QMS+3,ICOL)=(PM66+PM65)/UK
     GO TO 350
 320 DO 325 I=1,R
     PMAT(I,ICOL)=PM(I)
     PM66=PM66+I*PM(I)
 325 PM68=PM68+PM(I)
     DO 330 I=1,R
     KPLUS=R+I
     DO 340 J=KPLUS,MMAX,R
     PM68=PM68+PM(J)
     INW=(J-1)/R
     IF(J .LE. QMS) PMAT(J,ICOL)=PM(J)
     PM66=PM66+J*PM(J)
 340 PM65=PM65+DFLOAT(INW)*PM(J)
 330 CONTINUE
     PMAT(QMS+3,ICOL)=PM66/(UK*R)
 350 PMAT(QMS+1,ICOL)=PM68
     PMAT(QMS+2,ICOL)=PM65
     PMAT(QMS+8,ICOL)=Q1-PMAT(QMS+2,ICOL)
     PMAT(QMS+4,ICOL)=D-PMAT(QMS+3,ICOL)
     PMAT(QMS+5,ICOL)=DFLOAT(ITIM)
     PMAT(QMS+6,ICOL)=DFLOAT(MMAX)
     PMAT(QMS+7,ICOL)=LAM
     MMAXM=MAX0(MMAXM,MIN0(MMAX,QMS))
     IF (ICOL .LT. 15 .AND. ITIM .NE. MINS) GO TO 201
     IF (MMAXM .EQ. 0) GO TO 2261
2261 WRITE(21,7701)
     WRITE(21,8801) (PMAT(QMS+1,J), J=1,ICOL)
     WRITE(21,7702)
     WRITE(21,8805) (PMAT(QMS+2,J), J=1,ICOL)
     WRITE(22,8805) (PMAT(QMS+8,J), J=1,ICOL)
     WRITE(21,7703)
     WRITE(21,8805) (PMAT(QMS+3,J), J=1,ICOL)
     WRITE(21,8805) (PMAT(QMS+4,J) ,J=1,ICOL)
     WRITE (21,7705)
     WRITE (21,8802) (PMAT(QMS+5,J), J=1,ICOL)
     WRITE (21,7706)
     WRITE (21,8802) (PMAT(QMS+6,J), J=1,ICOL)
     WRITE (21,7707)
     WRITE (21,8801) (PMAT(QMS+7,J), J=1,ICOL)
     ICOL=0
     MMAXM=0
 201  CONTINUE
2000 CONTINUE
1000 CONTINUE
     STOP
3000 WRITE(21,3301)IND,IER,MMAX,TOL
     WRITE(21,3302)(C(I),I=1,24)
     STOP
     END
```

```
      SUBROUTINE QUEUE(N,X,Y,YPRIME)
      IMPLICIT DOUBLE PRECISION(A-H,L,O-Z)
      INTEGER L,R
      DIMENSION Y(351),YPRIME(351)
     -, P(350),DP(350)
      COMMON PMAT(65,15),ARR(145),PMATO(15),U,UK,INT,K,MMAX,L,R
      MMAXM=MMAX-1
      PO=Y(1)
      DO 101 I=1,MMAX
  101 P(I)=Y(I+1)
      XP=X-.0001D0
      IHR=IDINT(XP/INT)+1
      Z=ARR(IHR)+(ARR(IHR+1)-ARR(IHR))*DMOD(XP,DFLOAT(INT))/INT
      IF(L .GT. 1)GO TO 400
      IF(R .GT. 1)GO TO 500
C
C     THESE ARE THE EQUATIONS FOR M/M/K
      DPO = (-Z*PO) + U*P(1)
      IF (K .GE. 2) GO TO 301
      KK = 2
      DP(1) = Z*PO - (Z+U)*P(1) + U*P(2)
      GO TO 330
  301 DP(1) = Z*PO -(Z+U)*P(1) + 2*U*P(2)
      DO 310 I = 2,K,1
  310 DP(I) = Z*P(I-1) - (Z+I*U)*P(I) + (I+1)*U*P(I+1)
      KK = K
  330 DO 320 I = KK,MMAXM
  320 DP(I) = Z*P(I-1) - (Z+UK)*P(I) + UK*P(I+1)
      DP(MMAX)=Z*P(MMAXM)-UK*P(MMAX)
      GO TO 600
C     THESE ARE THE EQUATIONS FOR E(L)/M/1
 400  Z=L*Z
      DPO = (-Z*PO) + U*P(L)
      DP(1)=-(Z+U)*P(1)+Z*PO+U*P(L+1)
      IF(L .EQ. 1) GO TO 305
      DP(1)=DP(1)+U*P(1)
      IF(L .EQ. 2) GO TO 305
       LM=L-1
       DO 200 I=2,LM
  200 DP(I)=-Z*P(I)+Z*P(I-1)+U*P(I+L)
  305 NL=MMAX-L
      ML=MAXO(L,2)
      DO 220 I=ML,NL
  220 DP(I)=-(Z+U)*P(I)+Z*P(I-1)+U*P(I+L)
      IF(L .EQ. 1) GO TO 230
       NLP=NL+1
       DO 221 I=NLP,MMAXM
  221 DP(I)=-(Z+U)*P(I)+Z*P(I-1)
  230  DP(MMAX)=-U*P(MMAX)+Z*P(MMAXM)
      Z=Z/L
      GO TO 600
```

```
C     THESE ARE THE EQUATIONS FOR M/E(R)/1
 500  U=U*R
      DP0 = (-Z*P0) + U*P(1)
       DP(1)=-(Z+U)*P(1)+U*P(2)
       IF(R .LE. 2)GO TO 105
       LM=R-1
       DO 100 I=2,LM
  100 DP(I)=-(Z+U)*P(I)+U*P(I+1)
  105 NL=MMAX-R
       LP=R+1
       DP(R)=-(Z+U)*P(R)+Z*P0+U*P(R+1)
      DO 110 I=LP,NL
  110 DP(I)=-(Z+U)*P(I)+Z*P(I-R)+U*P(I+1)
      IF(R .EQ. 1) GO TO 125
       NLP=NL+1
       DO 120 I=NLP,MMAXM
  120 DP(I)=-U*P(I)+U*P(I+1)+Z*P(I-R)
 125   DP(MMAX)=-U*P(MMAX)+Z*P(MMAX-R)
      U=U/R
  600 YPRIME(1)=DP0
      DO 20 I=1,MMAX
   20 YPRIME(I+1)=DP(I)
      RETURN
      END
```

```
C     NUMERICAL SOLUTION OF M/D/K SYSTEM
      IMPLICIT DOUBLE PRECISION (A-G,L,O-Z)
      INTEGER CASES,STORE,QM,QMD,QMM,TPRINT
      DOUBLE PRECISION MU,MUH
      DIMENSION ARR(145),PMAT(16,10),PVEC(360),MUH(40)
      COMMON /A1/INT,INTER,INTMU,K,STORE/A2/W,Q1,MUH,ARR
      COMMON /A3/WTC,ST,IP20/A4/PMAT/A5/MPLUS/A6/KCT/A7/MMAX
      COMMON TPRINT
      OPEN(UNIT=21,ACCESS='SEQOUT',FILE='MDK.DAT')
      OPEN(UNIT=22,ACCESS='SEQOUT',FILE='PLTMDK.DAT')
 9900 FORMAT(4I,2F)
 9901 FORMAT(I4,X,2I3,X,3I1,X,I4)
 9903 FORMAT(2F)
 7701 FORMAT('0TIME')
 7704 FORMAT('0EXPECTED DELAY')
 7705 FORMAT('0ARRIVAL RATE')
 7706 FORMAT('0EXPECTED QUEUE LENGTH')
 8800 FORMAT ('1',6X,20A4,I3,' RWAYS, CAP = ',F5.1,' OPS/HR')
 8801 FORMAT(15F8.2)
 8804 FORMAT(15F8.4)
 8805 FORMAT(15F8.4)
C     NC=NUMBER OF CASES
      READ(5,9910)NC
 9910 FORMAT(1I)
      TPRINT=0
      INT=1400
      QMD=350
      INTMU=4444
      IWTC=1
      KO=1
      INTER=1
      QMM=350
      WTC=1.D0
      DO 3 CASES=1,NC
C     STORE=TIME INTERVAL BETWEEN PRINTED OUTPUT
C     MINS=TOTAL RUN LENGTH
C     Q1=EQUILIBRIUM EXPECTED QUEUE LENGTH
C     W=EQUILIBRIUM EXPECTED DELAY
      READ(5,23)K,STORE,MINS,Q1,W
   23 FORMAT(3I,2F)
      IQS=0
      ILAM=2+MINS/INT
      IMU=2+MINS/INTMU
C     MUH(I) IN CUSTOMERS/HOUR
      READ (5,9903) (MUH(I),I=1,IMU)
      IF(STORE .LT. 60./MUH(1))GO TO 2222
C     ARR(I)=LAMBDA(I) IN CUSTOMERS/HOUR
      READ(5,9903)(ARR(I),I=1,ILAM)
      DO 1 I=1,ILAM
    1 ARR(I)=ARR(I)/60.D0
      KPLUS=K+1
      XSTRT=0.D0
      IQ=0
      DO 233 I=1,360
  233 PVEC(I)=0.DO
C     INIT=INITIAL STATE(DETERMINISTIC)
      READ(5,999)INIT
  999 FORMAT(1I)
      PVEC(INIT)=1.DO
      PRINT=MINO(MINS,15*STORE)
      XFIN=XSTRT+PRINT
100   XFIN=DMIN1(XFIN,DFLOAT(MINS))
    2 CALL MDK(XSTRT,XFIN,PVE C,QMD,WTMD)
      WRITE(21,7701)
      WRITE(21,8801) (PMAT(I,2),I=1,KCT)
      WRITE(21,7704)
      WRITE(21,8804) (PMAT(I,1),I=1,KCT)
      WRITE(21,8804) K,(PMAT(I,8),I=1,KCT)
      WRITE(21,7705)
      WRITE(21,8805) (PMAT(I,4),I=1,KCT)
      WRITE(21,7706)
      WRITE(21,8856) (PMAT(I,5),I=1,KCT)
      WRITE(22,8857) (PMAT(I,7),I=1,KCT)
 8857 FORMAT(15F8.4)
 8856 FORMAT(15F8.4)
      XFIN=DMIN1(XFIN+PRINT,DFLOAT(MINS))
      IF(XSTRT .LT. MINS-STORE) GO TO 2
    3 CONTINUE
      STOP
 2222 WRITE(21,8888)
 8888 FORMAT(/,'STORE MUST BE MULTIPLE OF EXP SERVICE TIME')
      STOP
      END
      SUBROUTINE MDK(XSTRT,XFIN,PVEC,QM,WTMD)
      IMPLICIT DOUBLE PRECISION(A-G,O-Z)
      INTEGER LMAX,TPRINT
      INTEGER QM,STORE
      DOUBLE PRECISION LAM,L,MUH,LAMSUM,LM
      DIMENSION PM(350),PVEC(360),L(100),AMAT(16),DMAT(16),QMAT(16),
     -          GMAT(16),MUH(40),ARR(145),PMAT(16,10),
     -         BMAT(16),CMAT(16)
      COMMON /A1/INT,INTER,INTMU,K,STORE/A2/W,Q1,MUH,ARR
      COMMON /A3/WTC,ST,IP20/A4/PMAT/A5/MPLUS/A6/KCT/A7/MMAX
      COMMON TPRINT
      EQUIVALENCE (PMAT(1,1),DMAT(1)),(PMAT(1,3),GMAT(1))
      EQUIVALENCE (PMAT(1,4),AMAT(1)),(PMAT(1,5),BMAT(1))
      EQUIVALENCE (PMAT(1,7),CMAT(1)),(PMAT(1,8),QMAT(1))
      TPRINT=INT(XSTRT)
      MMAX=0
      KPLUS=K+1
      TIME=XSTRT
      TABLE=IDINT(0.5DO+TIME/STORE)*STORE+STORE
      KCT=0
    2 IHM=IDINT(TIME/INTMU)+1
      U=MUH(IHM)+(MUH(IHM+1)-MUH(IHM))*DMOD(TIME,DFLOAT(INTMU))/INTMU
      ST=K*60.DO/(U*WTC)
      IP20=(20.DO*U*WTC/60.DO)+K-.5DO
      TIME=TIME+ST
      IHR=IDINT(TIME/INT)+1
      LAM=ARR(IHR)+
     -    (ARR(IHR+1)-ARR(IHR))*INTER*DMOD(TIME,DFLOAT(INT))/INT
      LAM=LAM*ST*WTC
      LMAX=1
      EXPLM=DEXP(-LAM)
      L(1)=LAM
   10 LMAX=LMAX+1
      L(LMAX)=LAM/DFLOAT(LMAX)*L(LMAX-1)
      IF (L(LMAX) .GE. 1.D-6) GO TO 10
      PM0=PVEC(K)*EXPLM
      M=0
      PMM=PM0
      PM66=0.D0
   20 M=M+1
      PMI=0.D0
      JJ=MAX0(K,M-LMAX+K)
      J=MIN0(M,LMAX)
   21 PMI=PMI+PVEC(JJ)*L(J)
      JJ=JJ+1
      J=J-1
      IF(J .GE. 1) GO TO 21
      PMI=(PMI+PVEC(JJ))*EXPLM
      PMM=PMM+PMI
      PM66=PM66+M*PMI
      PM(M)=PMI
      IF (PMI .GT. 1.D-7 .AND. M .LT. QM) GO TO 20
      IF (PMM .LT. .1 .OR. M .LE. K) GO TO 20
      MMAX=MAX0(MMAX,M)
      PVEC(K)=PM0
      DO 30 I=1,K
      PM66=PM66-I*PM(I)
      PVEC(K)=PVEC(K)+PM(I)/PMM
   30 CONTINUE
      PM66=PM66+K*PM(K)+(1-(PVEC(K)-PM(K)))*(0.5D0-K)
      DO 40 J=KPLUS,M
  40  PVEC(J)=PM(J)/PMM
      PP20=0.D0
      DO 11 I=K,M
  11  PP20=PP20+(I-K)*PVEC(I)
      IF (TIME .LT. TABLE-ST/2.) GO TO 1
      TABLE=TABLE+STORE
      KCT=KCT+1
      DMAT(KCT)=PM66*ST/K
      AMAT(KCT)=LAM/ST
      BMAT(KCT)=PP20
      CMAT(KCT)=Q1-BMAT(KCT)
      QMAT(KCT)=W-DMAT(KCT)
      PMAT(KCT,2)=TPRINT+KCT*STORE
      GMAT(KCT)=TIME
    1 IF (TIME .LT. XFIN .AND. KCT .LT. 16) GO TO 2
      XSTRT=TIME+.01D0
      RETURN
      END
```

APPENDIX 3

A PRELIMINARY INVESTIGATION OF TRANSIENT BEHAVIOR OF k-SERVER AND FINITE-CAPACITY QUEUEING SYSTEMS

Chapters 3 and 4 constitute a thorough investigation of the transient behavior of the expected queue length, Q(t), for two particular classes of infinite-capacity, single-queue, single-server queueing systems. The dominating feature of the empirical results is that, in every case, for large t the transient decay is approximately exponential with a time constant that depends on ρ, μ, and the coefficients of variation of the interarrival and service time distributions.

In this appendix, we present a preliminary investigation of the effects of two system characteristics on this behavior. Specifically, for Markovian systems and for partially deterministic systems in which the embedded chain is a first-order Markov process, we examine the manner in which the functional form of the transient decay and the time required for the transient response to become negligible depend on the number of servers and on system capacity. Our intent here is only to provide groundwork for further research of these systems by describing some of the results that we have obtained to date.

A3.1 The Decay of Transients in M/M/k Queueing Systems

In this section, we examine the transient behavior of a collection of infinite-capacity M/M/k queueing systems to determine the dependence of the transient response on the number of servers in the system. We first consider systems that begin at rest. Then we vary the initial conditions to determine their effect on the transient decay.

Figure A3.1 illustrates $\log|Q(\infty) - Q(t)|$ for M/M/k systems with $\rho=.75$, $k\mu=1$, and $k=2,5$, and 9. All systems are initially idle. These

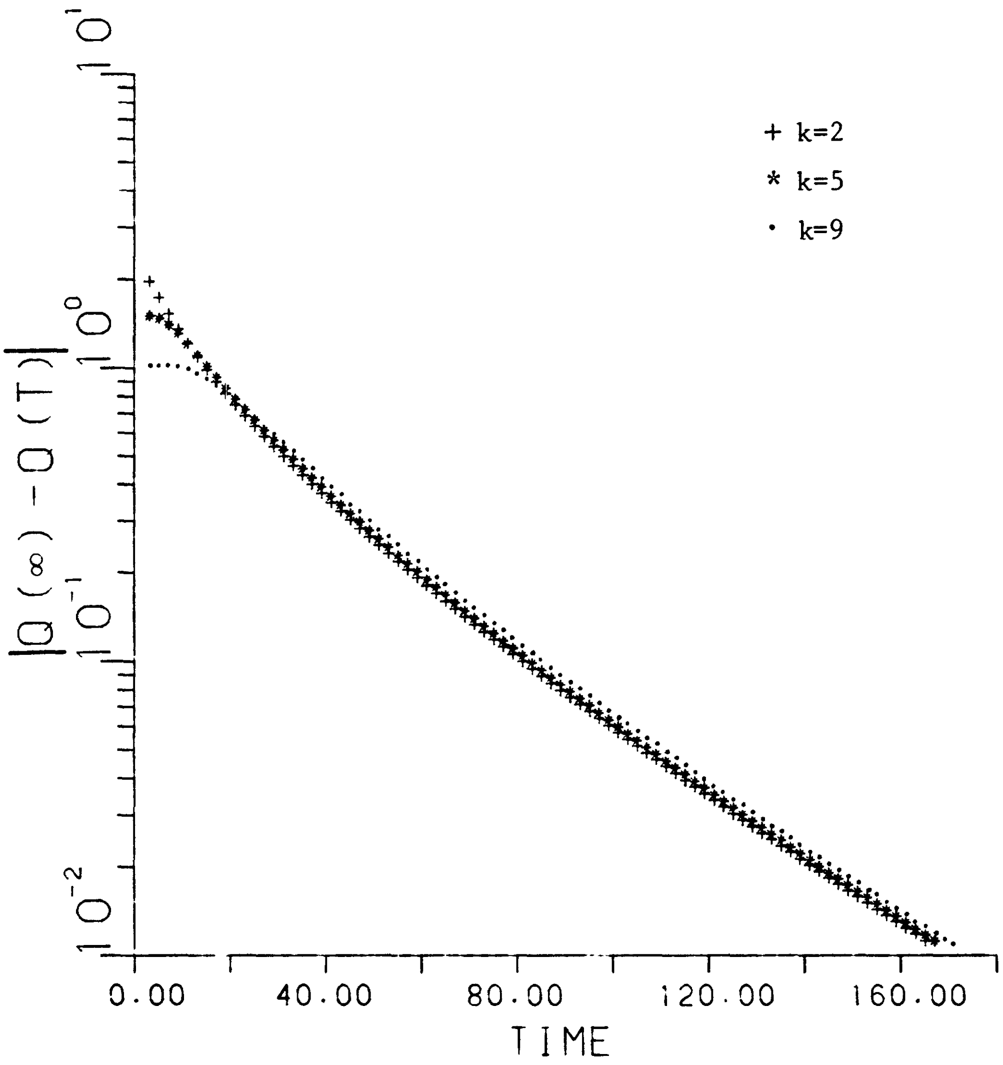


Figure A3.1: **Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus** Time for an M/M/k System With $\rho=.75$, $k\mu=1$, and $P_0(0)=1$; k=2, 5, and 9

cases were selected so that, independent of k, when all servers are busy customers enter the system according to a Poisson process with rate $\lambda=.75$ and are served according to a Poisson process with combined rate $k\mu=1$. Thus, with respect to expected value measures, as long as the service facility is fully utilized the systems are equivalent. Due to differences when one or more servers are idle, however, if λ and $k\mu$ remain fixed, $Q\ (\infty)$ is a decreasing function of k. This accounts for the differences in $\log|Q(\infty) - Q(0)|$ at time $t=0$.

Several features of the curves in Figure A3.1 merit comment. As in the single-server case, all curves become linear for large t. Thus, except for an initial period, transients decay in an approximately exponential manner. For large t, the slope of $\log|Q(\infty) - Q(t)|$ versus t appears to be independent of the number of servers in the system. This suggests that the time constant is a function of $k\mu$ but not of k alone.

Finally, note that for small t, $\log|Q(\infty) - Q(t)|$ remains constant for an amount of time that increases with k. This can be explained intuitively as follows. The expected queue length is clearly equal to zero as long as one or more servers are idle. Therefore, as the systems illustrated in Figure A3.1 begin in the empty state, $\log|Q(\infty) - Q(t)|$ will remain constant (equal to $\log Q(\infty)$)until a customer arrives to find all servers busy. As customers arrive at the same rate for all k and the service facility cannot possibly be saturated before the kth customer arrives, as k increases the service facility will require a proportionately longer time to fill. Thus, as illustrated in the figure, $\log|Q(\infty) - Q(t)|$ will remain constant for an initial time period of a length which is an increasing function of k.

This behavior for small t complicates estimation of the time to equilibrium. To avoid this problem, we use instead L(t), the number of customers in the system (in queue and in service) at time t as our representative measure of system behavior. Figure A3.2 illustrates $\log|L(\infty)-L(t)|$ versus t for the same three M/M/k systems with $\rho = .75$, $\mu = 1$, and k=2,5, and 9. As before, the systems begin at rest. (Note that $L(\infty)$ is an increasing function of k when λ and $k\mu$ remain fixed [21].) In each case, for large t, L(t) approaches $L(\infty)$ in an approximately exponential manner. The convexity of the early portion of the curve implies that initial transient decay occurs at a rate faster than this exponential function. Thus, it appears that for systems which begin at rest, L(t) can be approximated by

$$L(t) \cong L(\infty) + Ae^{-t/\tau}, \quad t \geq \hat{t}, \qquad (A3.1)$$

for some parameters $A < 0$ and $\hat{t}$, and is bounded below by

$$L(t) = L(\infty)\ [1 - e^{-t/\tau}], \quad t \geq 0, \qquad (A3.2)$$

for some parameter τ.

Closer examination of the slope of the $\log|L(\infty) - L(t)|$ versus t curves for large t suggests that our τ_R formula (3.24) with μ replaced by $k\mu$ (we refer to this modified τ_R expression as τ_R^*) provides a good approximation to this time constant. In Table A3.1 we show the experimental time constant τ_{exp}, τ_R^*, and the ratio τ_{exp}/τ_R^* for several representative M/M/k systems. In most cases, τ_R^* is within 10% of τ_{exp}. These results suggest that a good upper bound on the time to equilibrium is given by pure exponential decay, i.e., expression A3.2, with parameter τ_R^*.

An implication of these results is that for an M/M/k system which begins at rest, the time required for the transient effects to become

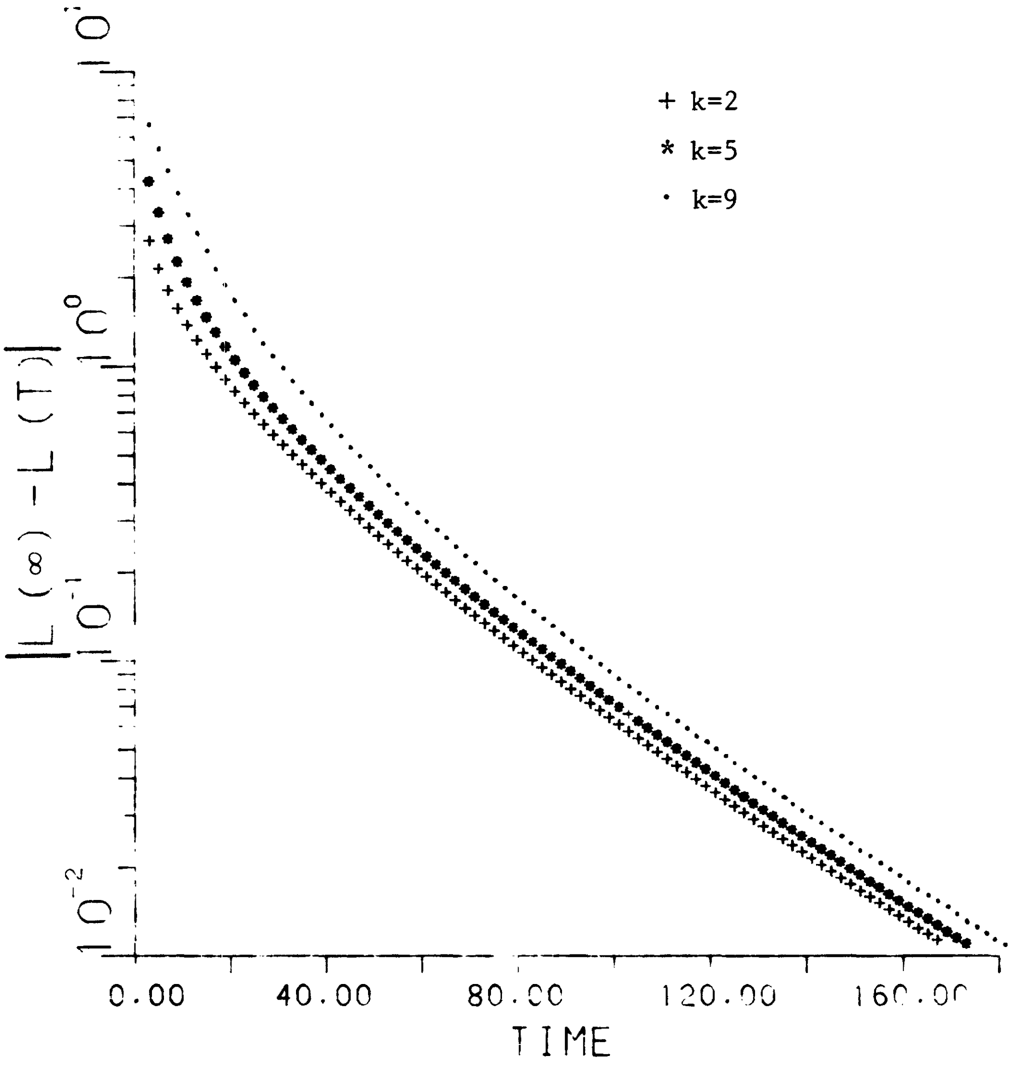


Figure A3.2: Semilogarithmic Plots of $|L(\infty) - L(T)|$ Versus Time for an M/M/k System With $\rho=.75$, $k\mu=1$, and $P_0(0)=1$; k=2, 5, and 9

Table A3.1: A Comparison of Estimated and Observed Time Constants for Two M/M/k Systems Which Begin at Rest

k	ρ	$k\mu$	τ_{exp}	τ_R	τ_{exp}/τ_R
2	.5	1	8.3	8.6	1.0
3			8.2	8.6	1.0
5			8.1	8.6	.9
7			8.6	8.6	1.0
9			9.9	8.6	1.2
2	.75	1	41	41	1.0
3			40	41	1.0
5			39	41	1.0
7			39	41	1.0
9			39	41	1.0

negligible is a function of λ and $k\mu$ but is not otherwise a function of the number of servers. Intuitively, we might expect this behavior to hold for other initial conditions, as well. The results that follow suggest that this is, in fact, the case.

Figures A3.3 - A3.4 illustrate $\log|L(\infty) - L(t)|$ versus time for an M/M/9 system with $\rho=.75$, $\mu=1/9$, and initial conditions of $i=2,6,7,8,11,20$, and 25 customers in the system. In all cases, for large t transients decay in an approximately exponential manner.

By examining graphs of L(t) versus t for these seven cases, we confirm that, for this system, the initial behavior of L(t) falls into the four groups indicated in Chapter 4. Figure A3.5 illustrates L(t) for small t when there are initially 2, 6,7, or 8 customers in the system. When i is less than 7, L(t) is a monotonic increasing function of time. When i equals 7 or 8, L(t) decreases for small t, and then increases monotonically to $L(\infty)$. When $i=8$, this initial decrease in L(t) overshoots the equilibrium value, $L(\infty)$.

These cases illustrate behavior in the first two categories indicated in Chapter 4. Intuitive justification of this behavior is similar to that in the single-server case. If, at time t=0, the first arrival is likely to occur before the first service completion, L(t) will increase in a monotonic manner. For an M/M/k system with i customers at time t=0, if $i \leq k$,

$$P\begin{pmatrix}\text{first customer arrival occurs} \\ \text{before first service completion}\end{pmatrix} = \frac{\lambda}{i\mu+\lambda} . \qquad \text{(A3.3)}$$

Therefore, L(t) will increase in a monotonic manner if

$$\frac{\lambda}{i\mu+\lambda} \geq .5 , \qquad \text{(A3.4)}$$

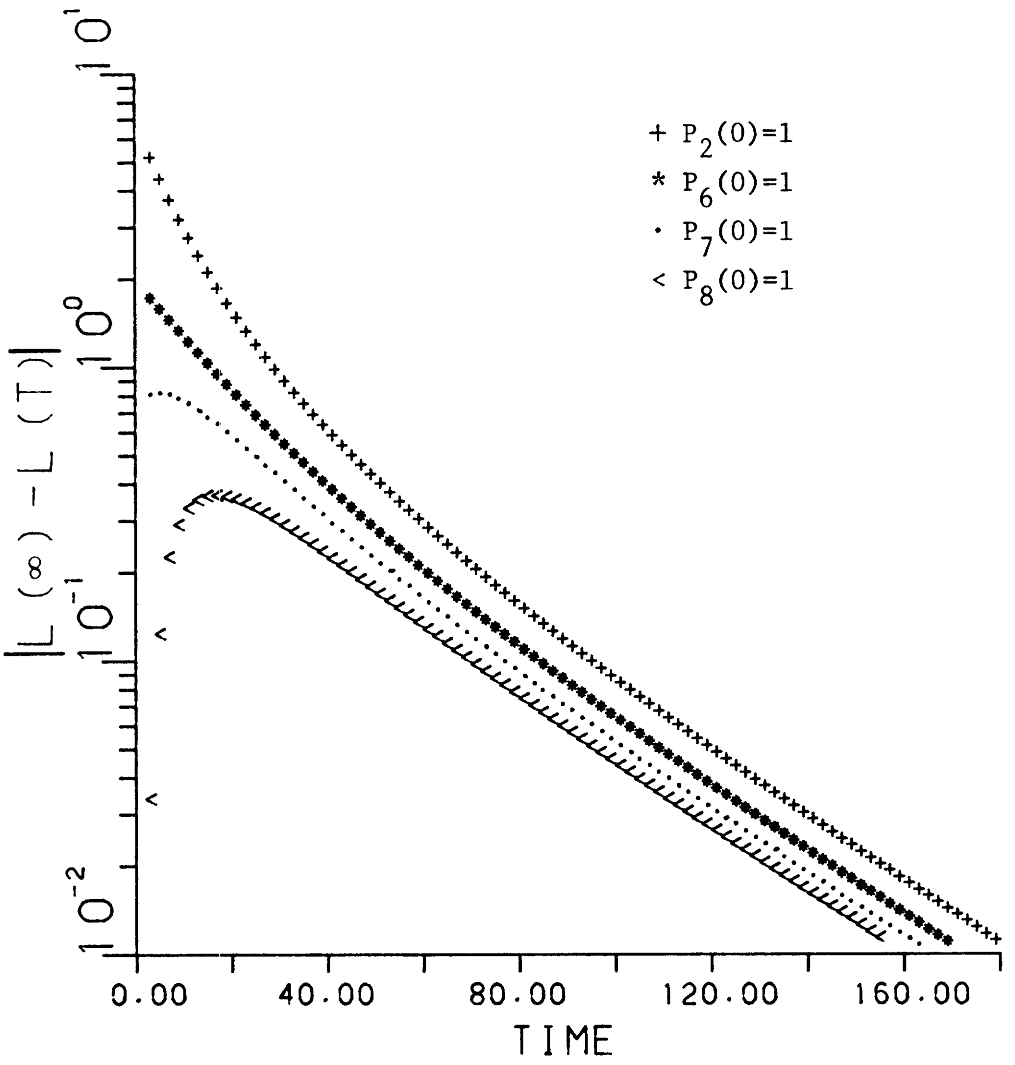


Figure A3.3: Semilogarithmic Plots of $|L(\infty) - L(T)|$ Versus Time for an M/M/9 Sytem With $\rho=.75$ and $\mu=1/9$; $P_2(0)=1$, $P_6(0)=1$, $P_7(0)=1$, and $P_8(0)=1$

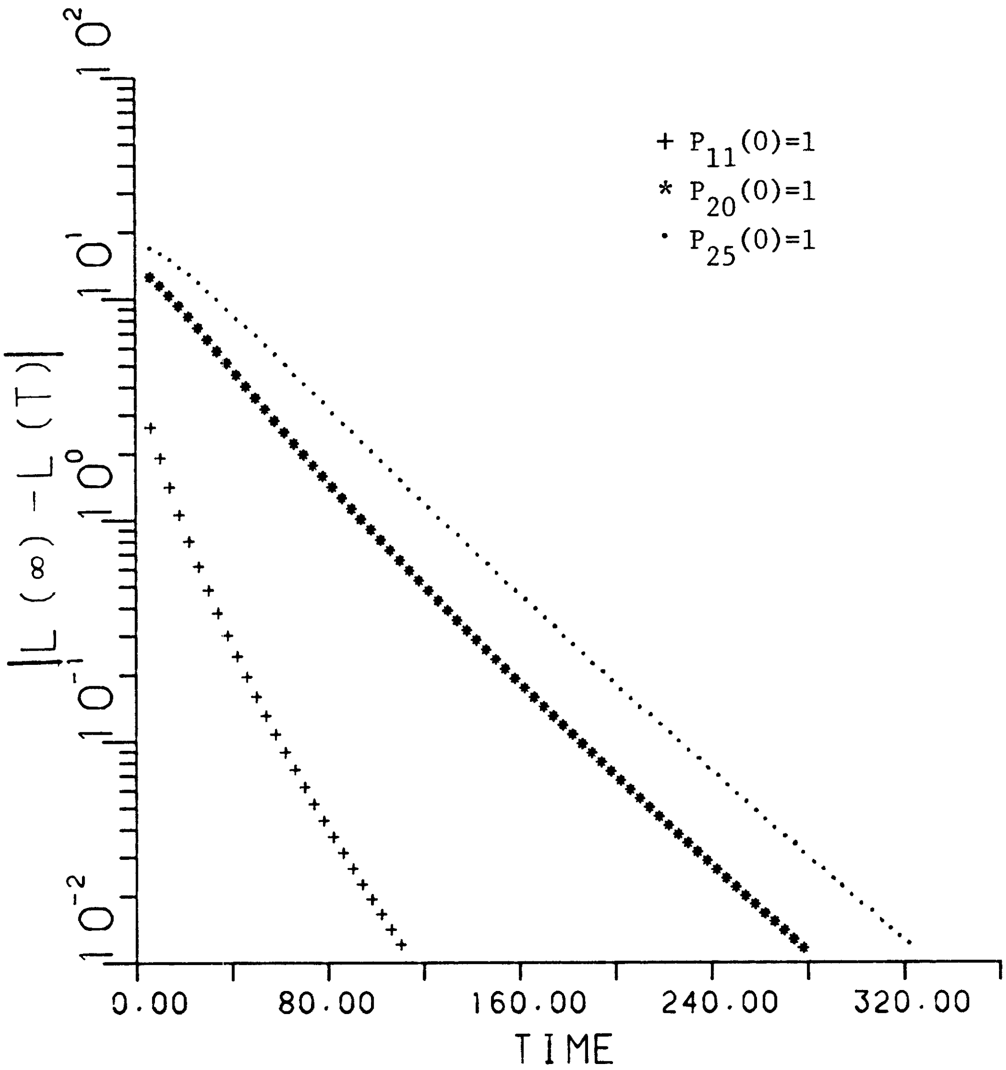


Figure A3.4: Semilogarithmic Plots of $|L(\infty) - L(T)|$ Versus Time for an M/M/9 System With $\rho=.75$ and $\mu=1/9$; $P_{11}(0)=1$, $P_{20}(0)=1$, and $P_{25}(0)=1$

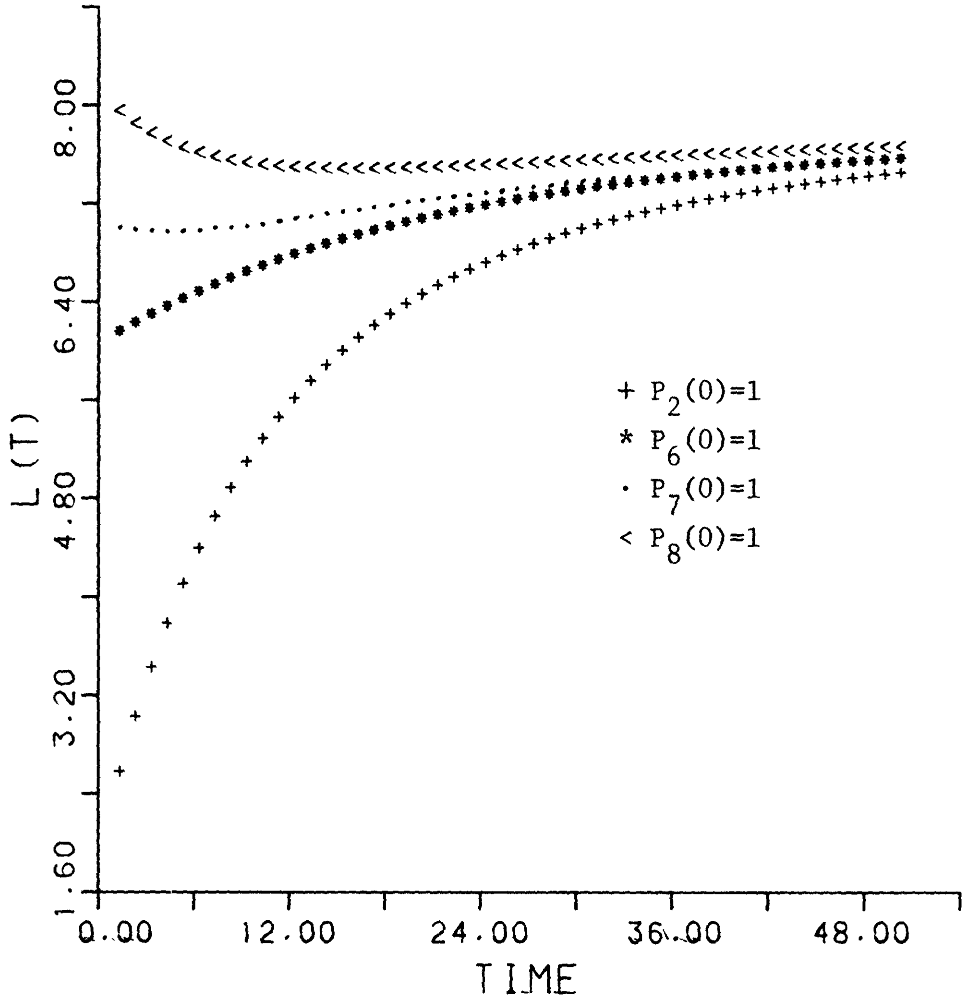


Figure A3.5: The Expected Number in the System Versus Time for an M/M/9 System With $\rho=.75$ and $\mu=1/9$; $P_2(0)=1$, $P_6(0)=1$, $P_7(0)=1$, and $P_8(0)=1$

or, equivalently, if

$$i \leq \frac{\lambda}{\mu} . \qquad \text{(A3.5)}$$

For the system illustrated in Figure A3.5, $\lambda=.75$ and $\mu=1/9$. Thus, we expect $L(t)$ to increase in a monotonic manner for all t only if $i \leq 6.75$. This is, in fact, the behavior indicated in Figure A3.5.

In Figure A3.6, we illustrate initial behavior of $L(t)$ when there are 11,20, or 25 customers in the system at time $t=0$. In all cases, $L(t)$ decreases monotonically to $L(\infty)$, but when $i=20$ or 25, $L(t)$ is initially a linear function of t. This is the behavior observed in Chapter 4 for systems with initial conditions in groups (iii) and (iv). The intuitive justification presented at that point carries over directly to k-server systems.

In Table A3.2, we compare τ_{exp} to τ_R^* for each of these cases. As before, τ_R^* appears to provide a good approximation for the actual time constant τ.

These results suggest that, as in the single-server case we can categorize transient behavior into four classes with respect to initial conditions. Let i be equal to the number of customers in the system at time $t=0$. Then:

(i) If $i \leq \frac{\lambda}{\mu}$, $L(t)$ approaches $L(\infty)$ monotonically from below. In addition, $L(t)$ is bounded from below by (A4.1) with a negative value of A and $\hat{t}=0$.

(ii) If $\frac{\lambda}{\mu} < i \leq \tilde{L}_1$, for some $\tilde{L}_1 > L(\infty)$, $L(t)$ will not be a monotonic function of t, but will initially decrease before increasing to $L(\infty)$ in a monotonic manner. If $L(\infty) \leq i \leq \tilde{L}_1$, this initial decrease will cause $L(t)$ to overshoot $L(\infty)$ exactly once. The time to equilibrium will be bounded above by that of a system that begins at rest.

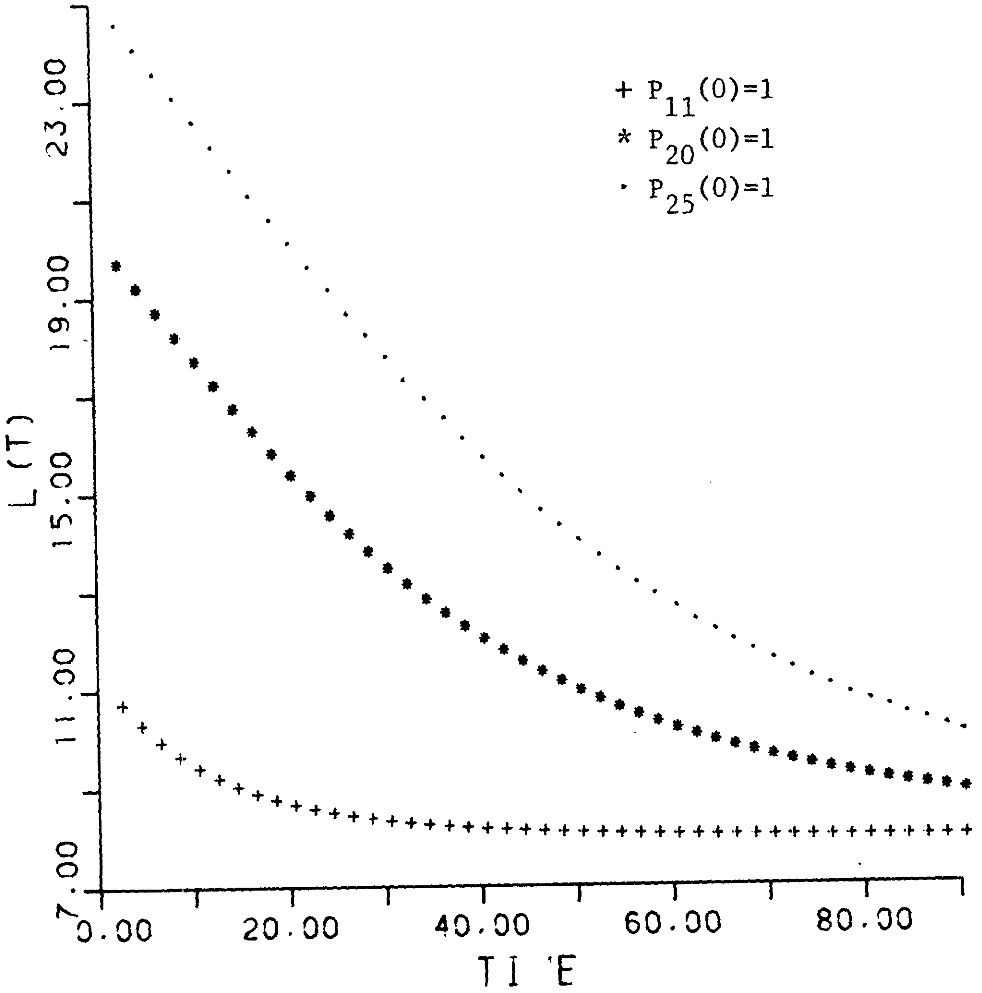


Figure A3.6: The Expected Number in the System Versus Time for an M/M/9 System With $\rho=.75$ and $\mu=1/9$; $P_{11}(0)=1$, $P_{20}(0)=1$, and $P_{25}(0)=1$

Table A3.2: A Comparison of Estimated and Observed Time Constants for an M/M/9 System Which Does Not Begin at Rest

k	ρ	$k\mu$	Initial Conditions	τ_{exp}	τ_R	τ_{exp}/τ_R
9	.75	1	$P_2(0) = 1$	37	41	.9
			$P_6(0) = 1$	41	41	1.0
			$P_7(0) = 1$	41	41	1.0
			$P_8(0) = 1$	41	41	1.0
			$P_{20}(0) = 1$	41	41	1.0
			$P_{25}(0) = 1$	45	41	1.1

(iii) If $\tilde{L}_1 < L(0) \leq \tilde{L}_2$, for some $\tilde{L}_2 > \tilde{L}_1 > L(\infty)$, L(t) will decrease monotonically to L(∞), bounded above by expression (A4.1) with a positive value of A.

(iv) If $\tilde{L}_2 < L(0)$, initially L(t) will decrease linearly, then L(t) will behave as in category (iii).

A3.2 The Effect of System Capacity on the Decay of Transients

To this point, we have examined only infinite-capacity queueing systems. Intuitively, since there are fewer states in the system, we might expect that when capacity is finite, equilibrium will be reached more quickly than in the corresponding infinite-capacity system. The following empirical results support this conjecture and indicate that, as with systems considered earlier, for large t transient decay is approximately exponential. The discussion here will be restricted to systems which begin at rest.

We consider first an M/M/1 system with $\rho=.75$, $\mu=1$, and a maximum of N customers allowed in the system at any time. Figures A3.7 and A3.8 illustrate $\log|Q(\infty) - Q(t)|$ versus t for ten values of N in the range from 2 to 20. In each case, after an initial period decay is approximately exponential. The time constant (equivalently, the magnitude of the reciprocal of the slope) of $\log|Q(\infty) - Q(t)|$ can be seen to increase with N confirming our intuitive feeling that the time to equilibrium should, in fact, be an increasing function of system capacity. In Table A3.3, we list the experimental time constants for these and other representative M/M/1 systems with varying system capacity.

It should be noted that it has been proven that, at least asymptotically, transients in these systems decay in an exponential manner. This is due to the theoretical result of Morse (see expression (3.2)) which

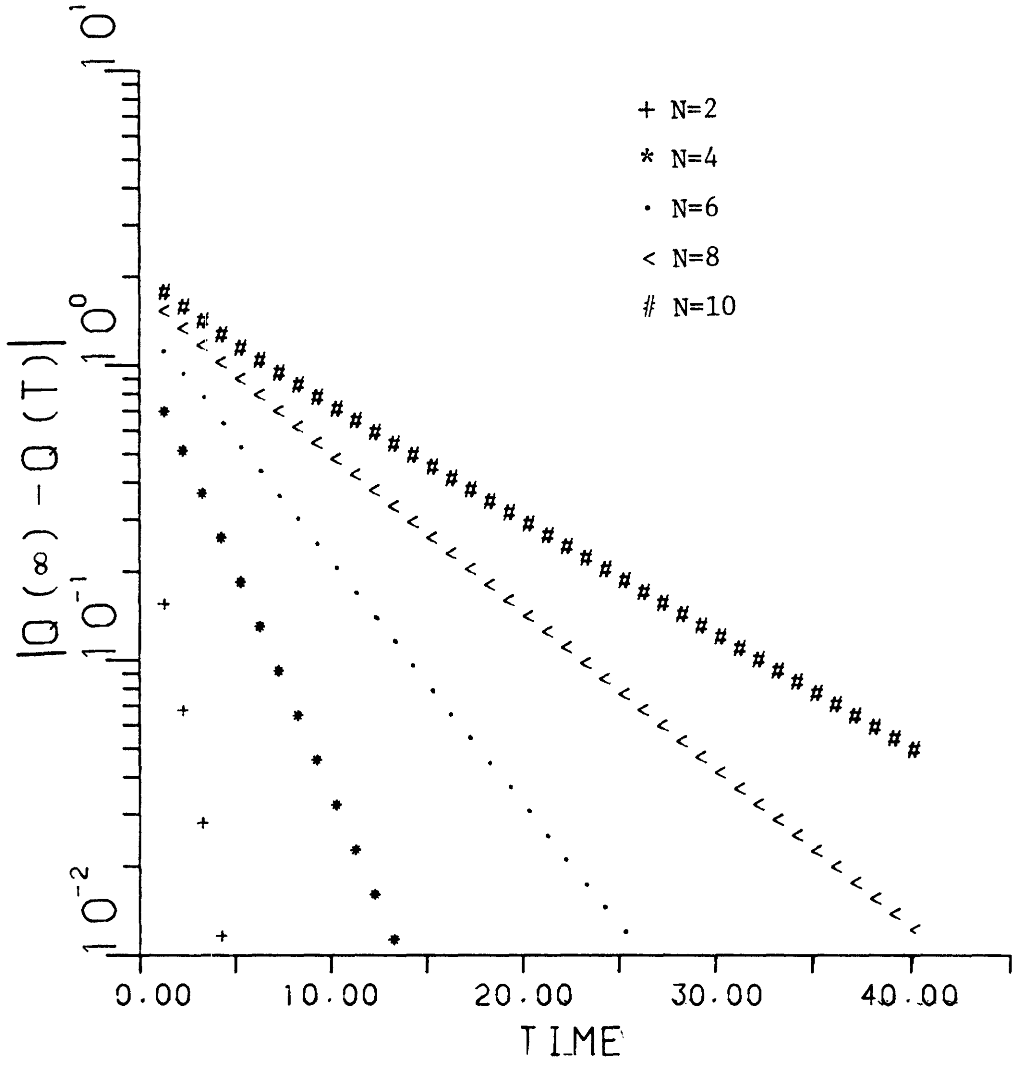


Figure A3.7: **Semilogarithmic Plots of** $|Q(\infty) - Q(T)|$ Versus Time for an M/M/1 System With $\rho=.75$, $\mu=1$, $P_0(0)=1$, and a Finite Capacity of N Customers; N=2, 4, 6, 8, and 10

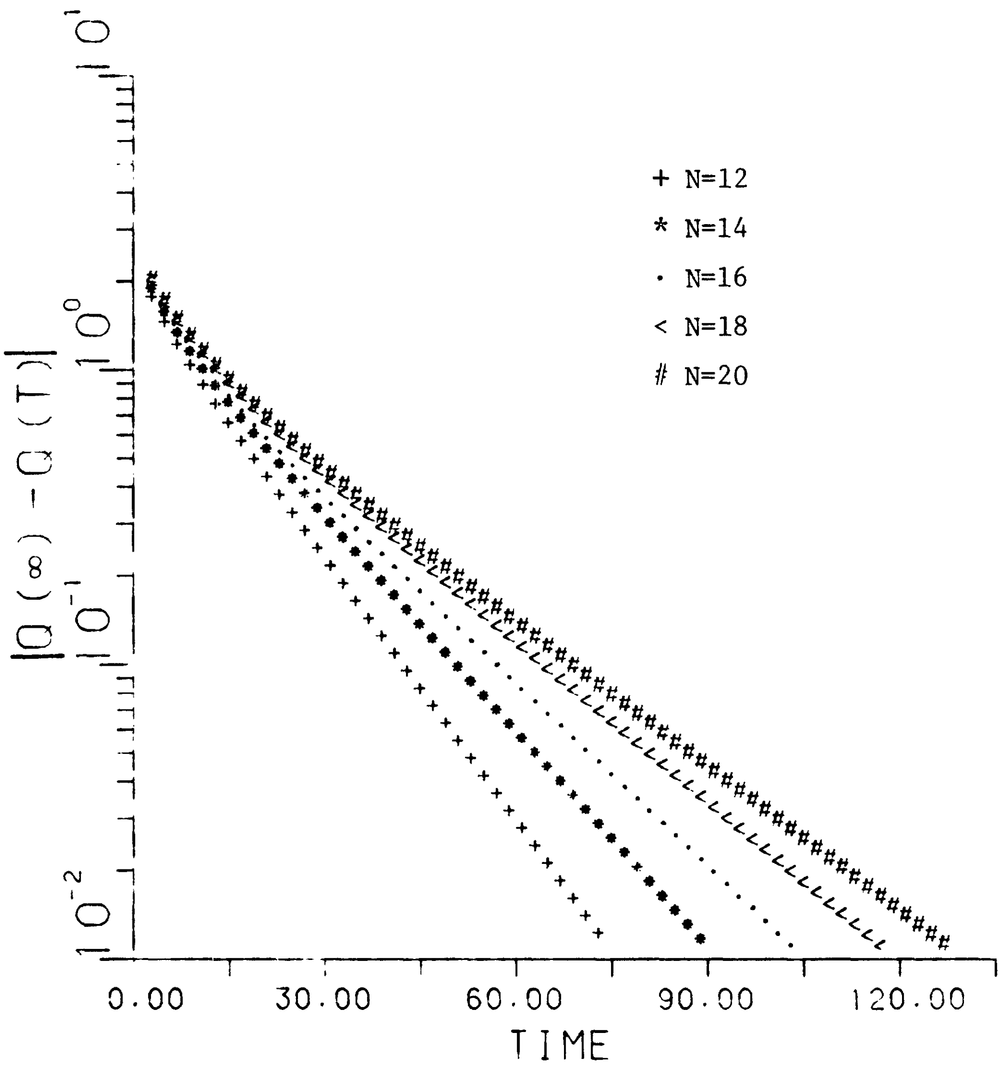


Figure A3.8: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an M/M/1 System With $\rho=.75$, $\mu=1$, $P_0(0)=1$, and a Finite Capacity of N Customers; N=12, 14, 16 18, and 20

Table A3.3: The Experimental Time Constant for an M/M/l System With ρ=.75, μ=1, and a Finite Capacity of N Customers

N	ρ=.5	ρ=.75	ρ=.85
2	1.3	1.1	1.1
4	2.8	2.8	2.8
6	4.3	5.3	5.3
8	5.8	8.0	8.5
10	7.2	10.8	12
12		15	17
14		18	22
16		21	26
18		24	32
20		25	37
∞	8.6	41	122

indicates that for a finite-capacity M/M/1 queueing system, Q(t) can be expressed as a weighted sum of decaying exponential terms. We do not, however, have prior knowledge on the amount of time necessary for one of these exponential terms to dominate. Our empirical results suggest that the length of the initial nonexponential period is an increasing function of N.

In Figures A3.9 and A3.10, we show that these results also hold for two Erlangian queueing systems, specifically $M/E_4/1$ and $E_4/M/1$ systems with $\rho=.75$ and $\mu=1$. In Table A3.4 the experimental time constants for these systems are listed.

These preliminary results indicate that the time to equilibrium of a finite-capacity queueing system is bounded above by that of the corresponding infinite-capacity system. To verify that this is, in fact, true, a more exhaustive study of systems with various interarrival and service time distributions under a range of initial conditions is needed. We suspect that with further work, a modified τ_R formula could be determined to account for system capacity.

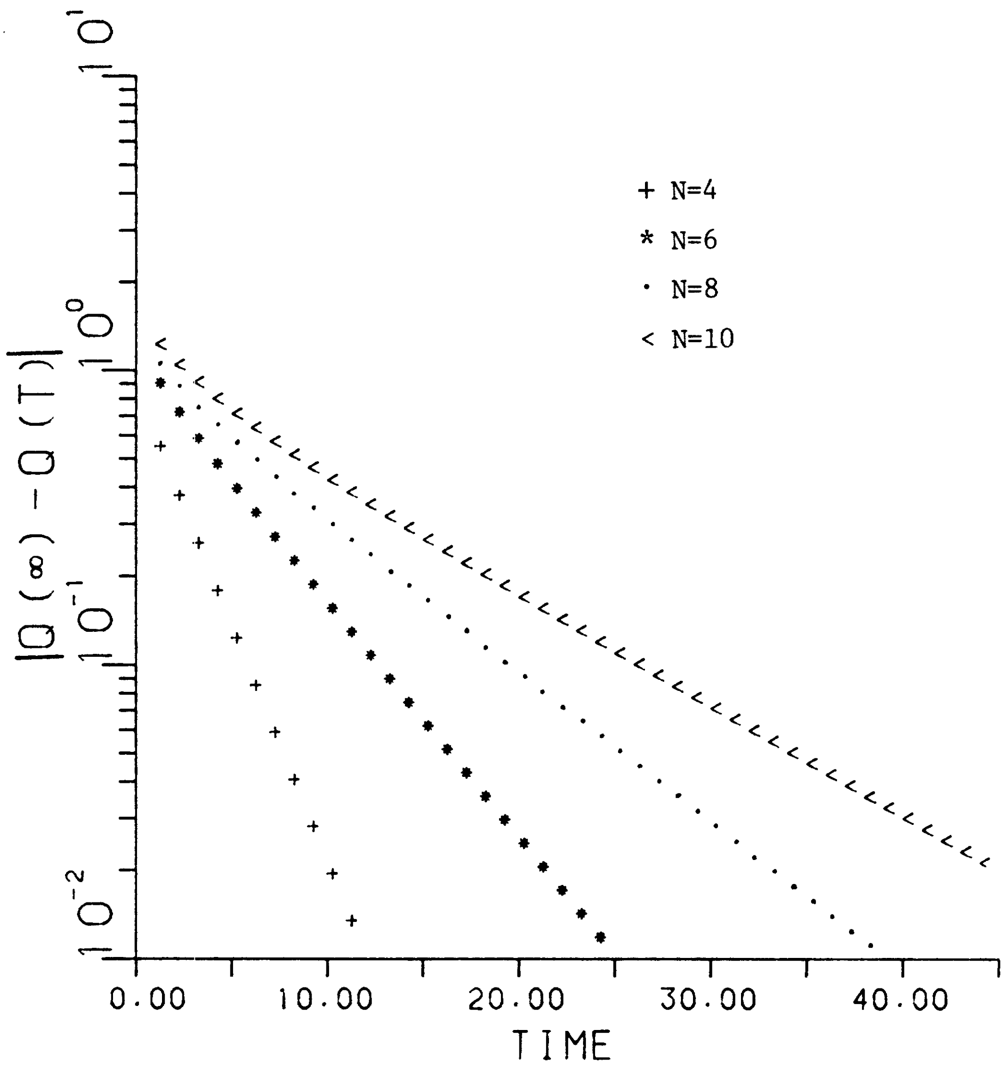


Figure A3.9: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an $M/E_4/1$ System With $\rho=.75$, $\mu=1$, $P_0(0)=1$, and a Finite Capacity of N Customers; N=4, 6, 8, and 10

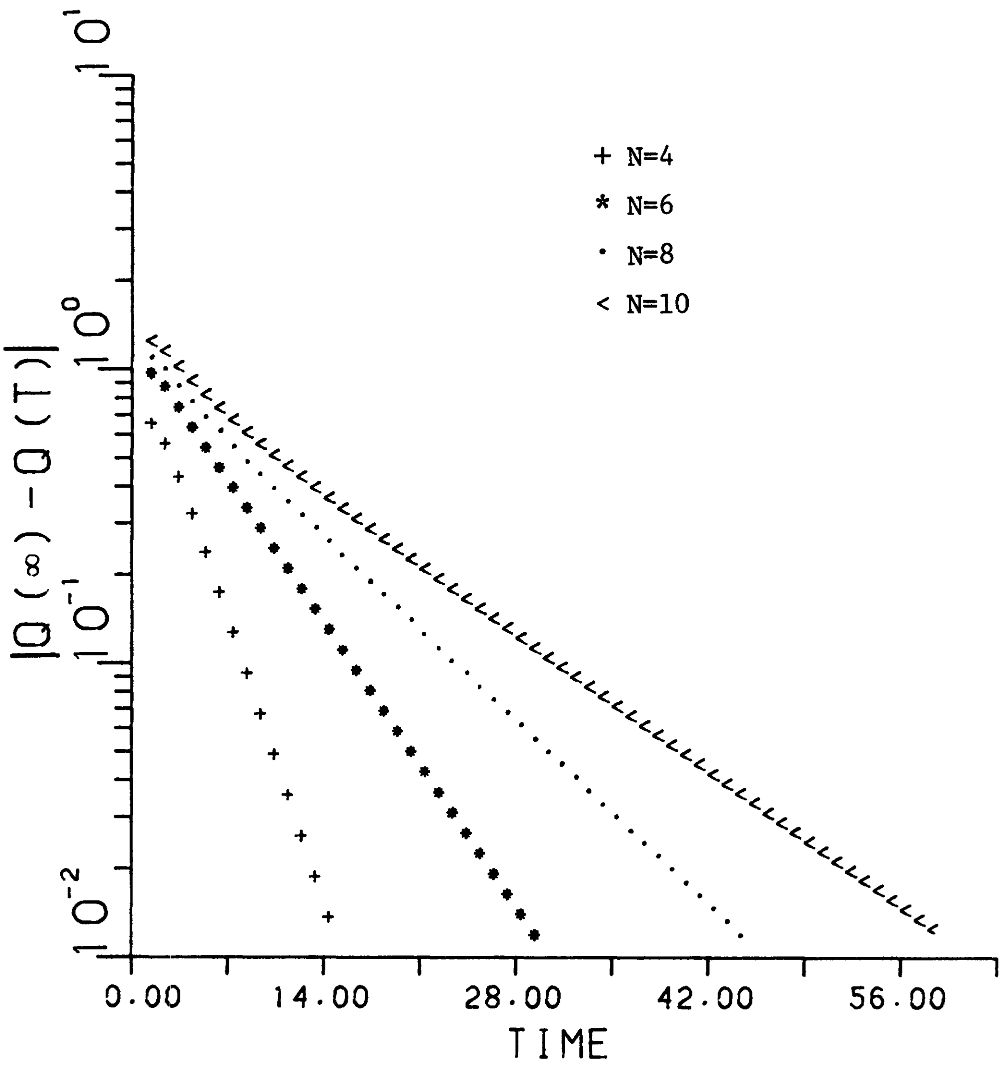


Figure A3.10: Semilogarithmic Plots of $|Q(\infty) - Q(T)|$ Versus Time for an $E_4/M/1$ System With $\rho=.75$, $\mu=1$, $P_0(0)=1$, and a Finite Capacity of N Customers; N=4, 6, 8, and 10

Table A3.4: The Experimental Time Constant for $M/E_4/1$ and $E_4/M/1$ Systems With ρ=.75, μ=1, and a Finite Capacity of N Customers

$M/E_4/1$		$E_4/M/1$	
N	τ_{exp}	N	τ_{exp}
2	.7	2	.8
4	2.6	4	3.2
6	5.5	6	6.0
8		8	9.9
10	11.9	10	13.4
∞	26	∞	26

REFERENCES

1. Barzily, Z. and Gross, D. "Transient Solutions for Repairable Item Provisioning." Washington, D.C.: The George Washington University, Technical Report, April 1979.

2. Bhat, U. N. "Sixty Years of Queueing Theory." Management Science, 15 (1969): B-280-B-294.

3. Bhat, U. N. "The Value of Queueing Theory-A Rejoinder." Interfaces, 8 (1978): 27-28.

4. Byrd, J., Jr. "The Value of Queueing Theory." Interfaces, 8 (1978): 22-26.

5. Chang, S. S. L. "Simulation of Transient and Time Varying Conditions in Queueing Networks," Proceedings of the Seventh Annual Pittsburgh Conference on Modeling and Simulation, University of Pittsburgh (1977): 1075-1078.

6. Daley, D. J. "Monte Carlo Estimation of the Mean Queue Size in a Stationary GI/M/1 Queue." Operations Research, 16 (1968): 1002-1005.

7. Drake, A. W. Fundamentals of Applied Probability Theory. New York: McGraw-Hill, 1967.

8. Gaver, D. P., Jr. "Diffusion Approximations and Models for Certain Congestion Problems." Journal of Applied Probability, 5 (1968): 607-623.

9. Grassman, W. K. and Servranckx, J. "The Que Package." Mimeographed. Saskatoon, Canada: University of Saskatchewan, 1979.

10. Grassmann, W. K. "Transient Solutions in Markovian Queueing Systems." Computers and Operations Research, 4 (1977): 47-53.

11. Grassmann, W. K. "Transient and Steady State Results for Two Parallel Queues." OMEGA, 8 (1980): 105-112.

12. Gross, D. and Harris, C. M. Fundamentals of Queueing Theory. New York: John Wiley and Sons, 1974.

13. Hengsbach, G. and Odoni, A. R. "Time Dependent Estimates of Delays and Delay Costs at Major Airports." Cambridge, Mass.: MIT, Flight Transportation Laboratory, Technical Report, January 1975.

14. Keilson, J. Markov Chain Models-Rarity and Exponentiality. New York: Springer-Verlag, 1979.

15. Kingman, J. F. C. "The Heavy Traffic Approximation in the Theory of Queues." Proceedings of the Symposium on Congestion Theory, University of North Carolina Press (1965): 137-169.

16. Kingman, J. F. C. "On Queues in Heavy Traffic." Journal of the Royal Statistical Society, 24 (1962): 383-392.

17. Kivestu, P. A. "Alternate Methods of Investigating the Time Dependent M/G/k Queue." S.M. thesis, Massachusetts Institute of Technology, 1976.

18. Kivestu, P. A. Private Communication.

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

20. Kleijnen, J. P. C. Statistical Techniques in Simulation Part 2. New York: Marcel Dekker, 1975.

21. Kleinrock, L. Queueing Systems. Vol. 1: Theory. New York: John Wiley and Sons, 1975.

22. Kleinrock, L. Queueing Systems. Vol. 2: Computer Applications. New York: John Wiley and Sons, 1976.

23. Kobayashi, H. "Application of the Diffusion Approximation to Queueing Networks II: Non-equilibrium Distributions and Application to Computer Modeling." Journal of the Association for Computing Machinery, 21 (1974): 459-469.

24. Kolesar, P. "A Quick and Dirty Response to the Quick and Dirty Crowd; Particularly to Jack Byrd's 'The Value of Queueing Theory'." Interfaces, 9 (1979): 77-82.

25. Koopman, B. O. "Air-Terminal Queues Under Time-Dependent Conditions." Operations Research, 20 (1972): 1089-1114.

26. Kotiah, T. C. T. "Approximate Transient Analysis of Some Queueing Systems." Operations Research, 26 (1978): 333-346.

27. Law, A. M. and Carson, J. S. "A Sequential Procedure for Determining the Length of a Steady-State Simulation." Operations Research, 27 (1979): 1011-1025.

28. Leese, E. L. and Boyd, D. W. "Numerical Methods of Determining the Transient Behavior of Queues with Variable Arrival Rates." Journal of the Canadian Operations Research Society, 4 (1966): 1-13.

29. Marks, N. B. "A Study of Times to Reach Equilibrium in an M/M/S Queueing System." Mimeographed. Coral Gables, Florida: University of Miami, 1979.

30. Moore, S. C. "Approximating the Behavior of Nonstationary Single-Server Queues." Operations Research, 23 (1975): 1011-1032.

31. Morse, P. M. Queues, Inventories and Maintenance. New York: John Wiley and Sons, 1958.

32. Neuts, M. F. "The Single Server Queue in Discrete Time-Numerical Analysis I." Naval Research Logistics Quarterly, 20 (1973): 297-304.

33. Newell, G. F. Applications of Queueing Theory. London: Chapman and Hall, 1971.

34. Newell, G. F. "Queues With Time-Dependent Arrival Rates I - The Transition Through Saturation." Journal of Applied Probability, 5 (1968): 436-451.

35. Rider, K. L. "A Simple Approximation to the Average Queue Size in the Time-Dependent M/M/1 Queue." Journal of the Association for Computing Machinery, 23 (1976): 361-367.

36. Roth, E. "An Advanced Time-Dependent Queueing Model for Airport Delay Analysis." Cambridge, Mass.: MIT, Flight Transportation Laboratory, Technical Report, October 1979.

37. Rothkopf, M. H. and Oren, S. S. "A Closure Approximation for the Non-stationary M/M/S Queue." Management Science, 25 (1979): 522-534.

38. Saaty, T. L. "Seven More Years of Queues: A Lament and a Bibliography." Naval Research Logistics Quarterly, 13 (1966): 447-476.

39. U.S., Department of Transportation, Federal Aviation Administration, "Techniques for Determining Airport Airside Capacity and Delay," No. FAA-RD-74-124, June 1976.

40. Vazsonyi, A. "To Queue or Not to Queue: A Rejoinder." Interfaces, 9 (1979): 83-86.

41. Wang, K.-K. G. "Continuous Simulation of Queueing Networks with Transient and Time Varying Conditions." Ph.D. dissertation, State University of New York at Stonybrook, forthcoming.

42. Weiss, W. E. "A Flexible Model for Runway Capacity Analysis." S. M. thesis, Massachusetts Institute of Technology, 1980.

43. Wilson, J. R. and Pritsker, A. A. B. "Evaluation of Startup Policies in Simulation Experiments." Simulation, 31 (1978): 79-89.

44. Wilson, J. R. and Pritsker, A. A. B. "A Survey of Research on the Simulation Startup Problem." Simulation, 31 (1978): 55-58.