COMPUTING TRANSIENT DISTRIBUTIONS IN GENERAL SINGLE-SERVER QUEUES

David M. Lucantoni, Gagan L. Choudhury and Ward Whitt

David M. Lucantoni, AT&T Bell Laboratories, Room 3K-601, Holmdel, NJ 07733-3030, USA
Gagan L. Choudhury, AT&T Bell Laboratories, Room 3K-603, Holmdel, NJ 07733-3030, USA
Ward Whitt, AT&T Bell Laboratories, 2C-178, Murray Hill, NJ, 07974-2070, USA

Abstract We present the two-dimensional transforms of the transient workload and queue-length distributions in the single-server queue with general service times and a batch Markovian arrival process (*BMAP*). This arrival process includes the familiar phase-type renewal process and the Markov modulated Poisson process as special cases, as well as superpositions of these processes, and allows correlated interarrival times and batch sizes. Numerical results are obtained via two-dimensional transform inversion algorithms based on the Fourier-series method. From the numerical examples we see that predictions of system performance based on transient and stationary performance measures can be quite different.

1. Introduction and Summary

In this paper we consider the single-server queue with unlimited waiting space, a work-conserving service discipline and i.i.d. (independent and identically distributed) service times that are independent of a general arrival process. Our purpose is to obtain computable transient results for this general model. These transient results are important for studying real-time control of communication networks and other systems.

In order to obtain computable results, we assume that the arrival process is a *batch Markovian arrival process* (*BMAP*), as in Lucantoni [1] [2]. The *BMAP* is a convenient representation of the *versatile Markovian point process* Neuts [3] [4] or *Neuts (N) process* Ramaswami [5]. The *BMAP* generalizes the *Markovian arrival process* (*MAP*), which was introduced by Lucantoni, Meier-Hellstern and Neuts [6]. The *MAP* includes as special cases both the phase-type renewal process (Neuts [7]) and the Markov-modulated Poisson process (Heffes and Lucantoni [8]). Indeed, stationary *MAP*s are dense in the family of all stationary point processes; see Asmussen and Koole [9]. An important property of *MAP*s and *BMAP*s is that superpositions of independent processes of these types are again processes of the same type; this property is exploited in Choudhury, Lucantoni and Whitt [10] to study the effect of statistically multiplexing a large number of bursty sources.

Hence, we consider the *BMAP/G/*1 queue and derive the two-dimensional transforms of the workload (or virtual waiting time) distribution at time t and the queue-length distribution at time t. As usual with the *BMAP/G/*1 queue, these quantities are actually $m \times m$ matrices, with the $(i,j)^{\text{th}}$ element specifying that the auxiliary phase is j at time t, conditioned upon the phase at time 0 being i.

These transient results can be regarded as matrix generalizations of transient results for the *M/G/*1 queue, which can be found in Takács [11], Abate and Whitt [12] and references cited there. As in the *M/G/*1 special case, a key role here is played by the busy-period distribution and the emptiness function. These are discussed in Sections 2.4 and 3 here.

0-7803-0917-0/93$03.00 © 1993 IEEE

In fact, there is a long history of transient results for single-server queueing models generalizing *M/G/*1, as can be seen from the books by Neuts [7],[3], Takács [11], Cohen [13] and Benes [14], and references therein. With regard to the present work, the 1967 papers by Çinlar [15], [16] and the early papers of Neuts (cited in [3]) are notable.

A distinctive feature of our paper, in relation to previous papers on transient behavior for these *M/G/*1-type queues, is that *we demonstrate that our formulas are computable*. In particular, we calculate the time-dependent probability distributions by *numerically inverting the two-dimensional transforms*. For this purpose, we apply the two-dimensional transform inversion algorithms in Choudhury, Lucantoni and Whitt [17]. These algorithms are based on the Fourier-series method [18], exploiting the two-dimensional Poisson summation formula, as in (5.44)–(5.48) of [18]. For this purpose, we obtain the busy-period transform by iterating the characterizing functional equation, drawing upon Choudhury, Lucantoni and Whitt [19].

The derivations and proofs of the present results are given in Lucantoni, Choudhury and Whitt [20], along with additional results on the transient distributions as well as expressions for the moment functions of the workload at time t. Additional numerical examples and details about the implementation of the algorithms are also presented in [20].

The remainder of this paper is organized as follows. In §2 we review the definition and basic properties of the Batch Markovian Arrival Process and the single server queue with this arrival process. In particular, we review the transform of the duration of the busy period which plays a fundamental role in the transient solution of this model. In §3 we derive the Laplace transform for the probability that the system is empty at time t. Sections 4 and 5 contain the main results on the transient distributions of the workload and queue length, respectively, and numerical examples are presented in §6.

2. The *BMAP/G/*1 Queue

The Batch Markovian Arrival Process The *BMAP* is a natural generalization of the Poisson process (see Lucantoni [1]). It is constructed by considering a two-dimensional Markov process $\{N(t), J(t)\}$ on the state space $\{(i,j): i \geq 0, 1 \leq j \leq m\}$ with an infinitesimal generator Q having the structure

$$Q = \begin{bmatrix} D_0 & D_1 & D_2 & D_3 & \cdots \\ & D_0 & D_1 & D_2 & \cdots \\ & & D_0 & D_1 & \cdots \\ & & & D_0 & \cdots \\ & & & & \cdots \end{bmatrix}, \tag{1}$$

where D_k, $k \geq 0$, are $m \times m$ matrices; D_0 has negative diagonal elements and nonnegative off-diagonal elements; D_k, $k \geq 1$, are nonnegative and D, defined by

$$D = \sum_{k=0}^{\infty} D_k, \tag{2}$$

is an irreducible infinitesimal generator. We also assume that $D \neq D_0$, which assures that arrivals will occur.

The variable $N(t)$ counts the number of arrivals in the interval $(0,t]$, and the variable $J(t)$ represents an auxiliary state or phase. Transitions from a state (i,j) to a state $(i+k,l)$, $k \geq 1$, $1 \leq j,l \leq m$, correspond to batch arrivals of size k, and thus the batch size can depend on j and l. The matrix D_0 is a stable matrix (see e.g., pg. 251 of Bellman [21]), which implies that it is nonsingular and the sojourn time in the set of states $\{(i,j): 1 \leq j \leq m\}$ is finite with probability one, for all i; see Lemma 2.2.1 of Neuts [7]. This implies that the arrival process does not terminate.

Let $\boldsymbol{\pi}$ be the stationary probability vector of the Markov process with generator D, i.e., $\boldsymbol{\pi}$ satisfies

$$\boldsymbol{\pi} D = 0, \qquad \boldsymbol{\pi e} = 1, \tag{3}$$

where $\boldsymbol{e}$ is a column vector of 1's. Then the component π_j is the stationary probability that the arrival process is in state j. The arrival rate of the process is then

$$\lambda = \boldsymbol{\pi} \sum_{k=1}^{\infty} k D_k \boldsymbol{e} = \boldsymbol{\pi d}, \tag{4}$$

where $\boldsymbol{d} = \sum k D_k \boldsymbol{e}$.

Intuitively, we think of D_0 as governing transitions in the phase process which do not generate arrivals and D_k as the rate of arrivals of size k (with the appropriate phase change). For other examples and further properties of the *BMAP* see [1].

A key quantity for analyzing the *BMAP/G/*1 queue is the matrix generating function

$$D(z) = \sum_{k=0}^{\infty} D_k z^k, \quad \text{for } |z| \leq 1.$$

Let $P_{ij}(n,t) = P(\, N(t) = n,\ J(t) = j \mid N(0) = 0,\ J(0) = i\,)$ be the (i,j) element of a matrix $P(n,t)$. That is, $P(n,t)$ represents the probability of n arrivals in $(0,t]$ plus the phase transition. Then the matrix generating function $P^*(z,t)$ defined by

$$P^*(z,t) = \sum_{n=0}^{\infty} P(n,t) z^n, \quad \text{for } |z| \leq 1,$$

is given explicitly by

$$P^*(z,t) = e^{D(z)t}, \quad \text{for } |z| \leq 1,\ t \geq 0, \tag{5}$$

where $e^{D(z)t}$ is an exponential matrix (see e.g., pg. 169 of Bellman, [21]). Note that for Poisson arrivals, $m = 1$, $D_0 = -\lambda$, $D_1 = \lambda$, and $D_k = 0$, $k \geq 2$, so that (5) reduces to $P^*(z,t) = e^{-\lambda(1-z)t}$ which is the familiar generating function of the Poisson counting process.

The Queueing Model Consider a single-server queue with a *BMAP* arrival process specified by the sequence $\{D_k,\ k \geq 0\}$. Let the service times be i.i.d. and independent of the arrival process; let the service time have an arbitrary distribution function H with Laplace-Stieltjes transform (*LST*) h and n^{th} moment α_n. We assume that the mean $\alpha \equiv \alpha_1$ is finite. Let the *traffic intensity*, $\rho \equiv \lambda \alpha$.

The Embedded Markov Renewal Process at Departures The embedded Markov renewal process at departure epochs is defined as follows. Define $X(t)$ and $J(t)$ to be the number of customers in the system (including in service, if any) and the phase of the arrival process at time t, respectively. Let τ_k be the epoch of the k^{th} departure from the queue, with $\tau_0 = 0$. (We understand that the sample paths of these processes are right continuous and that there is a departure at $\tau_0 = 0$.) Then $(X(\tau_k),\ J(\tau_k),\ \tau_{k+1} - \tau_k)$ is a semi-Markov process on the state space $\{\ (i,j): i \geq 0,\ 1 \leq j \leq m\ \}$. The semi-Markov process is *positive recurrent* when $\rho < 1$. The transition probability matrix of the semi-Markov process is given by

$$Q(x) = \begin{bmatrix} \hat{B}_0(x) & \hat{B}_1(x) & \hat{B}_2(x) & \cdots \\ \hat{A}_0(x) & \hat{A}_1(x) & \hat{A}_2(x) & \cdots \\ 0 & \hat{A}_0(x) & \hat{A}_1(x) & \cdots \\ 0 & 0 & \hat{A}_0(x) & \cdots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{bmatrix}, \quad x \geq 0, \tag{6}$$

where, for $n \geq 0$, $\hat{A}_n(x)$ and $\hat{B}_n(x)$ are the $m \times m$ matrices of mass functions with elements defined by

$[\hat{A}_n(x)]_{ij}$ = $P($ Given a departure at time 0, which left at least one customer in the system and the arrival process in phase i, the next departure occurs no later than time x with the arrival process in phase j, and during that service there were n arrivals),

$[\hat{B}_n(x)]_{ij}$ = $P($ Given a departure at time 0, which left the system empty and the arrival process in phase i, the next departure occurs no later than time x with the arrival process in phase j, leaving n customers in the system).

An embedded Markov renewal process with a transition probability matrix having the structure in (6) is called "*M/G/*1-type" (Neuts [3]) since it has matrix generalizations of the skip-free-to-the-left and spatial homogeneity properties of the ordinary *M/G/*1 queue.

We introduce the transform matrices

$$A_n(s) = \int_0^{\infty} e^{-sx}\, d\hat{A}_n(x), \qquad B_n(s) = \int_0^{\infty} e^{-sx} d\hat{B}_n(x),$$

$$A(z,s) = \sum_{n=0}^{\infty} A_n(s) z^n, \qquad B(z,s) = \sum_{n=0}^{\infty} B_n(s) z^n, \tag{7}$$

where $\mathrm{Re}(s) \geq 0$ and $|z| \leq 1$. It was shown in Lucantoni [1] that

$$A(z,s) = \int_0^{\infty} e^{-sx} e^{D(z)x} dH(x) \equiv h(sI - D(z)), \tag{8}$$

and

$$B(z,s) = z^{-1}[sI - D_0]^{-1}[D(z) - D_0]A(z,s). \tag{9}$$

The definition in (8) above is consistent with the usual definition of a scalar function evaluated at a matrix argument (see Theorem 2, pg. 113 of Gantmächer, [22]). In particular, since h is analytic in the right half-plane, the above function is defined by using the matrix argument in the power series expansion of h. This is well defined as long as the spectrum of the matrix argument also lies in the right half plane. Note that from (8) we see that $A(z,s)$ is a power series in $D(z)$. Thus, $A(z,s)$ and $D(z)$ commute. This property is used repeatedly in [20].

The Busy Period Following the general treatment of Markov chains of *M/G/*1-type in [3], we define $\hat{G}_{jj'}^{[r]}(x)$, $x \geq 0$, as the probability that the first passage from the state $(i+r,\ j)$ to the state (i,j'), $i \geq 1$, $1 \leq j,\ j' \leq m$, $r \geq 1$, occurs no later than time x, and that (i,j') is the first state visited in level i. The matrix with elements $\hat{G}_{jj'}^{[r]}(x)$ is $\hat{G}^{[r]}(x)$.

By a first passage argument, it was shown in Neuts [23] that the transform matrix $G(s)$, defined by

$$G(s) = \int_0^\infty e^{-sx} d\hat{G}^{[1]}(x), \qquad \text{for } \operatorname{Re}(s) \geq 0,$$

satisfies the nonlinear matrix equation

$$G(s) = \sum_{n=0}^{\infty} A_n(s) G(s)^n. \tag{10}$$

In the context of the *BMAP/G/*1 queue, $G(s)$ governs the duration of the busy period. It was also shown in [23] that the transform matrix governing the duration of a busy period starting with r customers, is given by $G(s)^r$. Equation (10) is the key equation in the matrix analytic solution to queues of the *M/G/*1 type.

It was shown in Lucantoni [1] that $G(s)$ is also the solution to

$$G(s) = \int_0^\infty e^{-sx} e^{D[G(s)]x} dH(x) \equiv h(sI - D[G(s)]), \tag{11}$$

where $D[G(s)] \equiv \sum_{k=0}^{\infty} D_k G(s)^k$. Equation (11) is the matrix analogue of the *Kendall functional equation*, (see (59) in Kendall, [24], and the discussion of I. J. Good on pg. 182 there). In particular, if $m=1$ then the *BMAP* is a Poisson process with $D_0 = -\lambda$, $D_1 = \lambda$, and $D_k = 0$ for $k \geq 2$, so that (11) reduces to $G(s) = h(s + \lambda - \lambda G(s))$ which is (59) in [24].

The matrix G is the key ingredient in the solution of the stationary version of this system. An efficient algorithm for computing this matrix based on uniformization is given in [1]. For the transient solution, we need to compute the matrix $G(s)$ for complex s. It is shown in Choudhury, Lucantoni and Whitt [19] that $G(s)$ may be computed by iterating in (11). Convergence is guaranteed if the iteration is started with either $G_0 = 0$ or $G_0 = G$ and, in fact, if both of these iterations are carried out, then by stopping the iteration at any point the matrices obtained correspond to the transforms of distributions which bound the true distribution. This extends results for the *M/G/*1 queue in Abate and Whitt [25].

In order to compute the right hand side of (11) in each iteration, two cases are considered in [19]. If the service-time distribution has a rational Laplace transform (e.g., phase-type or other distributions in the Coxian family), then the right hand side may be computed exactly with one matrix inversion and a few matrix multiplications. If the service time distribution is not rational, then a procedure similar to uniformization is used.

3. The Emptiness Functions

In this section we characterize the probability that the system is empty at time t. The key role of this function for general systems was demonstrated by Benes [14]. We distinguish several cases depending on what information is available at $t=0$. In particular, we consider starting with a fixed amount of work x, $x \geq 0$; starting with a fixed number of customers, i_0, where $t=0$ is an epoch of departure; and starting with an amount of work which is distributed according to an arbitrary distribution F.

Let $V(t)$ be the amount of work in the system at time t; let

$$P_{x0}^{ij}(t) = P(\ V(t)=0,\ J(t)=j \mid V(0)=x,\ J(0)=i\ );$$

and let the $m \times m$ matrix $P_{x0}(t)$ have (i,j)-entry $P_{x0}^{ij}(t)$. Also, let $p_{x0}(s) = \int_0^\infty e^{-st} P_{x0}(t)\,dt$, for $\operatorname{Re}(s) > 0$. Then we have the following generalization of the *M/G/*1 formula. (See (9) on pg. 52 of [11] and (34) and (36) in [12]).

Theorem 1: The matrix $p_{x0}(s)$ is given by

$$p_{x0}(s) = e^{-(sI - D[G(s)])x} (sI - D[G(s)])^{-1}, \tag{12}$$

for Re(s)>0.

Note that the exponential disappears when $x=0$. Since the components of the vector $G(s)\boldsymbol{e}$ are Laplace-Stieltjes transforms and $|G(s)\boldsymbol{e}| < 1$, for $\operatorname{Re}(s) > 0$, the eigenvalues of $D[G(s)]$ are in the left half-plane. Therefore, for $\operatorname{Re}(s) > 0$, the eigenvalues of $sI - D[G(s)]$ are in the right half-plane and the inverse appearing in (12) is well defined.

Let $\hat{P}_{i_0 0}(t)$ be the $m \times m$ matrix with (j,k) entry

$$\hat{P}_{i_0 0}^{jk}(t) = P(V(t)=0,\ J(t)=j \mid X(0)=i_0,\ J(0)=j,\ \tau_0=0). \tag{13}$$

As a consequence of Theorem 1, we immediately have

$$\hat{p}_{i_0 0}(s) \equiv \int_0^\infty e^{-st} \hat{P}_{i_0 0}(t)\,dt = G(s)^{i_0} p_{00}(s) \tag{14}$$

$$= G(s)^{i_0} (sI - D[G(s)])^{-1}.$$

The unconditional emptiness function, starting with initial workload distributed according to cdf F, defined by

$$P_0(t) \equiv \int_0^\infty P_{x0}(t)\,dF(x), \tag{15}$$

has Laplace transform

$$p_0(s) \equiv \int_0^\infty e^{-st} P_0(t)\,dt = f(sI - D[G(s)])(sI - D[G(s)])^{-1}, \tag{16}$$

where f is the *LST* of F. We show in [20] that

$$\lim_{t \to \infty} P_0(t) = \begin{cases} (1-\rho)\boldsymbol{eg} & \text{for } \rho \leq 1, \\ 0 & \text{for } \rho > 1. \end{cases} \tag{17}$$

4. The Workload

The Transient Results In this section we derive the transform of the workload (work in the system in uncompleted service time) at time t. We accomplish this in two steps. First, we assume a departure at time $t=0$ and derive the distribution of the work in the system at some fixed time t, conditioned on the number of customers left in the system after that departure. Using this result, we derive the more general distribution of the work in the system at time t, conditioned on the amount of work at time $t=0$, where this is not necessarily an epoch of departure. Although the second result is more general, from a practical viewpoint the first might be more useful. In particular, in a real system it might be easier to measure the number of customers, packets, etc., at departure times than to know the exact amount of work in the system.

Define the $m \times m$ matrix $W_{i_0}(t,x)$, whose (i,j) entry is

$$\left[W_{i_0}(t,x)\right]_{ij} = P(V(t) \le x, J(t)=j \mid X(0)=i_0, J(0)=i, \tau_0=0);$$

i.e., $W_{i_0}(t,x)$ is the conditional delay distribution at time t given the number of customers in the system following the departure at time $t=0$. Let the transform matrices be

$$w_{i_0}(t,s) = \int_0^\infty e^{-sx}\, d_x W_{i_0}(t,x), \quad w_{i_0}(\xi,s) = \int_0^\infty e^{-\xi t} w_{i_0}(t,s)\, dt,$$

where Re $(s) \ge 0$ and Re $(\xi) > 0$. In the following theorem, the inverse need not exist for all argument pairs (ξ,s); at these points the left side is defined by continuity.

Theorem 2: The matrix $w_{i_0}(\xi,s)$ is given explicitly by

$$w_{i_0}(\xi,s) = (h(s)^{i_0} I - s\hat{p}_{i_0 0}(\xi))[\xi I - sI - D(h(s))]^{-1}, \quad (18)$$

and the matrix $w_{i_0}(t,s)$ is given by

$$w_{i_0}(t,s) = \left[h(s)^{i_0} I - s\int_0^t \hat{P}_{i_0 0}(u)\, e^{-(sI+D(h(s)))u} du\right] e^{(sI+D(h(s)))t}, \quad (19)$$

where Re $(s) \ge 0$, Re $(\xi) > 0$, and $\hat{P}_{i_0 0}(u)$ is defined in (13).

Although we are able to express the transform of the delay explicitly in terms of t in (19), we note that this expression is not trivial to evaluate numerically. It involves numerically inverting a Laplace transform where the evaluation of the transform at a value of s requires the numerical integration of the emptiness function times an exponential matrix where the values of the emptiness function are themselves obtained by inverting a Laplace transform. The corresponding expression for the ordinary $M/G/1$ queue also suffers from the same difficulty. This may partly explain why the known formulas for that case have not been widely used for practical computations.

In contrast, however, the transform expression in (18) is relatively simple to evaluate, so that with an inversion algorithm for 2-dimensional Laplace transforms, we have a practical method for obtaining numerical results. We describe such an algorithm in Lucantoni, Choudhury and Whitt [20].

It can be shown using Rouché's theorem that for each s, Re$(s) \ge 0$, the determinant of the matrix $X(s,\xi) \equiv [\xi I - sI - D(h(s))]$ appearing in the inverse in (18) has exactly m roots in the region Re$(\xi) > 0$. (For similar arguments see Çinlar [15] and Neuts [26] [27].) Since w_{i_0} is a transform and is therefore analytic in the interior of the above region, see p.26 of Deutsch [28], these pairs of (ξ,s) must also be zeros of the first matrix on the right in (18). That is, they are removable singularities. The classical approach to this type of problem would then assume that the roots are distinct to obtain m independent linear equations for each row of the matrix on the left. In practice, the roots may not be distinct, or if they are close, there may be numerical difficulties in locating these roots. These technical problems are circumvented in the present case since we derived explicit results for the matrices in (18).

Let F be the cdf of the initial work at time 0 (where $t=0$ need not be an epoch of departure) and let f be its Laplace-Stieltjes transform. Let $W(t,x)$ be the matrix whose $(i,j)^{\text{th}}$ element is the probability that the work in the system is less than x and the phase is j at time t, given that at time 0 the phase was i and the initial workload (including the customer in service, if any) was distributed according to F. Let $w(t,s)$ and $w(\xi,s)$ be the Laplace transforms

$$w(t,s) = \int_0^\infty e^{-sx} d_x W(t,x), \quad w(\xi,s) = \int_0^\infty e^{-\xi t} w(t,s)\, dt.$$

Then we have the following theorem.

Theorem 3: The Laplace transform $w(\xi,s)$ is given by

$$w(\xi,s) = (f(s)I - sp_0(\xi))[\xi I - sI - D(h(s))]^{-1}, \quad (20)$$

and

$$w(t,s) = \left[f(s)I - s\int_0^t P_0(u)\, e^{-[sI+D(h(s))]u} du\right] e^{[sI+D(h(s))]t}, \quad (21)$$

for Re$(s) \ge 0$), Re$(\xi) > 0$, where $P_0(u)$ and $p_0(\xi)$ are given in (15) and (16), respectively.

Note that Theorem 2 is a special case of Theorem 3 where $f(s) = h(s)^{i_0}$. Note that (21) is the direct analogue of Equation (8) on pg. 51 of [11].

The Limiting Distribution of the Waiting Time Differentiating with respect to t in (21), we have

$$\frac{\partial}{\partial t} w(t,s) = w(t,s)[sI + D(h(s))] - sP_0(t).$$

Therefore, using (17) and assuming that the partial derivative approaches 0 as $t \to \infty$, we see that the transform of the limiting distribution of the workload is given by $w(s) \equiv \lim_{t \to \infty} w(t,s)$ and

$$w(s) = \begin{cases} s(1-\rho)eg[sI + D(h(s))]^{-1}, & \text{for } \rho < 1, \\ 0, & \text{for } \rho \ge 1, \end{cases}$$

which agrees with (44) in [1]. Hence, by [1], the partial derivative does indeed approach 0 as $t \to \infty$.

5. The Transient Queue Length

Let $Y^{jk}_{i_0 i}(t) = P(X(t)=i, J(t)=k \mid X(0)=i_0, J(0)=j, \tau_0=0)$, and let $Y_{i_0 i}(t)$ have (j,k)-entry $Y^{jk}_{i_0 i}(t)$. Recall that $\tau_0 = 0$ means that there is a departure at time 0. Then clearly,

$$Y_{i_0 0}(t) = W_{i_0}(t,0) = \int_0^\infty dM_{i_0 0}(u)\, e^{D_0(t-u)},$$

by conditioning on the last departure before time t. Let $y_{i_0 i}(s)$ be the Laplace transform of $Y_{i_0 i}(t)$. Then $y_{i_0 0}(s) = G(s)^{i_0} p_{00}(s) = p_{i_0 0}(s)$. The probability generating function of the queue length at time t is defined by

$$y_{i_0}(z,s) \equiv \sum_{i=0}^\infty y_{i_0 i}(s) z^i.$$

Theorem 4: The matrix $y_{i_0}(z,s)$ is given by

$$y_{i_0}(z,s) = \Big[z^{i_0+1}(I-A(z,s))(sI-D(z))^{-1}$$

$$+ (z-1)\hat{p}_{i_0 0}(s)A(z,s)\Big][zI-A(z,s)]^{-1}, \quad (22)$$

for $\mathrm{Re}(s)>0$ and $|z|<1$, where $\hat{p}_{i_0 0}(s)$ is given in (14) and $A(z,s)$ is given in (8).

Equation (22) is the matrix analogue of Equation (77) on pg. 74 in [11]. Let the Laplace transform of the complementary queue length distribution be defined by

$$y^*_{i_0 i}(s) = \int_0^\infty e^{-st} \sum_{n=i+1}^\infty Y_{i_0 n}(t)\,dt,$$

with the corresponding generating function

$$y^*_{i_0}(z,s) \equiv \sum_{i=0}^\infty y^*_{i_0 i}(s) z^i .$$

Then since $y^{\cdot}_{i_0}(1,s) = (sI-D)^{-1}$, we have the following corollary.

Corollary: The transform of the complementary queue length distribution, $y^*_{i_0}(z,s)$, is given by

$$y^*_{i_0}(z,s) = \frac{1}{1-z}\left[(sI-D)^{-1} - y_{i_0}(z,s)\right].$$

6. Numerical Results

In this section, we demonstrate the computability of our results. In particular, we calculate the time-dependent probability distributions by numerically inverting the two-dimensional transforms via the inversion algorithms in Choudhury, Lucantoni and Whitt [17]. These algorithms are based on the Fourier-series method [18] and are enhancements and generalizations of the Euler and lattice-Poisson algorithms described in [18]. We refer to [20] and [17] for further discussion and examples of these algorithms.

We consider a *BMAP* which is a superposition of four independent and identical *MMPP*s. Each *MMPP* alternates between a high-rate and a low-rate state where the ratio of the arrival rates in the two states is 4:1. The durations of each state are such that there is an average of five arrivals during the sojourns in each state. The individual arrival rates are scaled appropriately to achieve the desired traffic intensity, ρ. The auxiliary phase in the overall *BMAP* can be characterized by the number of individual *MMPP*'s that are in the high-rate state. Let j_0 be the initial number. The service time distribution is assumed to be Erlang of order 16, E_{16}, with unit mean so that the time units are in mean service times. The squared coefficient of variation of this service-time distribution is 1/16.

Figures 1-3 show several transient workload and queue length distributions on log scales. In each case the stationary distribution is shown by a solid line and the transient distributions are shown by dashed or dotted lines. In all figures, we assume that $j_0 = 2$, i.e., at time 0, two sources are in the high-rate state and two are in the low-rate state.

Figure 1 shows how the transient workload distribution approaches the stationary distribution as t increases. In particular, Figure 1 displays the workload tail probabilities $P(V(10)>x|X(0)=i_0, J(0)=j_0)$ as a function of t for two values of i_0, $i_0 = 0$ and $i_0 = 32$, with $j_0 = 2$. Note that the rate of convergence to steady-state clearly depends on the initial queue length.

The transient distributions are proper for $\rho \geq 1$, as well. This is demonstrated in Figure 2 where the workload tail probabilities are displayed for several values of t when $\rho = 2.0$. For each case in this example, $i_0 = j_0 = 2$, i.e., the initial queue length is two and two *MMPP*'s start out in the high-rate state. As $t \rightarrow \infty$, $V(t) \rightarrow \infty$ w.p. 1, so that $V(t) \rightarrow V(\infty)$, where $V(\infty)$ has the degenerate distribution $P(V(\infty)>x) = 1$ for all x, as is shown by the solid line. As expected, the transient distributions approach the limiting behavior as t increases, but note however, that if the overload is limited in duration, the system performance might well be acceptable. In particular, we believe that transient solutions can shed light on the problem of overload controls.

Finally, in Figure 3, we plot the transient queue-length probability mass function with an initial queue length of 32. We note that, as expected, as t increases, the initial distribution (concentrated at a point mass at 32) gradually spreads out to approach the stationary distribution. Note the striking qualitative differences between the stationary distribution and the transient results for moderate values of t. This is further indication that predictions of system performance based on stationary analysis could be very far from what is observed during the short run.

REFERENCES

1. Lucantoni, D. M., New results for the single server queue with a batch Markovian arrival process, *Stoch. Mod.*, **7** (1991) 1-46.
2. Lucantoni, D. M., The *BMAP/G/*1 queue: A tutorial, *Models and Techniques for Performance Evaluation of Computer and Communications Systems,*` L. Donatiello and R. Nelson Editors, Springer Verlag, 1993.
3. Neuts, M. F., *Structured Stochastic Matrices of M/G/*1 *Type and Their Applications,* New York: Marcel Dekker, 1989.
4. Neuts, M. F., A versatile Markovian point process, *J. Appl. Prob.*, **16** (1979) 764-79.
5. Ramaswami, V., The *N/G/*1 queue and its detailed analysis, *Adv. Appl. Prob.*, **12** (1980) 222-61.
6. Lucantoni, D. M., Meier-Hellstern, K. S, Neuts, M. F., A single server queue with server vacations and a class of non-renewal arrival processes, *Adv. Appl. Prob.*, **22** (1990) 676-705,
7. Neuts, M. F., *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* Baltimore: The Johns Hopkins University Press, 1981.
8. Heffes, H., and Lucantoni, D. M., A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, *IEEE J. on Selected Areas in Communication,* **SAC-4** (1986) 856-868.
9. Asmussen, S., and Koole, G., Marked point processes as limits of Markovian arrival streams, *J. Appl. Prob.*, **30** (1993) 365-72.
10. Choudhury, G. L., Lucantoni, D. M., Whitt, W., Squeezing the most out of ATM, submitted for publication, 1993.
11. Takács, L., *Introduction to the Theory of Queues,* New York: Oxford University Press, 1962.
12. Abate, J, and Whitt, W., Transient behavior of the *M/G/*1 workload process, to appear in *Oper. Res.*, 1993.
13. Cohen, J. W., *The Single Server Queue,* Amsterdam: North-Holland, 1969.
14. Benes, V., *General Stochastic Processes in the Theory of Queues,* Reading, MA: Addison-Wesley, 1963.

15. Çinlar, E, The time dependence of queues with semi-Markovian service times. *J. Appl. Prob.*, **4** (1967) 356-64.

16. Çinlar, E, Queues with semi-Markovian arrivals, *J. Appl. Prob.*, 4 (1967) 365-379.

17. Choudhury, G. L., Lucantoni, D. M., Whitt, W., Multi-dimensional transform inversion with applications to the transient *M/G/*1 queue, submitted for publication.

18. Abate, J. and Whitt, W., The Fourier-series method for inverting transforms of probability distributions, *Queueing Systems,* **10** (1992) 5-88.

19. Choudhury, G. L., Lucantoni, D. M., Whitt, W., The distribution of the duration and number served during a busy period in the *BMAP/G/*1 queue, in preparation.

20. Lucantoni, D. M., Choudhury, G. L., and Whitt, W., The transient *BMAP/G/*1 queue, to appear in *Stoch. Models*, 1994.

21. Bellman, R., *Introduction to Matrix Analysis,* New York: McGraw Hill, 1960.

22. Gantmacher, F. R., *The Theory of Matrices, Vol. I,* New York: Chelsea, 1977.

23. Neuts, M. F., Moment formulas for the Markov renewal branching process. *Adv. Appl. Prob.*, **8,** (1976) 690-711.

24. Kendall, D. G., Some problems in the theory of queues, *J. Roy. Statist. Soc.*, Ser. B **13** (1951) 151-185.

25. Abate, J., and Whitt, W., Solving probability transform functional equations for numerical inversion, *OR Letters*, **12** (1992) 275-281.

26. Neuts, M. F., The single server queue with Poisson input and semi-Markov service times, *J. Appl. Prob.*, **3** (1996) 202-230.

27. Neuts, M. F., Two queues in series with a finite, intermediate waitingroom, *J. Appl. Prob.*, **5** (1968) 123-42.

28. Deutsch, G., *Introduction to the Theory and Application of the Laplace Transformation,* New York: Springer-Verlag, 1974.

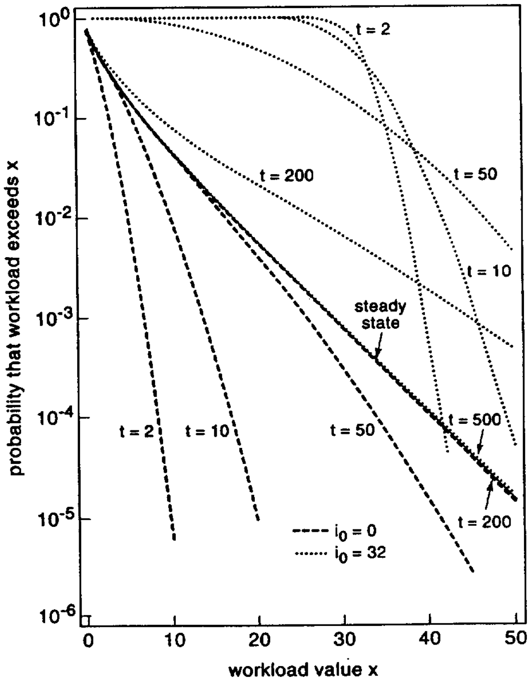

Figure 1. Numerical results for the workload tail probabilities as a function of the time t and initial queue length i_0 in the $\sum^4 MMPP_i/E_{16}/1$ queue with traffic intensity $\rho = 0.7$ and $j_0 = 2$ two of the four MMPPs starting in the high-rate state.

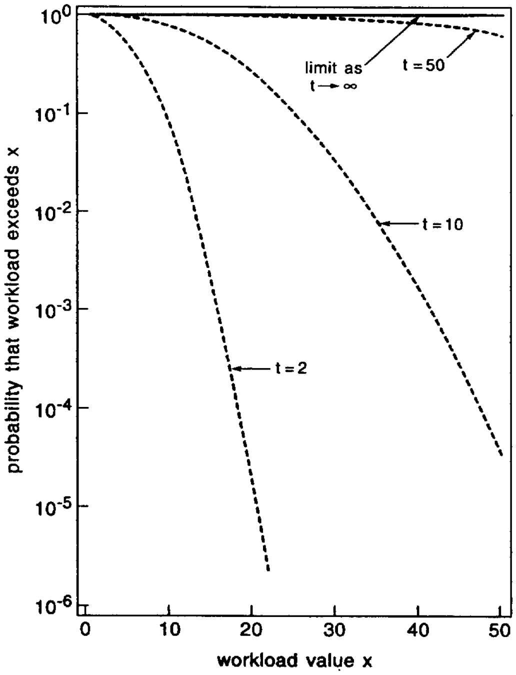

Figure 2. Numerical results for the workload tail probabilities as a function of time t in the unstable $\sum^4 MMPP_i/E_{16}/1$ model with traffic intensity $\rho = 2.0$. In each case, the initial queue length is $i_0 = 2$ and the number of the four MMPPs starting off in the high-rate state is $j_0 = 2$.

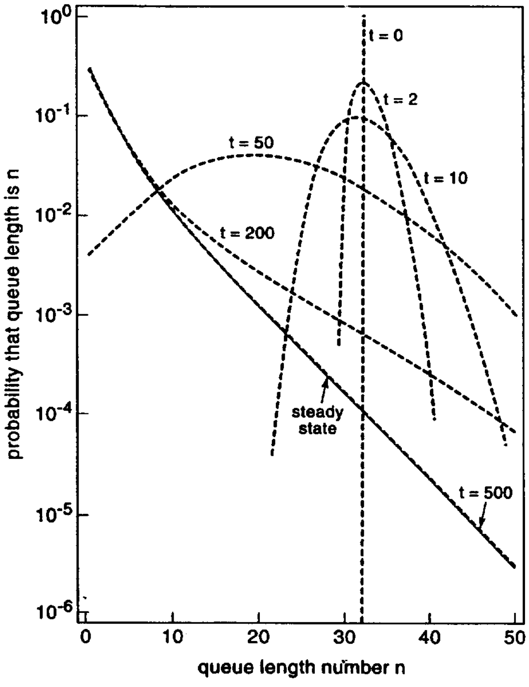

Figure 3. Numerical results for the transient queue-length probability mass function as a function of time t in the $\sum^4 MMPP_i/E_{16}/1$ model with traffic intensity $\rho = 0.7$, initial queue length $i_0 = 32$ and $j_0 = 2$ of the four MMPPs starting in the high-rate state.