COMPUTER SCIENCE
TECHNICAL REPORT SERIES

UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND
20742

IV 12

Technical Report TR-955
AFOSR-78-3654

October 1980

An Approximate Transient Analysis of the M(t)/M/1 Queue

Richard A. Upton
Satish K. Tripathi

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 81-0485	2. GOVT ACCESSION NO. AD-100 179	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) AN APPROXIMATE TRANSIENT ANALYSIS OF THE M(t)/M/1 QUEUE		5. TYPE OF REPORT & PERIOD COVERED INTERIM 6. PERFORMING ORG. REPORT NUMBER TR-955
7. AUTHOR(s) Richard A. Upton and Satish K. Tripathi		8. CONTRACT OR GRANT NUMBER(s) AFOSR-78-3654
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Computer Science University of Maryland College Park MD 20740		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS PE61102F 2304/A2
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB DC 20332		12. REPORT DATE OCTOBER 1980 13. NUMBER OF PAGES 50
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

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

18. SUPPLEMENTARY NOTES

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

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

An approach, based on recent work by Stern (STER 1979), is described for obtaining the approximate transient behavior of both the M/M/1 and the M(t)/M/1 queues, where the notation M(t) indicates an exponential arrival process with time-varying parameter λ(t). The basic technique employs an M/M/1/K approximation to the M/M/1 queue to obtain a spectral representation of the time-dependent behavior for which the eigenvalues and eigenvectors are real.

Lanba

(CONT.)

ITEM #20, CONT.

Following a general survey of transient analysis which has already been accomplished, Stern's M/M/1/K approximation technique is examined to determine how best to select a value for K which will yield both accurate and computationally efficient results. It is then shown how the approximation technique can be extended to analyze the M(t)/M/1 queue where we assume that the M(t) arrival process can be approximated by a discretely time-varying Poisson process.

An approximate expression for the departure process of the M/M/1 queue is also proposed which implies that for an M(t)/M/1 queue whose arrival process is discretely time-varying, so too the departure process can be approximated as discretely time-varying (albeit with a different time-varying parameter).

In all cases, the accuracy and validity of these results are examined by comparison with exact analytic results, simulation or alternative discrete-time approaches (e.g., the embedded Markov chain technique of Moore [MOOR 1972]).

AN APPROXIMATE TRANSIENT ANALYSIS OF THE M(t)/M/1 QUEUE†

by

Richard A. Upton* and Satish K. Tripathi

Dept. of Computer Science
University of Maryland
College Park, Maryland 20740

*Ric Upton is currently working for The Analytic Sciences Corporation, Six Jacob Way, Reading, MA 01867

†This research was supported in part by the Air Force Office of Scientific Research under grant AFOSR-78-3654[illegible].

ACKNOWLEDGEMENT

The authors would like to thank the members of the Technical Publications Department of TASC for their responsive preparation of this paper, and the Computer Science Center of the Dept. of Computer Science (University of Maryland) for the computer resources that were made available for this analysis. Comments from the members of the System Design and Analysis Group (SDAG), Dept. of Computer Science, University of Maryland were very helpful through the course of this work.

ABSTRACT

An approach, based on recent work by Stern [STER 1979], is described for obtaining the approximate transient behavior of both the M/M/1 and M(t)/M/1 queues, where the notation M(t) indicates an exponential arrival process with time-varying parameter $\lambda(t)$. The basic technique employs an M/M/1/K approximation to the M/M/1 queue to obtain a spectral representation of the time-dependent behavior for which the eigenvalues and eigenvectors are real.

Following a general survey of transient analysis which has already been accomplished, Stern's M/M/1/K approximation technique is examined to determine how best to select a value for K which will yield both accurate and computationally efficient results. It is then shown how the approximation technique can be extended to analyze the M(t)/M/1 queue where we assume that the M(t) arrival process can be approximated by a *discretely time-varying* Poisson process.

An approximate expression for the departure process of the M/M/1 queue is also proposed which implies that for an M(t)/M/1 queue whose arrival process is discretely time-varying, so too the departure process can be approximated as discretely time-varying (albeit with a different time-varying parameter).

In all cases, the accuracy and validity of these results are examined by comparison with exact analytic results, simulation or alternative discrete-time approaches (e.g., the embedded Markov chain technique of Moore [MOOR 1972] [MOOR 1975]).

1. INTRODUCTION

The concepts of stationarity, ergodicity, and convergence are fundamental to the analysis of queues in equilibrium. The basic result of such analysis is that it is possible to infer, from a finite number of observations of the arrival and departure processes of a queue, the stochastic properties of the queue over all time. Processes that occur in real life, however, seldom satisfy these conditions and consequently do not lend themselves to steady state analysis.

Despite this fact, the study of the transient behavior of queues has been largely ignored. The reasons for this include the facts that:

1) Very few transient solutions currently exist.

2) Those that do exist are extremely complicated to derive and manipulate.

3) Solutions are generally limited to single queues.

Transient analysis has, moreover, come to be viewed in two somewhat different ways. One view is that it deals primarily with the analysis of stationary queues approaching equilibrium. In this context, transient queueing analysis is concerned with deriving the behavior of queues as they converge to equilibrium as well as the speed with which the convergence occurs (see, for example, [JAIS 1960] [TAKA 1961] [FELL 1968] [BHAT 1968] [BS 1969]). A fundamental concept underlying this version of transient analysis is the relaxation time (also called the dominant time constant [GRIF 1978] [WSB 1975]) which is a measure of the time it takes a queue to "achieve" steady state. Paradoxically, the efforts put into this form of transient analysis have been generally directed at justifying the use of steady state solutions by showing that the relaxation time is negligible.

Another view of transient analysis is that it is concerned with studying the behavior of queues whose input and service processes are time dependent; more precisely, the parameters of the input and service processes are held to be functions of time. Such queues are obviously nonstationary and will never achieve equilibrium. The techniques for analyzing the behavior of nonstationary queues fall into two major categories: those which treat the parameters of the arrival and service processes as continuous functions of time, and those which treat (or approximate) the parameters as discrete (step) functions of time.

A number of the techniques treating the parameters as continuous functions of time rely upon the Kolmogorov difference-differential equations for birth-death processes. Using these equations, Reuter and Ledermann [RL 1953] examined the M(t)/M(t)/1 queue, where the argument t indicates that the process has a time varying parameter. They represented the queue by a Markov process with an enumerable set of states, and then provided sufficient conditions for the existence of a unique solution to the transition probability $P_{ik}(t)$, the conditional probability that the system is in state k at time t, given that it was in state i at time 0. These conditions include any one of the following:

1) $\lambda_n = 0$ for some $n \geq 1$

2) $\lambda_n > 0$ for $n \geq i$ and $\sum_{n=i}^{\infty} w_n = \infty$

$$\text{where } w_n = \frac{1}{\lambda_n} + \frac{\mu_n}{\lambda_n \lambda_{n-1}} + \ldots + \frac{\mu_n \cdots \mu_{i+1}}{\lambda_n \cdots \lambda_i} + \frac{\mu_n \cdots \mu_i}{\lambda_n \cdots \lambda_i} \quad (n \geq i)$$

3) $\lambda_n > 0$ for $n \geq i$ and $\sum_{n=i}^{\infty} \frac{1}{\lambda_n} = \infty$

4) $\lambda_n > 0$ for $n \geq i$ and $\sum_{n=1}^{\infty} \frac{\mu_n \mu_{n-1} \cdots \mu_i}{\lambda_n \lambda_{n-1} \cdots \lambda_i} = \infty$

5) $\lambda_n > 0$ for $n \geq i$ and $\mu_n = 0$ for infinitely many n

6) $\lambda_n > 0$ for $n \geq i$ and $\mu_n > 0$ for $n \geq N \geq i$ some N.

The λ_i and μ_i are, respectively, the birth and death transition rates associated with the underlying Markov process.

Takacs [TAKA 1955] investigated the virtual waiting time process of a single server queue. For such a queue with nonhomogeneous Poisson arrivals (i.e., with parameter $\lambda(t)$) and general service time distribution, $H(x)$, he was able to derive the integro-differential equation for the waiting time distribution, $F(t,x)$, at time t:

$$\frac{\partial F(t,x)}{\partial t} = \frac{\partial F(t,x)}{\partial x} - \lambda(t)F(t,x) + \lambda(t)\int_0^x H(x-y)d_y F(t,y)$$

He then derived an expression for $\phi(t,s)$, the Laplace-Stieltjes transform of the waiting time:

$$\phi(t,s) = e^{st-[1-\psi(s)]\Lambda(t)} \quad 1 - s\int_0^t F(u,0)\, e^{-su+[1-\psi(s)]\Lambda(u)} \, du$$

where $\psi(s)$ is the Laplace-Stieltjes transform of the service time distribution and $\Lambda(t) = \int_0^t \lambda(u)\, du$. Takacs showed that if the emptiness probabilities for the system can be determined uniquely, then the waiting time distribution follows directly. He also noted, however, that deriving the emptiness probabilities is usually quite difficult.

Clark [CLAR 1956] attempted to solve the forward Kolmogorov equations for the M(t)/M(t)/1 queue using a generating function approach. No exact representation of the queue length distribution was derived, but the problem was "reduced" to obtaining the solution of an integral equation. Expressions were derived, however, for the mean and variance of the queue length which depend on the emptiness probabilities for the system. Explicit results for the queue length were obtained only for the special cases where $\lambda(t)/\mu(t) = c$, some constant.

Another continuous time approach (proposed by Luchak [LUCH 1956]) involves the queue $M(t)/E^Y/1$, where E^Y here denotes a weighted-sum erlang distribution whose integer parameter is a non-negative random variable Y. Luchak found that if the arrival process parameter can be expressed as a polynomial in time, then the queue length distribution can be formulated recursively; only in selected cases, however, can a closed form solution be found. Luchak also noted that periodic arrival rates can be dealt with more easily than most since the queue length distribution need only be found for the initial period. The state at the end of each period can then be used as the initial condition for the next period. He further suggested that if the initial state of a queue with periodic arrivals was taken to be the steady state solution of a queue with constant arrival rate equal to the average rate over a period, then a "quasi-stationary" steady state could be reached in only a few periods.

Keilson and Kooharian [KK 1960] examined a multi-dimensional phase space model for a queue with time-dependent Poisson arrivals and general service time distribution, and later extended their analysis to general time-dependent arrivals [KK 1962].

Hasofer [HASO 1964] [HASO 1965], using the results of Reich [REIC 1958] [REIC 1959] in which it was shown that the time dependent emptiness probabilities are the unique solution of a Volterra equation of the first kind, derived explicit results for the M(t)/G/1 queue. Assuming a special form for the Poisson parameter $\lambda(t)$, and taking $\int_0^t \lambda(u)\, du = t - zb(t)$, Hasofer was able to write the emptiness probabilities for the queue at time t as a power series in z, $P(x,t) = \sum_{n=0}^{\infty} z^n F_n(t)$. Using Reich's equations, some complex analysis and the added assumption that $b(t)$ and $b'(t)$ are uniformly bounded, Hasofer was able to find a general expression for the Laplace transform of the F_n.

More concise results were obtained when the Poisson parameter was assumed to be periodic. For $\lambda(t) = t - z \sum_{n=1}^{n} a_n \sin(nwt+\phi_n)$, Hasofer showed that the functions F_n have the asymptotic form:

$$F_n(t) \cong \sum_{K=1}^{\infty} [A_{Kn}\cos (Kwt) + B_{Kn}\sin(Kwt)]$$

and the Laplace-Stieltjes transform of the waiting time has a similar asymptotic form.

Leese and Boyd, in a later work [LB 1966], revisited the numerical method proposed by Luchak and showed that it becomes intractable if the system of interest is observed for any significant period of time. They also analyzed a number of other approaches ranging from direct solutions of the Kolmogorov difference-differential equations and generating function techniques, to a Taylor series expansion of the queue length probabilities, a matrix approach using step function approximations of the arrival parameter and an integral equation technique developed by Wragg [WRAG 1963] which is somewhat similar to that described by Clarke. In all cases, based on computer resources available at that time, the computational requirements were found to be excessive.

Neuts [NEUT 1970] examined the transient behavior of two queues in tandem, where the second queue is assumed to have finite buffer capacity. The input to the first queue is assumed to be Poisson, while the service times of the first queue are assumed to be derived from a general distribution. The analysis relies on several embedded Markov renewal processes.

Rider [RIDE 1976] described an approximation to the expected queue length of a time-dependent M/M/1 queue. He reduces the difference-differential equations associated with the M/M/1 queue to a single equation for the expected queue length which is dependent on $P_o(t)$, the probability that the system is empty at time t. Exact expressions for $P_o(t)$ under restricted conditions are provided as well as an approximation relating $P_o(t)$ to the expected queue length at time t for those cases when the rate of change of the queue length is less than the service data.

Ross [ROSS 1978] examined a non-stationary M/G/1 queue where the time-dependent arrival parameter $\lambda(t)$ is itself governed by a stochastic process. Ross conjectures that the expected time a customer spends in this queue is greater than the expected waiting time for a customer in a queue whose arrival process parameter is stationary.

Kotiah [KOTI 1978] proposed approximations for the time-dependent expected queue length of an M/M/1 queue. He discusses iterative approaches that yield rational approximations to the complex root U* which lies inside the unit circle of a quadratic obtained by a Laplace transform-generating function technique. Extension of the methods to the queues M/M/2 and $M/E_K/1$ are also discussed.

Middleton [MIDD 1979] also examined the numerical solution of inverse transform, and established a number of tables which can be used in the solution of time-dependent M/G/1 queues.

McClish [MCCL 1979] examined the M(t)/G/1 queue where it was assumed that the arrival parameter is of the form $\lambda(t) = \lambda_o\ (1 + \varepsilon\phi(t))$, and the formulas are series expansions in ε. It is shown that under a variety of circumstances, periodicity of the input is sufficient to guarantee that a quasi-limiting distribution of the queue size exists.

Kambo and Bhalaik [KB 1979] examined two $M/M/\infty$ queues in tandem without feedback, where the arrivals and departures have time-dependent parameters. They provide necessary and sufficient conditions for the two queue length distributions to be independent. They also observe that with certain restrictions on the service rate, the system is Poisson in the limiting case.

Agrawala and Tripathi [AT 1979] present a method of obtaining the transient solution of a general single server queue which relies on characterizing the queueing system in terms of the virtual waiting time. The approach makes no assumptions about the independence of the arrivals and provides exact expressions for the virtual waiting time distribution and expected virtual waiting time.

An example of the application of transient analysis to a real problem is described by Bookbinder and Martell [BM 1979]. They employ a finite buffer, time-dependent single server queue to model helicopter allocation for forest fires. A straightforward Runge-Kutta method is used to solve the associated finite set of difference-differential equations.

One final continuous time technique of note involves the diffusion approximation ([NEWE 1968] [NEWE 1971]). The technique approximates queue processes and characteristics by continuous functions, but is accurate only for heavily utilized (ρ near 1) queues.

Of the techniques employing a discrete time approach, one of the first was offered by Galliher and Wheeler [GW 1958]. In their technique, a system is observed at fixed, equidistant points in time. The number of customers in the system at each of these instants is examined via a Markov chain analysis whose transition probabilities are derived from a Poisson distribution.

Leese and Boyd [LB 1966] also employed a discrete time approach to study the M(t)/M/1 queue. Their technique involves the juxtaposition of a series of M/M/1 queues; a finite time analysis then has to be performed during each interval in which the arrival parameter is constant. The solution they proposed, however, involves the use of infinite sums of Bessel functions to derive the transient characteristics of each queue, and quickly becomes intractable.

Eisen and Tainiter [ET 1963] and Yechiali and Naor [YN 1969] investigated queues for which the arrival (and possibly service) rate is not a deterministic function of time. The arrival rate is instead assumed to be a heterogeneous Poisson process governed by an extraneous phase process which is itself a continuous time Markov chain. Customers arrive in a Poisson stream with the arrival rate dependent upon the phase of the queue; the amount of time spent in each phase is exponentially distributed. Service times are either constant or vary with each phase. The major results derived are for the case of a process with two phases, where the service time is exponentially distributed.

Neuts [NEUT 1971a] extended the work of Yechiali and Naor to examine a sequence of M/G/1 queues each having parameters which are fixed over time intervals whose lengths are exponentially distributed and where the parameters attain new values in accordance with a given set of probabilities. This approach is generally inapplicable to most useful queueing problems where significant parameter changes occur in a more deterministic fashion.

Another technique proposed by Neuts [NEUT 1971b] for the GI/G/1 queue consists of dividing the time axis into units which correspond to the shortest possible service time and estimating the distribution of arrivals during each such interval. A sequence of queue length distributions can then be derived by studying a Markov chain whose states are 2-tuples comprised of the number of customers in the system and the number of units of service remaining for the customer currently being served.

More recently, generalized version of the methods initially developed by Luchak and Clarke have appeared. Koopman

[KOOP 1972] studied a system which has a periodic arrival process and a finite queue. The arrival and service time parameters are, therefore, dependent both upon time and the current queue length. Koopman's technique then involved a numerical solution of a finite set of differential equations.

Moore [MOOR 1972] [MOOR 1975] employed an embedded Markov chain approach to solve the $M^X(t)/E^Y/1$ queue, where the notation M^X denotes a compound Poisson process for which customers arrive in groups of random size X which are exponentially distributed. Since the compound Poisson process permits the ratio of the variance to the mean to be greater than 1, a wider range of input processes can be approximated. The generality of the E^Y service distribution is well known and was established in [GAVE 1954] [LUCH 1956]. Moore's basic approach uses the Chapman-Kolmogorov equations for an M/G/1 queue. The regeneration points for the imbedded Markov chain are the departure instants of customers from the queue.

In a later paper, Minh [MINH 1978] explores the $M^X(t)/G/1$ queue in discrete time. In this technique, the system is observed at equally spaced intervals and all events of the system (e.g., arrivals, transfers from queue to service, and departures) are assumed to occur at instants immediately prior to these epochs. Minh obtains the Chapman-Kolmogorov difference equations for a multivariate Markov chain which are then used to express the vitual waiting time probabilities in terms of the emptiness probabilities. A recurrence relation for calculating the emptiness probabilities is also derived. The technique allows the calculation of such measures as expected departure and waiting times for each customer, expected number of customers in the system at each epoch, and residual service time probabilities.

It has become apparent that the most viable techniques for analyzing the transient behavior of queues center about a discrete time approach ([MOOR 1975] [MINH 1978]). The rigid assumptions, mathematical intractability and computational complexity inherent to continuous time techniques limit their flexibility and extensibility. Yet, while the discrete time approaches reduce (if not overcome) these difficulties, they still present sizable challenges in trying to accurately represent queues via discretely time-varying processes and analyzing the behavior of complex networks of such queues; in fact, little work has been performed to date in the latter area.

Our objective in this paper, then, is to begin to lay the groundwork for the approximate (yet effective) transient analysis of queueing networks. We propose to do this in two stages. In Section 2.1, we will develop a modified version of the finite time analysis initially proposed by Leese and Boyd [LB 1966] for the M(t)/M/1 queue. Our modification centers about providing an approximate method (based on work by Stern [STER 1979]) for deriving the transient behavior of an M/M/1 queue which will replace the exact solution technique for obtaining the transient behavior based upon infinite sums of Bessel functions. This technique is both computationally efficient and can yield results to any desired level of accuracy.

In Section 2.2 we will show that by making certain assumptions about the ergodicity of the input and service processes of the M(t)/M/1 queue over finite intervals, an input process which is discretely time-varying exponential (i.e., M(t)) yields a departure process which is also of the form M(t). This result, which is an approximation, greatly facilitates the integration of a set of M(t)/M/1 queues into a network, and the subsequent transient solution of that network. It is worth noting that to date, only one result relating to departure processes of time dependent queues has been derived; this when Mirasol [MIRA 1963] and Kendall [KEND 1964] showed that the departure process for the $M(t)/G/\infty$ queue is $M(t)$.

Following these two developments, the results of several experiments validating our analysis will be provided in Section 3. A summary of our current work and a description of future research are developed in Section 4.

2. APPROXIMATE TRANSIENT ANALYSIS OF THE M(t)/M/1 QUEUE

2.1 THE APPROXIMATION TECHNIQUE

The precise situation with which we are concerned is shown in Fig. 2-1. In consonance with other discrete time approaches, we will make the assumption that the arrival rate parameter, $\lambda(t)$, varies (or can be approximated as varying) with time in a step-wise fashion, and that between those instants when $\lambda(t)$ changes, it can be assumed to be stationary. We will further assume that the service parameter remains constant for all intervals.

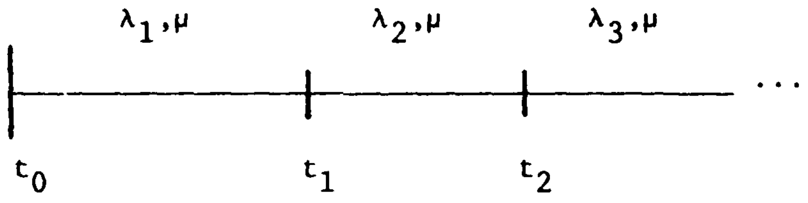


Figure 2-1

The instants at which $\lambda(t)$ changes are given by t_0, t_1, ... and the intervals bounded by those instants are $I_1 = [t_0,t_1)$, $I_2 = [t_1,t_2)$, etc. Since the service parameter remains constant for all intervals, we have within each interval an M/M/1 queue. Our analysis, therefore, conveniently decomposes into the sequential analysis of a set of (stationary) M/M/1 queues over finite intervals; it is the analysis of these individual queues which concerns us in this section.

An approximate transient analysis of the stationary M/M/1 queue has been performed by Stern [STER 1979] and relies heavily upon techniques developed by Keilson [KEIL 1974] [KEIL 1964] [KEIL 1965] [KEIL 1966]. In trying to overcome the complexity associated with deriving exact expressions for such quantities as the transient expected queue length or throughput, which involve infinite sums of Bessel functions [SAAT 1961], Stern approximates the M/M/1 queue by a finite buffer, M/M/1/K queue. The justifications for such an approximation include the facts that:

- Over a finite interval, only a limited number of arrivals can occur; hence, providing infinite buffer capacity is unnecessary.
- Few real systems have infinite buffer space and consequently do not exhibit true M/M/1 behavior.

Stern begins with the fundamental transition rate equation

$$p' = Q * p \qquad (1)$$

where Q is the (K+1)*(K+1) infinitesimal transition rate matrix associated with the finite Markov chain underlying the M/M/1/K queue, p is the state probability vector, and p' is the derivative of p. By changing variables, it is possible to obtain

$$u' = E{**}(-1/2) * p \qquad (2)$$

where

$$E = \begin{bmatrix} e_1 & & & 0 \\ & e_2 & & \\ & & \ddots & \\ 0 & & & e_K \end{bmatrix}$$

and

$$e_n = \rho{*}{*}n\ (1-\rho)/(1-\rho{*}{*}(K+1)),\ \rho = \lambda/\mu$$

which is just the steady state solution of being in state n $(\leq K)$ for the M/M/1/K queue.

Proceeding, we can define

$$u' = -\mu * A * u \tag{3}$$

where

$$A = \begin{bmatrix} \rho\ \sqrt{\rho}\ 0\ \ldots & & \\ -\sqrt{\rho}\ (1+\rho)\ -\ \sqrt{\rho}\ 0\ \ldots & & \\ & \ddots & \\ & & -\sqrt{\rho}\ 1 \end{bmatrix}$$

Now, the transformation in (2) symmetrizes the matrix Q in (1). Consequently, the finite Markov chain associated with Q is time reversible and a spectral representation of the time dependent behavior of the chain is available for which the eigenvalues and eigenvectors are real. In fact, both Q and A have real eigenvalues and the solution to equation (3) may, therefore, be written as

$$u(t) = \left[\sum_{i=0}^{K} \exp(-\mu{*}l_i{*}t)\ u_i\ u_i^T\ \ u(0)\right]. \tag{4}$$

where l_i and u_i are the i^{th} eigenvalue and eigenvector (arranged in ascending order) of A, respectively. It is worth noting that $l_o = 0$ and $l_i > 0,\ i \neq 0$.

The transient probability distribution for the number in system can then be given by

$$p(t) = p_o + \sum_{i=1}^{K} c_i p_i \exp(-\mu * l_i * t) \qquad (5)$$

where p_o is the steady state queue length distribution, and

$$p_i = E**(1/2) * u_i \qquad (5.1)$$

$$c_i = u_i^T * E**(-1/2) * p(0) \qquad (5-2)$$

The expected queue length at time t is seen to be given by

$$\bar{n}(t) = \beta^T * p(t) \ , \ \beta = [0,1,2,\ldots,K]^T$$

or

$$\bar{n}(t) = n_o + \sum_{i=1}^{K} c_i \ \beta^T \ p_i \ \exp(-\mu * l_i * t) \qquad (6)$$

where n_o is the steady state expected queue length.

The application of (5) and (6) to our particular problem is immediate. Since we have assumed stationarity within each of the intervals I_j, we are dealing with a sequence of M/M/1 queues which we can approximate as a sequence of M/M/1/K queues. Some fundamental questions remain, however. The first concerns the way in which the behavior of the queue is approximated when $\rho \geq 1$. The technique developed by Stern is valid only when $\rho < 1$, since the associated Markov chain is otherwise not ergodic and the time reversibility argument does not apply. As a result, we will employ the following alternative approximation. If we assume that the queue length at the beginning of some interval I_j for which $\rho \geq 1$ is given by $\bar{n}(t_{j-1})$, then we will approximate the transient expected queue length by:

$$\bar{n}(\tau) = \bar{n}(t_{j-1}) + (\lambda - \mu) *(\tau - t_{j-1}) , \tau \varepsilon I_j$$

A second question centers about the selection of the value of K_j in a particular interval I_j (for which $\rho < 1$). Our objectives in selecting a value for K_j are threefold:

- To ensure that the limiting (i.e., steady state) expected queue lengths of our M/M/1/K approximation and the ideal M/M/1 queue are within a sufficient distance of each other.
- To ensure that the rate at which our M/M/1/K approximation approaches its steady state value is comparable to that associated with the M/M/1/∞ queue. The term commonly used to quantify this rate of approach to steady state is the relaxation time.
- To ensure that computational tractability is maintained.

In this regard, two alternative methods of selecting K_j were considered. The first sought to define K_j based on a 2σ estimate of the queue length during the interval I_j. Thus, if at the beginning of I_j, $\bar{n}(t_{j-1}) = 0$, then K_j would be defined as:

$$K_j = \bar{n}_{oj} + 2\sigma_{n_{oj}}$$

where

$$\bar{n}_{oj} = \rho_j/(1-\rho_j)$$

= the steady state expected queue length

and

$$2\sigma_{n_{oj}} = 2 \sqrt{\rho_j(1+\rho_j^2)/(1-\rho_j)^3}$$

$$= 2\rho_j/(1-\rho_j)\ \sqrt{(1/\rho_j + \rho_j)/(1 - \rho_j)}$$

$$= \text{twice the standard deviation of } \bar{n}_{oj}$$

On the other hand, if at the beginning of an interval I_j, $\bar{n}(t_{j-1}) > 0$, a choice for K_j of

$$K_j = \bar{n}(t_{j-1}) + \bar{n}_{oj} + 2\sigma_{n_{oj}}$$

will do the same.

A second method of defining K_j is closely tied to the relaxation time of the associated approximation. By Stern's method, the relaxation time of a particular M/M/1/K approximation is given by:

$$\tau \cong 1/\mu\ (1+\rho - 2\sqrt{\rho}\ \cos\theta)$$

where

$$\theta = \pi/K+1$$

and the expression

$$1+\rho - 2\sqrt{\rho}\ \cos\theta$$

is equal to the least non-zero eigenvalue associated with the matrix A defined earlier. Obviously, as $K \rightarrow \infty$, ℓ approaches its minimum value (and τ its maximum).

It was decided that K_j would be defined such that the relaxation time associated with the M/M/1/K queue was equal to 2/3 that of the M/M/1 case. Thus, if at the beginning of I_j, $\bar{n}(t_{j-1}) = 0$, then

$$K_j = \text{INT}\ (\cos^{-1}(1.5- .25\ \frac{(\mu+\lambda_j)}{\sqrt{\mu\lambda_j}})) +1$$

where INT is a function which yields the greatest integer less than or equal to its argument. If at the beginning of I_j, $\bar{n}$ $(t_{j-1}) > 0$, then we define

$$K_j = \text{INT}\ (\bar{n}\ (t_{j-1}) + \cos^{-1}\ (1.5- .25\ \frac{(\mu+\lambda_j)}{\sqrt{\mu\lambda_j}}\ )) +1 \qquad (7)$$

Tables 2-1 and 2-2 show how the M/M/1/K approximation associated with the first and second definitions, respectively, of K_j compare with the M/M/1 case and each other. The initial queue length is assumed to be zero in all cases. As can be seen, it is not until $\rho > .88$ that the 2σ-based definition of K_j yields values for the expected steady state queue length and relaxation time which are closer to the exact M/M/1 case than those of equation (7). As a consequence, equation (7) was chosen to serve as our general definition for K_j.

The third question stems from the fact that the value of K_j usually varies over different intervals, creating a problem for the effective approximate analysis of a series of such queues. From (5.2) and (7), it is evident that the distribution p(t) plays an important part in the transition between distinct intervals. In fact, the initial queue length distribution p(0) appearing in equation (5.2) for a particular interval will be set equal to the distribution p(t) at the end of the preceeding interval. It will therefore be necessary to "translate" the K_{j+1}-dimensional queue length distribution vector $p_j(t)$ at the end of some interval I_j into the $K_{j+1}+1$-dimensional queue length distribution vector $p_{j+1}(0)$ at the beginning of the succeeding interval I_{j+1}.

TABLE 2-1

2σ-BASED DEFINITION RESULTS

ρ	K_j	$\bar{N}_{APP}$*	$\bar{N}_{EXACT}$**	$\bar{N}_{APP}/\bar{N}_{EXACT}$
0.10	0		0.11111	
0.20	1	0.16667	0.25000	0.66667
0.30	2	0.34532	0.42857	0.80576
0.40	3	0.56158	0.66667	0.84236
0.50	5	0.90476	1.00000	0.90476
0.60	8	1.40838	1.50000	0.93892
0.70	14	2.26178	2.33333	0.96933
0.80	29	3.96282	4.00000	0.99070
0.81	32	4.23161	4.26316	0.99260
0.82	35	4.52711	4.55556	0.99376
0.83	38	4.85510	4.88235	0.99442
0.84	42	5.22614	5.25000	0.99545
0.85	47	5.64701	5.66667	0.99653
0.86	52	6.12496	6.14286	0.99709
0.87	59	6.67820	6.69231	0.99789
0.88	67	7.32192	7.33333	0.99844
0.89	77	8.08211	8.09091	0.99891
0.90	89	8.99314	9.00000	0.99924
0.91	105	10.10628	10.11111	0.99952
0.92	126	11.49680	11.50000	0.99972
0.93	155	13.28382	13.28571	0.99986
0.94	196	15.66567	15.66667	0.99994
0.95	259	18.99958	19.00000	0.99998
0.96	363	23.99986	24.00000	0.99999
0.97	560	32.33331	32.33333	1.00000
0.98	1029	49.00000	49.00000	1.00000
0.99	2899	99.00001	99.00001	1.00000

*Expected steady state queue length for this approximation.

**Expected steady state queue length for M/M/1 queue.

TABLE 2-2

RELAXATION TIME-BASED DEFINITION RESULTS

ρ	K_j	$\bar{N}_{APP}$*	$\bar{N}_{EXACT}$**	$\bar{N}_{APP}/\bar{N}_{EXACT}$
0.10	3	0.11071	0.11111	0.99640
0.20	5	0.24962	0.25000	0.99846
0.30	7	0.42805	0.42857	0.99878
0.40	9	0.66562	0.66667	0.99843
0.50	12	0.99841	1.00000	0.99841
0.60	17	1.49817	1.50000	0.99878
0.70	24	2.32998	2.33333	0.99856
0.80	39	3.99468	4.00000	0.99867
0.81	42	4.25816	4.26316	0.99883
0.82	44	4.54960	4.55556	0.99860
0.83	47	4.87609	4.88235	0.99872
0.84	50	5.24299	5.25000	0.99866
0.85	54	5.65945	5.66667	0.99873
0.86	58	6.13480	6.14286	0.99860
0.87	63	6.68360	6.69231	0.99871
0.88	69	7.32424	7.33333	0.99876
0.89	76	8.08115	8.09091	0.99879
0.90	84	8.98902	9.00000	0.99878
0.91	94	10.09890	10.11111	0.99879
0.92	106	11.48572	11.50000	0.99876
0.93	122	13.26937	13.28571	0.99877
0.94	143	15.64722	15.66667	0.99876
0.95	173	18.97685	19.00000	0.99878
0.96	217	23.97024	24.00000	0.99876
0.97	291	32.29327	32.33333	0.99876
0.98	439	48.93933	49.00000	0.99876
0.99	884	98.87863	99.00001	0.99877

*Expected steady state queue length for this approximation.

**Expected steady state queue length for M/M/1 queue.

A problem arises, however, when for two contiguous intervals, I_j and I_{j+1}, $K_j \neq K_{j+1}$. In this case, the dimensions of the associated $p_j(t)$ and $p_{j+1}(t)$ are different, and an exact assignment cannot be made. Some possible approaches for solving this problem include:

1) Setting all K_j's equal to a number which is greater than or equal the largest possible value of any K_j over the intervals of interest. The advantage of this approach is that it guarantees that all $p_j(t)$'s have the same dimension. The disadvantages, however, include:

   - Computational inefficiency, since one is always working with the maximum M/M/1/K approximation between intervals.

   - It will usually be impossible to determine just what the appropriate maximum value for K should be. Recall that the K's are determined dynamically as the analysis proceeds and rely on information obtained not only from the current interval but the preceding interval as well.

2) If $K_j \leq K_{j+1}$, projecting (in the strict mathematical sense) $p_j(t_j - t_{j-1})$ directly in $p_{j+1}(0)$. For example, if we let $p_j(t) = (p_{j,0}(t), p_{j,1}(t), \ldots, p_{j,m}(t))$ and $p_{j+1}(t) = (p_{j+1,0}(t), \ldots, p_{j+1,n}(t))$ where $m \leq n$, then projecting $p_j(t)$ into $p_{j+1}(t)$ implies that $p_{j,0}(t) = p_{j+1,0}(t)$, $p_{j,1}(t) = p_{j+1,1}(t)$, ..., $p_{j,m}(t) = p_{j+1,k}(t)$ and $p_{j+1},\ell(t) = 0 \; \forall \ell$ such that $k < \ell \leq n$. On the other hand, if $K_j > K_{j+1}$, then we must "scale" the dimension of $p_j(t)$ down to that of $p_{j+1}(t)$ ensuring, however, that $\bar{n}(t_j) = E[p_j(t_j - t_{j-1})]$ equals $\bar{n}(t_j) = E[p_{j+1}(0)]$.

One possible method of performing this scaling is as follows. Let nmax be the least integer which is greater than or equal to $\bar{n}(t_j)$. Define $p_{j+1,0}(0) = 1 - \bar{n}(t_j)/\text{nmax}$, $p_{j+1,\text{nmax}}(0) = \bar{n}(t_j)/\text{nmax}$, and $p_{j+1,i}(0) = 0$ for all $i \neq 0$ or nmax. It is obvious then that with this representation,

$$\sum_{i=0}^{K_{j+1}} p_{j+1,i}(0) = 1 \quad \text{and} \quad E[p_{j+1}(0)] = \tilde{n}(t_j).$$

and we have achieved the required mapping between p_j and p_{j+1}.

Using Stern's basic technique together with the solutions described above, a useful and computationally efficient method becomes available for analyzing the approximate transient behavior of both a stationary M/M/1 queue and the (discretely time-varying) M(t)/M/1 queue. In the next section, we investigate those aspects of the M(t)/M/1 queue which will permit us to extend the analysis described above to networks of such queues.

2.2 APPROXIMATE CHARACTERIZATION OF THE DEPARTURE PROCESS

This section develops an approximate characterization of the departure process for the M(t)/M/1 queue described above and details how this characterization can potentially facilitate the solution of general networks comprised of such queues. Although several papers [CD 1974] [NATV 1975] [DALE 1976] have examined the topic of departure processes for a wide variety of queues, none have discussed in any depth the nature of the departure process during a transient period.

In order for us to begin to characterize the departure process, we make the following observations and approximations:

- Using the techniques described in Section 2.1, we can obtain an approximate value for $\tilde{n}(t)$ for any t. Given that each customer has a mean service time of $1/\mu$, the expected virtual waiting time for the queue at the beginning of some interval I_{j+1} is $\bar{n}(t_j)/\mu$. This means that during the interval $[t_j, t_j+(\bar{n}(t_j)/\mu)) = [t_j,t_j+v)$, the server is expected to be busy. The departure process will, therefore, appear as the service process and have parameter μ.

- Given that the expected virtual waiting time at the beginning of some interval I_{j+1} is non-zero and that $\lambda_{j+1} \neq 0$, then the departure process will have parameter μ up to the instant T^i when the server first goes idle (which may exceed t_j+v above).
- After the first instant T^i, $t_j+v \leq T^i < t_{j+1}$, such that the server goes idle, we approximate the departure process as the input process with parameter λ_{j+1}. This approximation makes liberal use of Burke's Theorem [BURK 1956] which states that M => M; its accuracy will depend in part upon the relaxation time.

Our first objective then is to determine when (and if) the first (expected) instant, T^i_j, occurs in interval I_j such that the server goes idle. We will assume that $p_j(t_{j-1})$ and $\bar{n}(t_{j-1})$ are known. While these last parameters are assumed to be derived by the methods presented in Section 2.1, our subsequent analysis deals explicitly with the M/M/1 queue and does not involve its M/M/1/K approximation.

Recalling that fundamental to any M/M/1 queue is a birth-death process with transition rates λ and μ, we find ourselves dealing with an infinite Markov chain whose states are just the number in system. If we denote the passage time probability density function (pdf) from a state n to state 0 by $s_{n,0}(\tau)$, and the corresponding random variable by $T_{n,0}$, it is immediate that

$$s_{n,0}(\tau) = s_{\bar{n}}(\tau) \circledast s_{\overline{n-1}}(\tau) \circledast \dots \circledast s_{\bar{1}}(\tau) \qquad (8)$$

where $s_{\bar{n}}(\tau) = s_{n,n-1}(\tau)$ and $\circledast$ is the convolution operator. In terms of random variables we have

$$T_{n,0} = T_{n,n-1} + T_{n-1,n-2} + \dots + T_{1,0}$$

A recursive probabilistic argument obtains $s_{n,0}(\tau)$ as follows. Let $v = \lambda+\mu$. If we assume that we are in some state n, the dwell time in that state has a pdf given by v*exp(-v*t). Furthermore, with probability $\lambda/\lambda+\mu$ there will be a transition to state n+1, while with probability $\mu/\lambda+\mu$ the transition will be to state n-1. Hence,

$$s_{\bar{n}}(\tau) = \frac{\mu}{v}\, ve^{-v\tau} + \frac{\lambda}{v}\, ve^{-v\tau} \circledast s_{\overline{n+1}}(\tau) \circledast s_{\bar{n}}(\tau) \qquad (9)$$

The recursion is obtained, apart from inversion difficulties, by using Laplace transforms. The transform of (9) is then seen to be

$$\begin{aligned} \sigma_{\bar{n}}(s) &= \frac{\mu}{s+\lambda+\mu} + \frac{\lambda}{s+\lambda+\mu} \cdot \sigma_{\overline{n+1}}(s) \cdot \sigma_{\bar{n}}(s) \qquad (10) \\ &= \frac{\mu}{s+\lambda+\mu} \Big/ \left[1 - \frac{\lambda}{s+\lambda+\mu} \cdot \sigma_{\overline{n+1}}(s)\right] \\ &= \frac{\lambda}{s+\lambda+\mu - \lambda\sigma_{\overline{n+1}}(s)} \end{aligned}$$

from which one can further derive

$$\frac{d}{ds}\,\sigma_{\bar{n}}(s) = \frac{-\mu\left(1 - \lambda \frac{d}{ds}\,\sigma_{\overline{n+1}}(s)\right)}{\left(s+\lambda+\mu - \lambda\sigma_{\overline{n+1}}(s)\right)^2} \qquad (11)$$

Clearly,

$$- \frac{d}{ds} \sigma_{\bar{n}}(0) = E[T_{\bar{n}}] = \bar{T}_{\bar{n}} \qquad (12)$$

where

$$T_{\bar{n}} = T_{n,n-1}$$

Hence,

$$\bar{T}_{\bar{n}} = \mu^{-1}(1 + \lambda \; \bar{T}_{\bar{n}}+1) \qquad (13)$$

The mean time from state n to n-1 may be described in more compact form with the aid of the parameters π_n defined by:

$$\pi_o = 1 \; ; \; \pi_n = (\lambda/\mu)^n = \rho^n$$

where

$$\lambda\pi_n = \mu\pi_{n+1}$$

Multiplying (13) by $\mu\pi_n$, we obtain

$$\mu\pi_n\bar{T}_{\bar{n}} - \mu\pi_{n+1}\bar{T}_{\bar{n}+1} = \pi_n \qquad (14)$$

from which we can derive

$$\bar{T}_{\bar{n}} = \frac{1}{\mu\pi_n} \mathop{}\limits_{j=n}^{\infty} \pi_j \qquad (15)$$

It follows that

$$
\begin{aligned}
\bar{T}_{n,o} &= \sum_{m=0}^{n-1} \frac{1}{\mu\pi_m} \sum_{j=m}^{\infty} \pi_j \\
&= \sum_{m=0}^{n-1} \frac{1}{\mu\rho^m} \sum_{j=m}^{\infty} \rho^j \\
&= \sum_{m=0}^{n-1} \frac{1}{\mu\rho^m} \left(\frac{1}{1-\rho} - \sum_{j=0}^{m-1} \rho^j \right) \\
&= \sum_{m=0}^{n-1} \frac{1}{\mu\rho^m} \cdot \frac{\rho^m}{1-\rho} \\
&= \sum_{m=0}^{n-1} \frac{1}{\mu-\lambda} \\
&= \frac{n}{\mu-\lambda} \qquad (16)
\end{aligned}
$$

Thus within a particular interval I_j, we see that the time of first idle is given by:

$$
\begin{aligned}
T_j^i &= t_{j-1} + \sum_k p_{j,k}(t_{j-1}) \frac{k}{\mu-\lambda_j} \\
&= t_{j-1} + \bar{n}(t_{j-1})/(\mu-\lambda) \\
&= t_{j-1} + \bar{T}_{n,o} \qquad (17)
\end{aligned}
$$

Now if $T_j^i \geq t_j$ or if $\lambda_j/\mu \geq 1$ (i.e., $\rho_j \geq 1$), then it is expected that the server will not be idle during I_j. We can therefore make the following observations:

1) In the case where either $T_j^i \geq t_j$ or $\lambda_j \geq \mu$, the departure process during I_j can be expected to appear as the service process with parameter μ.

2) In the case where $T_j^i = t_{j-1}$ (i.e., where $\bar{n}(t_{j-1}) = 0$) and $\lambda_j < \mu$, the departure process can be approximated as the input process with parameter λ_j.

3) When $t_{j-1} < T_j^i < t_j$ and $\lambda_j < \mu$, the departure process can be expected to appear as the service process with parameter μ up to time T_j^i after which is can be approximated as the input process with parameter λ_j.

Based on these observations, a general expression for the interdeparture time pdf, d(s), within an interval I_j becomes:

$$\begin{aligned} d(s) = {} & p'_j(s)\mu e^{-\mu(s-t_{j-1})} + (1-p'_j(s))\,[\rho_j \mu e^{-\mu(s-t_{j-1})} \\ & + (1-\rho_j)\lambda_j e^{-\lambda_j(s-t_{j-1})} \circledast \mu e^{-\mu(s-t_{j-1})}] \end{aligned} \tag{18}$$

where the density function $p'_j(s)$ is defined as:

$$p'_j(s) = \begin{cases} 1 & \text{if} \quad t_{j-1} \leq s < T_j^i \\ 0 & \text{otherwise} \end{cases}$$

This expression really implies, however, that d(s) is a stepwise time-varying exponential distribution; hence, the departure process is of the form M(t). It is worth noting, though, that the form of M(t) for the departure process will usually not be equal to the form of M(t) for the input process. More specifically, the instants at which the parameters associated with the arrival and departure process change, as well as the values they acquire between those instants, will usually be different.

There are several implications of these results. Probably the most significant is the fact that the approximate analysis of networks of M(t)/M/1 queues now becomes tractable. Since the output of any M(t)/M/1 queue is now in the same form as the input to all others, the interconnection of such queues is straightforward.

A second major implication is that it is now possible to approximately determine (for networks of such queues) just how long it takes a change in the arrival rate of one M(t)/M/1 queue to propagate through a series of such queues, and moreover, when equilibrium will be re-established for the entire network.

Finally, from the point of view of potential practical applications, the ability to approximately analyze the transient behavior of networks of such queues lends itself immediately to such problems as the modeling of dynamic routing strategies, internetworking problems and network stability analysis.

3. EXPERIMENTAL RESULTS

This section summarizes the results of several experiments that were performed to validate the approximation techniques presented in Section 2. In all cases, the latter techniques were compared to the output of simulations or analytically-derived exact values. The experiments had four objectives:

1) *Given stationary values for λ and μ, to* determine the relative accuracy of the M/M/1/K approximation developed in section 2.1 compared to the exact (analytical) M/M/1 case and simulation.

2) To evaluate the proposed (approximate) method of analyzing the behavior of a nonstationary M(t)/M/1 queue as a sequence of stationary M/M/1/K queues.

3) To determine the validity of approximately characterizing the departure process of an M(t)/M/1 queue as a discretely time-varying M(t) process.

4) To provide a comparison of the M/M/1/K approximation technique with Moore's [MOOR 1975] embedded Markov chain approach.

In sections 3.1 - 3.4 below, each of these objectives is separately addressed.

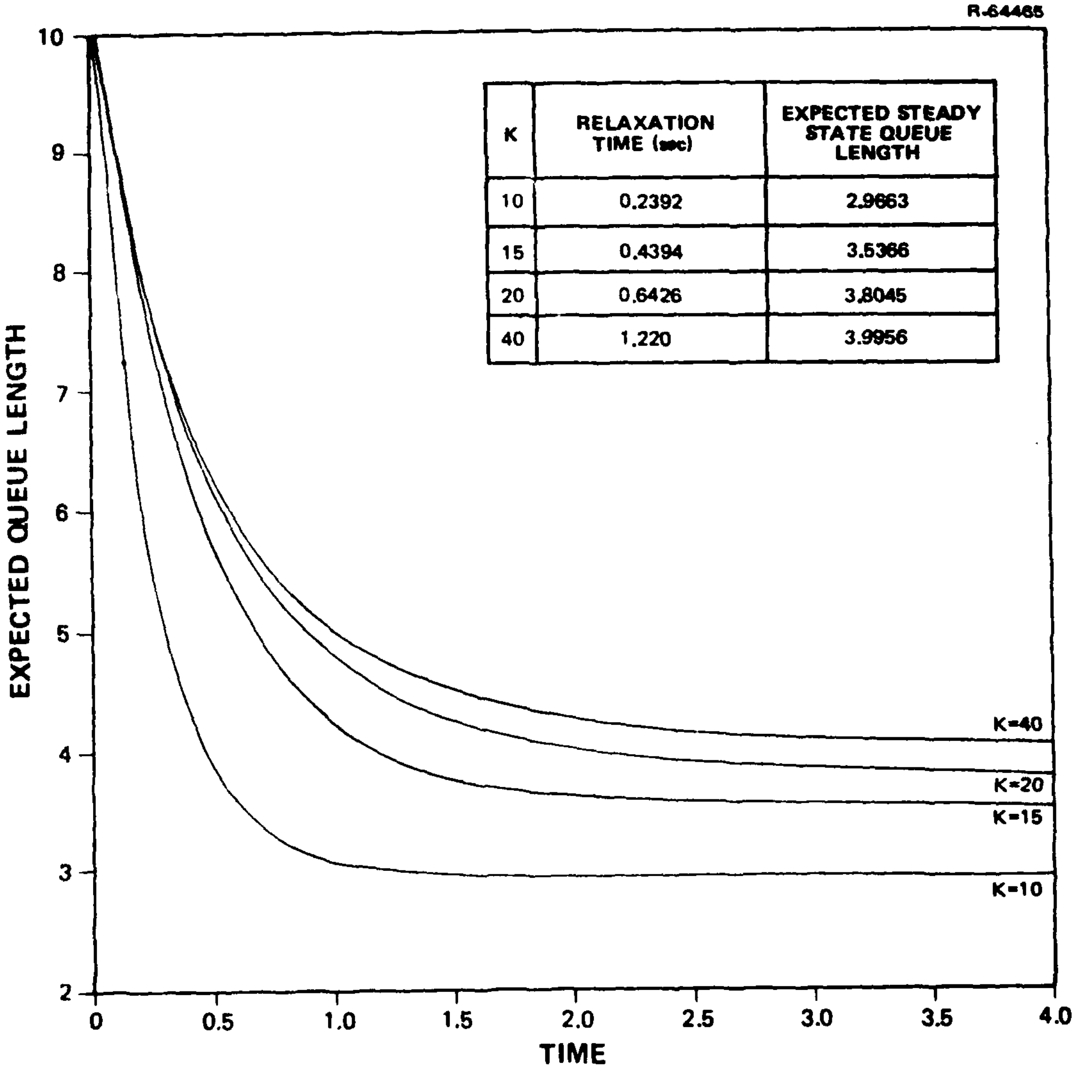
3.1 M/M/1/K APPROXIMATION RESULTS

A number of experiments were run to determine the sensitivity and accuracy of the M/M/1/K approximation described in section 2.1 to variations in the values of K, initial queue length and ρ. Figure 3-1 illustrates the way in which the transient expected queue length varies over time for K=10, 15, 20 and 40. In all cases, the initial queue length is assumed to be equal to 10 and $\rho=.8$ (with $\mu=50$).

The variations that can be seen in the curves associated with the different values of K derive from two sources. The first centers about the fact that the steady state values for the expected queue length are monotonically increasing with K; hence, the M/M/1 value is asymptotically approached as $K \rightarrow \infty$. The second involves the relaxation times associated with particular M/M/1/K approximations and the fact that they too are monotonically increasing with K.

The combined effect of these two conditions is that, in comparison to the exact M/M/1 case, an M/M/1/K approximation based on a small value of K approaches steady state more rapidly and with an asymptotic expected queue length that is less than an approximation based on a larger value of K. It was for these reasons that, in attempting to select a value for K (given a fixed initial queue length and ρ), we required that the M/M/1/K approximation provide an expected steady state queue length within .2% of the expected M/M/1 value together with a relaxation time within 33% of the M/M/1 value while at the same time maintaining computational tractability.

Figure 3-2 illustrates how our technique performs in comparison to simulations based on the same underlying parameters of $\rho=.8$, $\mu=50$ and an initial queue length of 10. Given these conditions, our technique specifies that we use an M/M/1/K approximation for which K=49. As can be seen, close agreement is obtained; in fact, the average difference between the results



K	RELAXATION TIME (sec)	EXPECTED STEADY STATE QUEUE LENGTH
10	0.2392	2.9663
15	0.4394	3.5366
20	0.6426	3.8045
40	1.220	3.9956

Figure 3-1

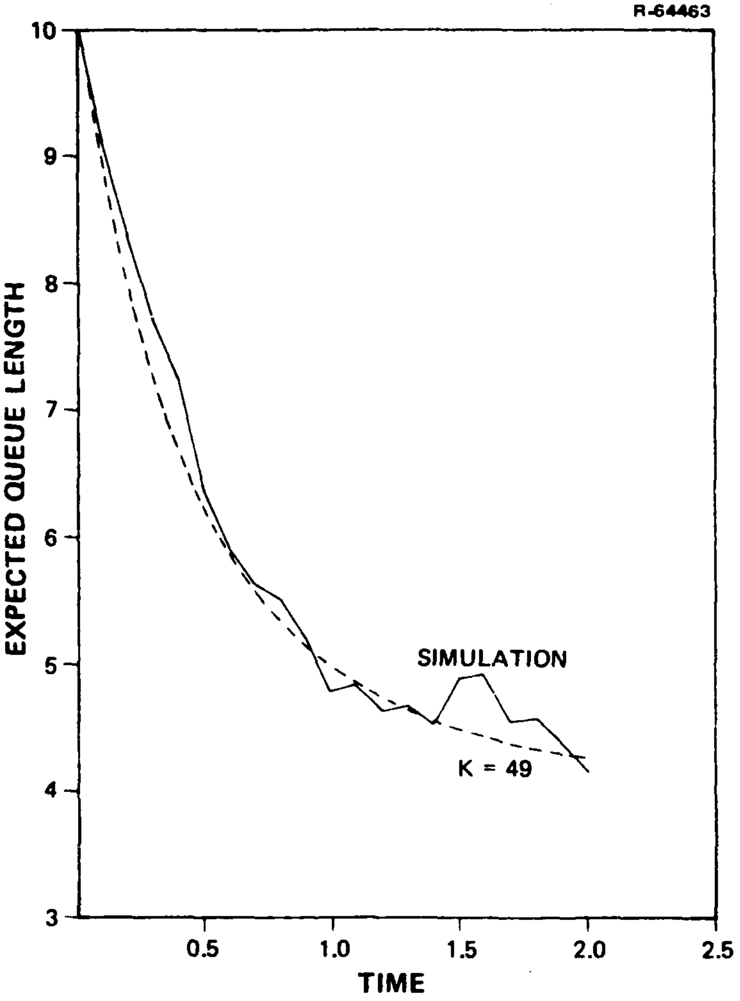

R-64463
10
9
8
7
6
5
4
3
EXPECTED QUEUE LENGTH
0.5
1.0
1.5
2.0
2.5
TIME
SIMULATION
K = 49

Figure 3-2

associated with the approximation and those associated with the simulation was only 0.179, with a standard deviation of 0.153. Similar agreement was obtained in other experiments.

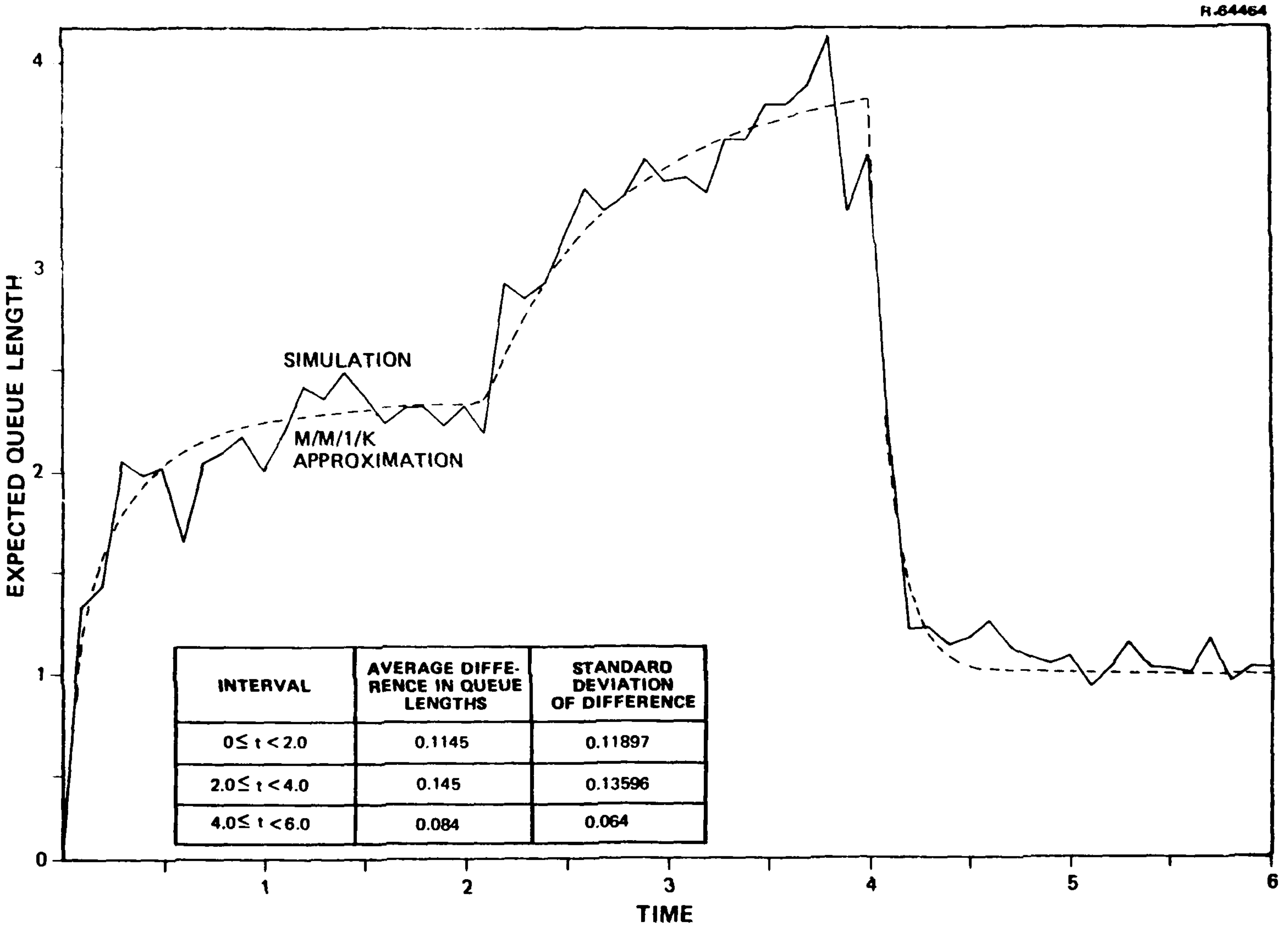
3.2 M(t)/M/1 APPROXIMATION RESULTS

A series of experiments was run to compare the transient expected queue lengths associated with an M(t)/M/1 queue derived both by simulation and the approximate sequential M/M/1/K technique described in section 2.1. Figure 3-3 illustrates the outcome of analyzing one particular M(t)/M/1 queue for which the arrival process was permitted to vary such that λ = 35, 40 and 25 for three consecutive intervals, each of 2 second duration. For all intervals, μ = 50 and was fixed.

Table 3-1 compares the expected queue lengths at the end of each interval obtained by our approximation and the simulation with those values associated with an M/M/1 queue. Since for each interval, the interval length is greater than its associated relaxation time, the (absolute) difference between the expected queue length obtained by simulation ($\bar{N}_{sim}$) and the M/M/1 value ($\bar{N}_{ss}$), and the difference between $\bar{N}_{sim}$ and the expected queue length obtained by our approximation ($\bar{N}_{app}$), serve as relative figures of merit.

TABLE 3-1

END OF ...	ρ	$\bar{N}_{sim}$	$\bar{N}_{app}$	$\bar{N}_{ss}$	$\lvert\bar{N}_{sim}-\bar{N}_{ss}\rvert$	$\lvert\bar{N}_{sim}-\bar{N}_{app}\rvert$
INTERVAL 1	0.7	2.308	2.319	2.333	0.025	0.011
INTERVAL 2	0.8	3.545	3.818	4.000	0.455	0.273
INTERVAL 3	0.5	1.018	0.999	1.000	0.018	0.019



INTERVAL	AVERAGE DIFFE-RENCE IN QUEUE LENGTHS	STANDARD DEVIATION OF DIFFERENCE
0≤ t < 2.0	0.1145	0.11897
2.0≤ t < 4.0	0.145	0.13596
4.0≤ t < 6.0	0.084	0.064

Figure 3-3

To further demonstrate the "goodness of fit" of the approximation results with those obtained by simulation, the table contained in Figure 3-3 shows how the mean queue lengths of each differed over the three intervals and includes standard deviations for each. Once again, the close agreement observed was repeated in other similar experiments.

3.3 DEPARTURE PROCESS APPROXIMATION RESULTS

Several experiments were run to validate our contention that during an interval whose initial queue length is greater than zero, the departure process of an M/M/1 queue can be approximated as the discretly time-varying M(t) process described in section 2.2. Specifically, our contention is that between the beginning of the interval and the time of first idle, T^i, the departure process has a probability density function which can be approximated as $\mu e^{-\mu t}$ $(0 \leq t < T^i)$. After T^i, we assume that the departure process has a pdf which can be approximated by $\lambda e^{-\lambda t}$ $(T^i \leq t)$.

Table 3-2 illustrates how the expected time of first idle obtained via simulation (T^i_{sim}) compares with our analytic result (T^i_{an}) for ρ = .5, .8 and .9, μ = 50 (and fixed), and initial queues lengths of 5 and 10 customers. In all cases, the close agreement predicted by Eqs. 16 and 17 was observed.

Table 3-3 shows how the interdeparture distributions derived from simulations associated with several intervals having different initial queue lengths compare to the predicted approximate values. Both the interdeparture distribution mean and standard deviation for the "pre-T^i" period and "post-T^i" period are given. Once again, the close agreement observed, combined with similar results from other experiments, lend

TABLE 3-2

INITIAL QUEUE LENGTH	ρ (μ=50)	T^i_{sim}	T^i_{an}	$\|T^i_{sim} - T^i_{an}\|$
10	0.5	0.3979	0.4	0.0021
10	0.8	0.9867	1.0	0.0133
10	0.9	1.8734	2.0	0.1266
5	0.5	0.2027	0.2	0.0027
5	0.8	0.4735	0.5	0.0265
5	0.9	0.9851	1.0	0.0149

TABLE 3-3

T-4475

		PRE - T^i			POST - T^i		
INITIAL QUEUE LENGTH	ρ (μ=50)	$\bar{d}^*$	σ^*_d	APPROX. MEAN=S.D.	$\bar{d}$	σ_d	APPROX. MEAN=S.D.
10	0.5	0.0198	0.0194	0.02	0.0405	0.0405	0.04
10	0.8	0.0204	0.0206	0.02	0.0254	0.0256	0.025
10	0.9	0.0201	0.0200	0.02	0.0230	0.0228	0.0222
5	0.5	0.0201	0.0196	0.02	0.0405	0.0406	0.04
5	0.8	0.0200	0.0206	0.02	0.0255	0.0254	0.025
5	0.9	0.0202	0.0202	0.02	0.0229	0.0228	0.0222

*$\bar{d}$ = mean interdeparture time; σ_d = standard deviation of interdeparture time.

increasing confidence to the accuracy and robustness of the approximation.

3.4 COMPARISON WITH MOORE'S TECHNIQUE

In this section, we provide some comparative results between the methods presented in this paper and those proposed by S.C. Moore [MOOR 1972] [MOOR 1975]. As was mentioned in the introduction, Moore's technique centers about the $M^X(t)/E_Y/1$ queue and employs an imbedded Markov chain approach to analyze its behavior. In applying this approach, the system status is observed only at departure instants. If a simple imbedded Markov chain approach (which assumed finite buffer size) were used, computational results could be obtained very quickly; however, all connection with the real or continuous time axis would be lost.

Realizing this, Moore developed an exact technique for maintaining the time axis connection in the case of a stationary arrival process, and an approximate technique for doing the same when the arrival process is nonstationary. The latter approximation resides in determining the times of the nth services initiation and completion. If one begins with $T_0 = 0$ and a known queue length distribution, it is necessary to apply the equations

$$T_n^1 = T_{n-1} + X_n$$

$$T_n = T_n^1 + S_n$$

alternately in order to maintain exact correspondence between the imbedded Markov chain behavior (associated with the queue) and the continuous time axis. In the equations, T_n is the time of the nth customers departure, X_n is the server idle time in the interval (T_{n-1}, T_n), and S_n is the length of the nth customer's service time.

In order to avoid the complexity of obtaining exact distributions for all terms involved, Moore develops approximations based on the expected values of T_n^1 and T_n. These are then used to derive both the time from departure n to the next arrival (which depends on T_{n-1}) and the number of arrivals during service n (which depends on T_n^1). The approximation is due to the fact that once the arrival parameters shift, the expected values $E[T_n^1]$ and $E[T_n]$ are only approximate.

Figure 3-4 shows how Moore's technique compares with that developed in this paper. As in Figure 3-3, the analysis proceeds over three intervals, the first two of which of 2 second duration each and the third which is infinite. The ρ associated with the intervals is .7, .8 and .5, respectively. As is seen, the two methods exhibit quite close agreement.

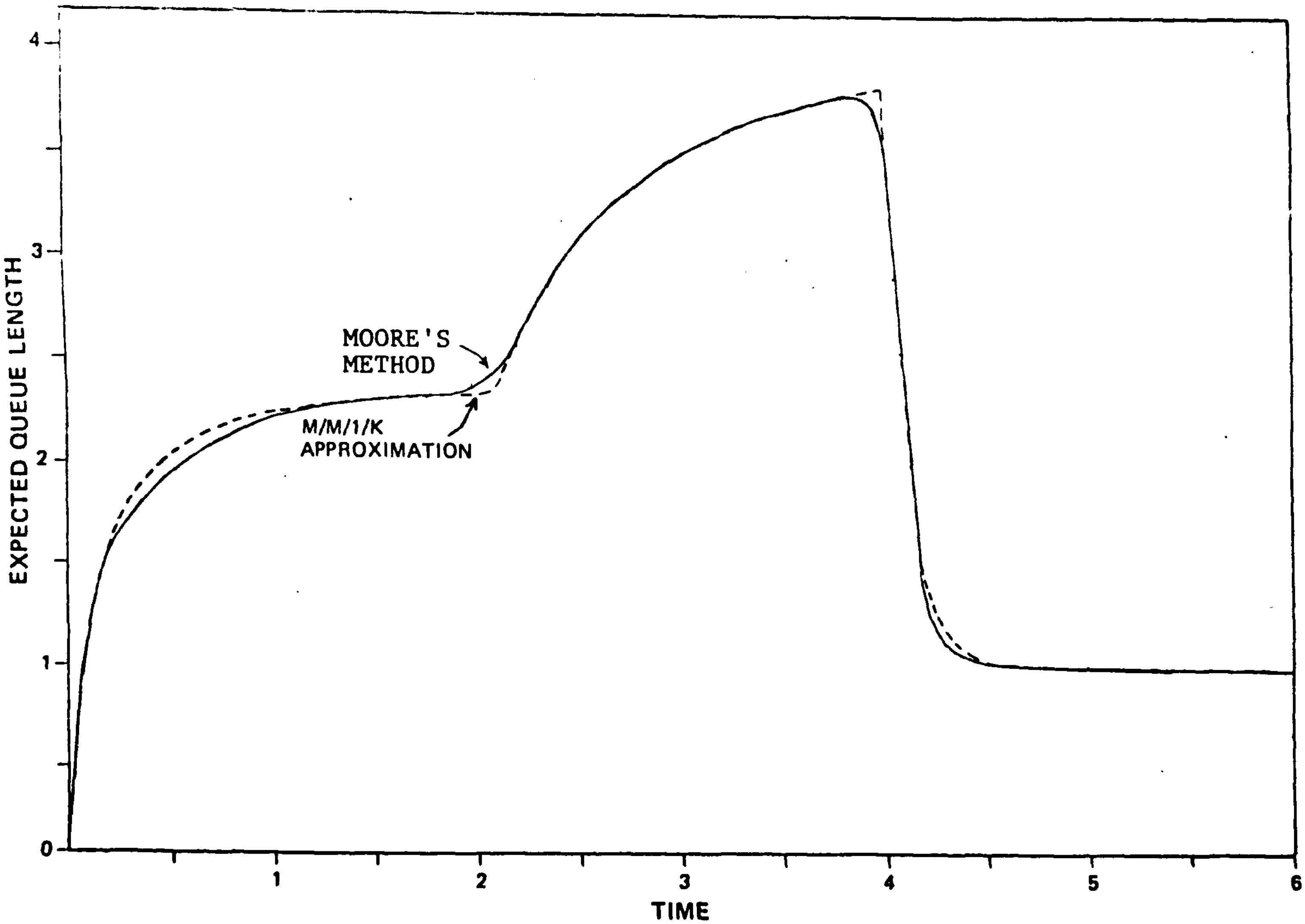

4
3
2
1
0
EXPECTED QUEUE LENGTH
MOORE'S METHOD
M/M/1/K APPROXIMATION
1
2
3
4
5
6
TIME

Figure 3-4

4. SUMMARY OF RESULTS AND FUTURE WORK

In the previous sections, we have examined in some detail Stern's M/M/1/K approximation technique for obtaining the transient behavior of an M/M/1 queue and proposed methods by which it can be applied in a practical, computationally efficient manner.

We have also shown how the latter technique can be extended to accommodate the analysis of the M(t)/M/1 queue where we assume that the M(t) arrival process can be approximated by a discretely time-varying Poisson process. The analysis of the M(t)/M/1 queue then becomes the analysis of a sequence of stationary M/M/1/K queues over finite intervals whose boundaries are defined by the instants when the arrival parameter changes.

Together with the above results, we have proposed an approximate expression for the departure process of the M/M/1 queue which implies that for an M(t)/M/1 queue whose arrival process is discretely time-varying M(t), so too the departure process can be approximated as discretely time-varying (albeit of a different form). The simulation results of Section 3.3 have shown this approximation to be accurate.

Finally, we have compared our technique with that proposed by S.C. Moore using an imbedded Markov chain approach. Although Moore's technique permits the modeling of a wider range of queues than our technique at this time, in the particular case of the M(t)/M/1 queue it is felt that our technique is more accurate. This is true especially during transitions to and from intervals for which $\rho \cong 1$ or $\rho > 1$ when the arrival process may change and a number of arrivals occur between two consecutive departures. The fact that our technique is tightly coupled to the instants when the parameters associated with the underlying Poisson processes change allows it to "react" more rapidly to varying queue dynamics.

The results of this paper will serve as the basis for further research into a number of areas including:

- Extension and refinement of the finite buffer approximation technique to more general queues including the $M(t)/E_y/1$ $M^x(t)/E_y/1$ and $E_y^x(t)/E_y/1$ queues.
- Extension of the methods and results described herein to methods for analyzing feed forward and general queueing networks.
- Examination of the effects that different service disciplines and multiple classes have on the behavior of such queues.
- Identification of the critical error sources associated with the approximation techniques and better quantification of their effects.
- Application of the methods to several practical problems including internetworking problems, network stability analysis, design and modeling of dynamic and other routing strategies, and optimal resource sharing in a transient environment.

REFERENCES

1. [AT 1979] Agrawala, A.K. and Tripathi, S.K., "Transient Solution of the Virtual Waiting Time of a Single Server Queue and Its Applications," Int. Journal of Information Sciences, Vol. 21, July 1980, pp. 141-158.

2. [BHAT 1968] Bhat, U.N., "Transient Behavior of Multiserver Queues with Recurrent Input and Exponential Service Times," JAP 5, 1968, pp. 158-168.

3. [BM 1979] Bookbinder, J.H. and Martell, D.L., "Time-Dependent Queueing Approach to Helicopter Allocation for Forest Fire Initial-Attack," INFOR, Vol. 17, No. 1, February 1979, pp. 58-70.

4. [BS 1969] Bhat, U.N. and Sahin, I., "Transient Behavior of the Queueing Systems M/D/1, M/Ek/1, D/M/1, and Ek/M/1," Tech. Mem. 135, Dept. of Op. Res., Case Western Reserve University, 1969.

5. [BURK 1956] Burke, P.J., "The Output of a Queueing System," Op. Res. 4, 1956, pp. 699-704.

6. [CD 1974] Cherry, W.P. and Disney, R.L., "Some topics in queueing network theory," in Mathematical Methods in Queueing Theory, ed. A.B. Clarke, Lecture Notes in Economics and Mathematical Systems 98, Springer-Verlag, New York, pp, 23-44.

7. [CLAR 1956] Clarke, A.B., "A Waiting Line Process of Markov Type," Annals of Mathematical Statistics, Vol. 27, 1956, pp. 452-459.

8. [DALE 1976] Daley, D.J., "Queueing output processes," Adv. App. Prob. 8, 1976, pp. 395-415.

9. [ET 1963] Eisen, M. and Tainiter, M., "Stochastic Variations in Queueing Processes," Op. Res. 11, 1963, pp. 922-927.

10. [FELL 1968] Feller, W., An Introduction to Probability Theory and Its Applications, Vol. 1., John Wiley & Sons, Inc., 1968 (Third Edition).

REFERENCES (Continued)

11. [GAVE 1954] Gaver, D.P., "The Influence of Servicing Times in Queueing Processes," Op. Res. 2, 1954, pp. 139-149.

12. [GRIF 1978] Griffin, W.C., Queueing, Basic Theory and Applications, Grid, Inc., Columbus, Ohio, 1978.

13. [GW 1958] Galliher, H.P. and Wheeler, R.C., "Non-stationary queueing probabilities for landing congestion of aircraft," Op. Res. 6, 1958, pp. 607-623.

14. [HASO 1964] Hasofer, A.M., "On the Single-Server Queue with Nonhomogeneous Poisson Input and General Service Time," JAP 1, 1964, pp. 369-384.

15. [HASO 1965] Hasofer, A.M., "Some Perturbation Results for the Single Server Queue With Poisson Input," JAP 2, 1965, pp. 462-466.

16. [JAIS 1960] Jaiswal, N.K., "Time-Dependent Solution of the Bulk Service Queueing Problem," Op. Res. 8, 1960, pp. 773-781.

17. [JH 1966] Jackson, R.R.P. and Henderson, J.C., "The Time-Dependent Solution to the Many Server Poisson Queue," Op. Res. 14, No. 4, 1966, pp. 720-722.

18. [KB 1979] Kambo, N.S. and Bhalaik, H.S., "Non-Homogeneous $M/M/\infty$ Queues in Series," INFOR, Vol. 17, No. 3, August 1979, pp. 262-275.

19. [KEIL 1964] Keilson, J., "A Review of Transient Behavior in Regular Diffusion and Birth-Death Processes, Part I," JAP 1, 1964, pp. 247-266.

20. [KEIL 1965] Keilson, J., "A Review of Transient Behavior in Regular Diffusion and Birth-Death Processes, Part II," JAP 2, 1965, pp. 405-428.

21. [KEIL 1966] Keilson, J., "The Ergodic Queue Length Distribution for Queueing Systems with Finite Capacity," No. 1, 1966, pp. 190-201.

22. [KEIL 1974] Keilson, J., Markov Chain Models - Rarity and Exponentiality," Rep. No. CSS 74-01, Center for System Science, University of Rochester, Rochester, N.Y., Sept. 1974.

REFERENCES (Continued)

23. [KEND 1953] Kendall, D.G., "Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Embedded Markov Chain," Annals of Math. Stat. 24, 1953, pp. 338-354.

24. [KK 1960] Keilson, J. and Kooharian, A., "On Time-Dependent Queueing Processes," Annals of Math. Stat. 31, 1960, pp. 104-11.

25. [KK 1962] Keilson, J., and Kooharian, A., "On the General Time-Dependent Queue with a Single Server," Annals of Math. Stat. 33, 1962, pp. 767-791.

26. [KOOP 1972] Koopman, B.O., "Air Terminal Queues Under Time Dependent Conditions," Op. Res. 20, 1972, pp. 1089-1114.

27. [KOTI 1978] Kotiah, T.C.T., "Approximate Transient Analysis of Some Queueing Systems," Op. Res. 26, No. 2, 1978, pp. 333-346.

28. [LB 1966] Leese, E.L. and Boyd, D.W., "Numerical Methods of Detemining the Transient Behavior of Queues with Variable Arrival Rates," Jour. Canadian Op. Res. Soc, 4, 1966, pp. 1-13.

29. [LUCH 1956] Luchak, G., "The Solution of the Single-Channel Queueing Equations Characterized by a Time-Dependent Poisson-Distributed Arrival Rate and a General Class of Holding Times," Op. Res. 4, 1956, pp. 711-732.

30. [MCCL 1979] McClish, D., "Queues and Stores with Nonhomogeneous Input," Institute of Statistics Mimeo Series #1225, Dept. of Statistics, Chapel Hill, N.C., April 1979.

31. [MIDD 1979] Middleton, M.R., "Transient Effects in M/G/1 Queues: An Empirical Investigation," Tech. Rep. No. 85, Dept. of Operations Research, Stanford University, Stanford, CA, June 1979.

32. [MINH 1978] Minh, D.L., "The Discrete-Time Single-Server Queue with Time Inhomogeneous Compound Poisson Input and General Service Time Distribution," JAP 15, 1978, pp. 590-601.

REFERENCES (Continued)

33. [MIRA 1963] Mirasol, N.M., "The Output of an M/G/∞ Queueing System is Poisson," Op. Res. 11, 1963, pp. 282-284.

34. [MOOR 1972] Moore, S.C., "Approximate Techniques for Nonstationary Queues," Ph.D. Thesis, SMU, 1972.

35. [MOOR 1975] Moore, S.C., "Approximating the Behavior of Non-Stationary Single-Server Queues," Op. Res. 23, No. 5, 1975, pp. 1011-1032.

36. [NATV 1975] Natvig, B., "On the Input and Output Processes for a General Birth-and-Death Queueing Model," Adv. Applied Prob. 7, 1975, pp. 576-592.

37. [NEUT 1970] Neuts, M.F., "Two Servers in Series, Studied in Terms of a Markov Renewal Branching Process," Advances in Applied Prob. 2, 1970, pp. 110-149.

38. [NEUT 1971a] Neuts, M.G., "The Single Server Queue in Discrete Time," Mimeograph Series No. 270, Dept. of Stat., Purdue University, 1971.

39. [NEUT 1971b] Neuts, M.G., "A Queue Subject to Extraneous Phase Changes," Adv. in App. Prob. 3, 1971, pp. 78-119.

40. [NEWE 1968] Newell, G.F., "Queues with Time-Dependent Rates," JAP 5, 1968.

    I - The transition through saturation pp. 436-451.

    II - The maximum queue and return to equilibrium, pp. 579-590.

    III - A mild rush hour, pp. 591-606.

41. [NEWE 1971] Newell, G.F., *Applications of Queueing Theory*, Chapman and Hall Ltd., Great Britain, 1971.

42. [REIC 1958] Reich, E., "On the Integro-Differential Equation of Takacs, I," Annals of Maths. Stat. 29, 1958, pp. 563-570.

43. [REIC 1959] Reich, E., "On the Integro-Differential Equation of Takacs, II," Annals of Math. Stat. 30, 1959, pp. 143-148.

REFERENCES (Continued)

44. [RIDE 1976] Rider, K.L., "A Simple Approximation to the Average Queue Size in the Time-Dependent M/M/1 Queue," JACM, Vol. 23, No. 2, April 1976.

45. [RL 1953] Reuter, G.E.H. and Ledermann, W., "On the Differential Equations for the Transition Probabilities of Markov Processes with Enumerably Many States," Proc. of the Cambridge Philosophical Society 49, 1953, pp. 247-262.

46. [ROSS 1978] Ross, S.M., "Average Delay in Queues with Non-Stationary Poisson Arrivals," Jour. of App. Prob. 15, 1978, pp. 602-609.

47. [SAAT 1961] Saaty, T.L., Elements of Queueing Theory, New York, McGraw Hill, 1961.

48. [SATY 1966] Satymurty, P.R., "Queueing with Balking - A Simple Method of Study the Transient Behavior," Op. Res. 14, No. 2, 1966, pp. 329-333.

49. [STER 1979] Stern, T.E., "Approximations of Queue Dynamics and Their Applications to Adaptive Routing in Computer Communication Networks," IEEE Trans. On Communications, Vol. Com-27, No. 9, Sept. 1979.

50. [TAKA 1955] Takacs, L. "Investigation of Waiting Time Problems by Reductions to Markov Processes," Acta Math. Hungar, 6., 1955, pp. 101-129.

51. [TAKA 1961] Takacs, L., "The Transient Behavior of a Single Server Queueing Process with Poisson Input," Proc. 4th Berkeley Symp., Vol. II University of California Press 1961, pp. 535-56.

52. [WRAG 1963] Wragg, A., "The Solution of an Infinite Set of Difference-Differential Equations Occurring in Polymerization and Queueing Problems," Proc. Cambridge Phil. Soc., Vol. 59, 1963, pp. 117-124.

53. [WSB 1975] White, J.A., Schmidt, J.W. and Bennett, G.K., Analysis of Queueing Systems, Academic Press, New York, 1975.

54. [YN 1969] Yechiali, U. and Naor, P. "Queueing Problems with Heterogeneous Arrivals and Service," O.R. Stat. and Econ. Mimeograph Series No. 49, Technion-Israel Institute of Tech. 1969.

DATE

FILMED

- 8