COMMUN. STATIST.-STOCHASTIC MODELS, 10(1), 145-182 (1994)

THE TRANSIENT *BMAP/G/1* QUEUE

David M. Lucantoni[1]
Gagan L. Choudhury[2]
Ward Whitt[3]

AT&T Bell Laboratories

ABSTRACT

We derive 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.

KEYWORDS: transient behavior of queues, transient waiting times, *N/G/1*, busy period, emptiness function, numerical inversion of transforms, two-dimensional transform inversion

---

[1] AT&T Bell Laboratories, Room 3K-601, Holmdel, NJ 07733-3030, email: dave@buckaroo.att.com

[2] AT&T Bell Laboratories, Room 3K-603, Holmdel, NJ 07733-3030, email: gagan@buckaroo.att.com

[3] AT&T Bell Laboratories, Room 2C-178, Murray Hill, NJ 07974-0636, email: wow@research.att.com

1. Introduction

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.

In order to obtain computable results, we assume that the arrival process is a *batch Markovian arrival process* (*BMAP*), as in Lucantoni [1]. The *BMAP* is a convenient representation of the *versatile Markovian point process* (Neuts [2], [3]) or *Neuts (N) process* (Ramaswami [4]). The *BMAP* generalizes the *Markovian arrival process* (*MAP*), which was introduced by Lucantoni, Meier-Hellstern and Neuts [5]. A tutorial on the *BMAP/G/*1 queue is presented in Lucantoni [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], [11] 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 [12], Abate and Whitt [13] 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 3.1 here.

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 [2],[7], Takács [12], Cohen [14] and Benes [15], and references therein. With regard to the present work, the 1967 papers by Çinlar [16], [17] and the early papers of Neuts (cited in [2]) 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 [18]. These algorithms are based on the Fourier-series method [19], and are enhancements and generalizations of the Euler and Lattice-Poisson algorithms described in [19]. For this purpose, we evaluate the busy-period transform by iterating the characterizing functional equation, drawing upon Choudhury, Lucantoni and Whitt [20].

Further transient results for the $BMAP/G/1$ queue have recently been derived and will be reported in Lucantoni [21]. In particular, we derived explicit expressions for the transform of the queue length at the n-th departure, assuming a departure at time $t=0$, and the workload at the n-th arrival (keeping track of the appropriate phase changes). The departure process is characterized by the double transform of the probability that the n-th departure occurs at time less than or equal to time x (similar to that derived by Saito [22]). This leads to an explicit expression for the *LST* of the expected number of departures up to time t. All of these expressions are direct matrix analogues of the corresponding $M/G/1$ results in Takács [12].

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. The algorithm for inverting multidimensional Laplace transforms is presented in §6 and several details regarding the implementation of the algorithm presented in §7. This algorithm is then used for computing the numerical examples in §8. All of the proofs are presented in §9.

2. The $BMAP/G/1$ Queue

2.1 The Batch Markovian Arrival Process

The *BMAP* is a natural generalization of the Poisson process (see Lucantoni [1],[6]). 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 [23]), 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

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

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

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

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

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] and [6].

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

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

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

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

is given explicitly by

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

where $e^{D(z)t}$ is an exponential matrix (see e.g., pg. 169 of Bellman, [23]). Note that for Poisson arrivals, $m=1$, $D_0 = -\lambda$, $D_1 = \lambda$, and $D_k = 0$, $k \ge 2$, so that (5) reduces to $P^*(z,t) = e^{-\lambda(1-z)t}$ which is the familiar generating function of the Poisson counting process.

2.2 The Queueing Model

Consider a single-server queue with a *BMAP* arrival process specified by the sequence $\{D_k,\ k \ge 0\}$. 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*, $\rho \equiv \lambda\alpha$.

2.3 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 $(X(\tau_k),\, J(\tau_k),\, \tau_{k+1} - \tau_k)$, for $k \ge 0$, is a semi-Markov process on the state space $\{\, (i,j) : i \ge 0,\ 1 \le j \le m \,\}$. 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 [2]) since it has matrix generalizations of the skip-free-to-the-left and spatial homogeneity properties of the ordinary $M/G/1$ queue.

Following the treatment of the $M/G/1$ queue in §1.5 of [2] we define the Markov renewal function $M(t)$ whose elements $M_{(i,j),(k,l)}(t)$ are the expected number of visits to state (k,l) in $[0,t]$, starting in state (i,j) at time $t=0$. Let $U(t)$ be the infinite diagonal matrix where the diagonal entries are 0 for $t<0$ and 1 for $t \geq 0$. Then the matrix $M(t)$ satisfies

$$M(t) = U(t) + (Q*M)(t) = U(t) + (M*Q)(t), \quad \text{for} \quad t \geq 0, \tag{7}$$

where $*$ denotes matrix convolution. For background on Markov renewal processes, see Chapter 10 of Çinlar [24].

It is well known that $M(t)$ is given by the Neumann series

$$M(t) = \sum_{n=0}^{\infty} Q^{(n)}(t),$$

where $Q^{(n)}(t)$, $n \geq 1$, is the n-fold matrix convolution of $Q(\cdot)$ with itself and

$Q^{(0)}(\cdot) \equiv U(\cdot)$. If we partition $M(t)$ and $U(t)$ into $m\times m$ blocks then we see from (6) and (7) that the blocks satisfy

$$M_{i_0 i}(t) = \delta_{i_0 i} U_{ii}(t) + (M_{i_0 0} * \hat{B}_i)(t) + \sum_{n=1}^{i+1} (M_{i_0 n} * \hat{A}_{i-n+1})(t), \tag{8}$$

where $\delta_{ii} = 1$, and $\delta_{ij} = 0$ for $i \neq j$.

We introduce the transform matrices

$$\begin{aligned} A_n(s) &= \int_0^\infty e^{-sx}\, d\hat{A}_n(x), & \tilde{A}(z,s) &= \sum_{n=0}^\infty A_n(s) z^n, \\ B_n(s) &= \int_0^\infty e^{-sx} d\hat{B}_n(x), & \tilde{B}(z,s) &= \sum_{n=0}^\infty B_n(s) z^n, \\ m_{i_0 i}(s) &= \int_0^\infty e^{-st} dM_{i_0 i}(t), & \tilde{m}_{i_0}(z,s) &= \sum_{i=0}^\infty m_{i_0 i}(s) z^i, \end{aligned} \tag{9}$$

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

$$\tilde{A}(z,s) = \int_0^\infty e^{-sx} e^{D(z)x} dH(x) \equiv h(sI - D(z)), \tag{10}$$

and

$$\tilde{B}(z,s) = z^{-1}[sI - D_0]^{-1}[D(z) - D_0]\tilde{A}(z,s). \tag{11}$$

The definition in (10) above is consistent with the usual definition of a scalar function evaluated at a matrix argument (see Theorem 2, pg. 113 of Gantmacher, [25]). 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 (10) we see that $\tilde{A}(z,s)$ is a power series in $D(z)$. Thus, $\tilde{A}(z,s)$ and $D(z)$ commute. This property is used repeatedly in the proofs.

Using (8)-(11), we have

$$\begin{aligned} \tilde{m}_{i_0}(z,s)[zI - \tilde{A}(z,s)] &= z^{i_0+1} I + m_{i_0 0}(s)[z\tilde{B}(z,s) - \tilde{A}(z,s)] \\ &= z^{i_0+1} I + m_{i_0 0}(s)(sI - D_0)^{-1}[D(z) - sI]\tilde{A}(z,s), \end{aligned} \tag{12}$$

since

$$(sI-D_0)^{-1}[D(z)-D_0] - I = (sI-D_0)^{-1}[D(z)-D_0-(sI-D_0)]$$
$$= (sI-D_0)^{-1}[D(z)-sI].$$

2.4 The Busy Period

Following the general treatment of Markov chains of $M/G/1$-type in [2], 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 $\boldsymbol{i}$ where level $\boldsymbol{i} \equiv \{(i,j);1 \leq j \leq m\}$ for $i \geq 0$. The matrix with elements $\hat{G}_{jj'}{}^{[r]}(x)$ is $\hat{G}^{[r]}(x)$.

By a first passage argument, it was shown in Neuts [26] 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

$$G(s) = \sum_{n=0}^{\infty} A_n(s) G(s)^n. \tag{13}$$

In the context of the $BMAP/G/1$ queue, $G(s)$ governs the duration of the busy period. It was also shown in [26] that the transform matrix governing the duration of a busy period starting with r customers, is given by $G(s)^r$. Equation (13) 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

$$G(s) = \int_0^\infty e^{-sx} e^{D[G(s)]x} dH(x) \equiv h(sI-D[G(s)]), \tag{14}$$

where $D[G(s)] \equiv \sum_{k=0}^{\infty} D_k G(s)^k$. Equation (14) is the matrix analogue of the *Kendall functional equation,* (see (59) in Kendall [27], 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 (14) reduces to $G(s) = h(s+\lambda-\lambda G(s))$ which is (59) in [27].

The matrix $D[G]$, where $G \equiv G(0)$ has a nice probabilistic interpretation which was originally pointed out in Lucantoni, Meier-Hellstern and Neuts [5]. Since G is strictly positive, it follows that the off-diagonal entries of $D[G]$ are nonnegative. When the queue is stable, G is stochastic so that $D[G]\boldsymbol{e} = \mathbf{0}$; that is, $D[G]$ is the infinitesimal generator of a finite-state, irreducible Markov process. From the structure of the matrix we see that starting in some state i, there will be an exponential sojourn time with rate $|(D_0)_{ii}|$. Then there will either be a transition to state j, with rate $(D_0)_{ij}$ (i.e., without an arrival), or a transition to state j with rate $(\sum_{k=1}^{\infty} D_k G^k)_{ij}$. That is, a batch of size k arrives followed by k busy periods which end in phase j, corresponding to an instantaneous phase change from i to j in this process. It is clear that this process is the phase of the arrival process observed only during idle periods, i.e., the time during the busy periods are *excised*. In the unstable case, i.e., $\rho > 1$, G is strictly substochastic so that $D[G]$ is a stable matrix. In other words, in this case the total amount of idle time observed before the final busy period (that never ends) is *phase* − *type* with representation $(\boldsymbol{a}, D[G])$, where $\boldsymbol{a}$ is the vector of initial phase probabilities at time 0; (see, e.g., [7]).

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 [20] that $G(s)$ may be computed by iterating in (14). If $\rho < 1$, convergence is guaranteed if the iteration is started with either $G_0 \equiv 0$ or $G_0 \equiv G$ 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 [28]. If $\rho \geq 1$, then $G(s)$ is evaluated by iterating in (14) starting with $G_0 \equiv 0$.

In order to compute the right hand side of (14) in each iteration, two cases are considered in [20]. 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 (see §7). If the service time distribution is not rational, then a procedure similar to uniformization is used.

2.5 Simplifications for the $M^X/G/1$ Queue

We end this and the next three sections by displaying the main results for the special case in which there are batch Poisson arrivals. In this case all the matrix equations reduce to scalar equations. In particular, if the arrival rate of batches is δ and the batch-size probability mass function is $\{\gamma_n, n \geq 1\}$, with probability generating function $\gamma(z)$ and mean $\bar{\gamma}$, then $\lambda = \delta\bar{\gamma}$ and $D(z) = -\delta + \delta\gamma(z)$. Therefore, from (14), we have

$$G(s) = h(s+\delta-\delta\gamma(G(s))), \quad \mathrm{Re}(s) \geq 0. \tag{15}$$

If the batch size distribution is identically equal to 1 then $\gamma(z) = z$, $\lambda = \delta$, and these results agree with those in [12] for the ordinary $M/G/1$ queue.

3. Preliminary Results

3.1 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 [15]. 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 $\mathrm{Re}(s) > 0$. Then we have the following generalization of the $M/G/1$ formula. (See (9) on pg. 52 of [12] and (34) and (36) in [13]).

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

$$p_{x0}(s) = e^{-(sI - D[G(s)])x}(sI - D[G(s)])^{-1}, \tag{16}$$

for $\mathrm{Re}(s) > 0$.

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 $\mathrm{Re}(s) > 0$, the eigenvalues of $D[G(s)]$ are in the left half-plane. Therefore, for $\mathrm{Re}(s) > 0$, the eigenvalues of $sI - D[G(s)]$ are in the right half-plane and the inverse appearing in (16) is well defined.

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

$$\hat{P}_{i_0 0}^{jk}(t) = P(\ V(t)=0,\ J(t)=j \mid X(0)=i_0,\ J(0)=j\ ,\ \tau_0=0\ ). \tag{17}$$

As a consequence of Theorem 1, we immediately have

$$\hat{p}_{i_0 0}(s) \equiv \int_0^\infty e^{-st}\hat{P}_{i_0 0}(t)\,dt = G(s)^{i_0} p_{00}(s) = G(s)^{i_0}(sI - D[G(s)])^{-1}. \tag{18}$$

For later use, we note that, by conditioning on the last departure before time t, we can write

$$\hat{P}_{i_0 0}(t) = \int_0^t dM_{i_0 0}(u) e^{D_0(t-u)}.$$

Taking Laplace transforms leads to

$$\hat{p}_{i_0 0}(s) = m_{i_0 0}(s)(sI - D_0)^{-1}. \tag{19}$$

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

$$P_0(t) \equiv \int_0^\infty P_{x0}(t)\, dF(x), \tag{20}$$

has Laplace transform

$$p_0(s) \equiv \int_0^\infty e^{-st} P_0(t)\, dt = f(sI - D[G(s)])(sI - D[G(s)])^{-1} \tag{21}$$

where f is the *LST* of F.

We now apply (21) to derive known steady state results. Recall that $\lim_{t\to\infty} P_0(t) = \lim_{s\to 0} sp_0(s)$. Let $R(s) = sp_0(s)$. Multiplying both sides of (20) by $s(sI - D[G(s)])$, we have

$$\begin{aligned} sR(s) - R(s)D[G(s)] &= sR(s) - D[G(s)]R(s) \\ &= sf(sI - D[G(s)]), \end{aligned} \tag{22}$$

since $(sI - D[G(s)])$ commutes with $f(sI - D[G(s)])$ as seen by expanding f in a power series. Letting $s \to 0$ in (22), we have $R(0)D[G] = D[G]R(0) = 0$. Therefore the columns of $R(0)$ are right eigenvectors of $D[G]$ corresponding to the eigenvalue 0. Similarly, the rows of $R(0)$ are left eigenvectors of $D[G]$ corresponding to the eigenvalue 0. Since $D[G]$ is the infinitesimal generator of an irreducible, finite state Markov process, its left and right eigenvectors corresponding to the eigenvalue 0 are unique up to a scalar constant by application of the Perron-Frobenious theorem; see e.g., Theorem 2, pg. 53 of Gantmacher, [29], Since it was shown in [1] that $\mathbf{g}D[G] = \mathbf{0}$, we have $R(0) = c\mathbf{e}\mathbf{g}$, for some constant c. However, since we know that for any $G/G/1$ queue with $\rho \leq 1$ the stationary probability that the system is empty is $1-\rho$, we have

$$\lim_{t\to\infty} P_0(t) = \begin{cases} (1-\rho)\,\boldsymbol{eg} & \text{for } \rho \le 1, \\ 0 & \text{for } \rho > 1. \end{cases} \tag{23}$$

3.2 Simplifications for the $M^X/G/1$ Queue

For the $M^X/G/1$ queue with the notation in §2.5, we have

$$p_{x0}(s) = \frac{e^{-(s+\delta-\delta\gamma(G(s)))x}}{s+\delta-\delta\gamma(G(s))}, \quad m_{x0}(s) = (s+\lambda)p_{x0}(s), \tag{24}$$

$$p_0(s) = \frac{f(s+\delta-\delta\gamma(G(s)))}{s+\delta-\delta\gamma(G(s))}, \quad \hat{p}_{i_0 0}(s) = \frac{G(s)^{i_0}}{s+\delta-\delta\gamma(G(s))}, \tag{25}$$

for Re $(s)>0$.

4. The Workload

4.1 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 zero 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

$$w_{i_0}(t,s) \equiv \int_0^\infty e^{-sx} d_x W_{i_0}(t,x), \quad \text{and} \quad \tilde{w}_{i_0}(\xi,s) \equiv \int_0^\infty e^{-\xi t} w_{i_0}(t,s)\,dt,$$

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 $\tilde{w}_{i_0}(\xi,s)$ is given explicitly by

$$\tilde{w}_{i_0}(\xi,s) = (h(s)^{i_0} I - s\hat{p}_{i_0 0}(\xi))[\xi I - sI - D(h(s))]^{-1}, \qquad (26)$$

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

$$w_{i_0}(t,s) = \left[h(s)^{i_0} I - s\int_0^t \hat{P}_{i_0 0}(u) e^{-(sI+D(h(s)))u} du \right] e^{(sI+D(h(s)))t}, \quad (27)$$

where Re $(s) \geq 0$, Re $(\xi) > 0$, and $\hat{P}_{i_0 0}(u)$ is defined in (17).

Although we are able to express the transform of the delay explicitly in terms of t in (27), 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 (26) 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 §6.

It can be shown using Rouché's theorem that for each s, Re$(s) \geq 0$, the determinant of the matrix $X(s,\xi) \equiv [\xi I - sI - D(h(s))]$ appearing in the inverse in (26) has exactly m roots in the region Re$(\xi) > 0$. (For similar arguments see Çinlar [16] and Neuts [30], [31].) Since $\tilde{w}_{i_0}$ is a transform and is therefore analytic in the interior of the above region, see p.26 of Deutsch [32], these pairs of (ξ,s) must also be zeros of the first matrix on the right in (26). 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 (26).

As a consequence of Theorem 2, we can easily treat the workload at time t given that a departure occurred at time 0 and a random number of customers are present. Let $\tilde{w}^*(\xi,s)$, be the double Laplace transform

$$\tilde{w}^*(\xi,s) = \sum_{i=0}^{\infty} \phi_i \tilde{w}_i(\xi,s),$$

where $\{\phi_i\}$ is the probability mass function of the number of customers in the system after the departure at $t=0$. If $\Phi(z)$ is the probability generating function of $\{\phi_i\}$, then

$$\tilde{w}^*(\xi,s) = \left[\Phi(h(s))I - s\Phi[G(\xi)]\left(\xi I - D[G(\xi)]\right)^{-1}\right][\xi I - sI - D(h(s))]^{-1},$$

where Re $(s) \ge 0$ and Re $(\xi) > 0$.

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 $\tilde{w}(\xi,s)$ be the Laplace transforms

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

Then we have the following theorem.

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

$$\tilde{w}(\xi,s) = (f(s)I - sp_0(\xi))[\xi I - sI - D(h(s))]^{-1}, \tag{28}$$

and

$$w(t,s) = \left[f(s)I - s\int_0^t P_0(u) e^{-[sI + D(h(s))]u} du\right] \times e^{[sI + D(h(s))]t}, \tag{29}$$

for Re$(s) \ge 0$), Re$(\xi) > 0$, where $P_0(u)$ and $p_0(\xi)$ are given in (20) and (21), respectively.

Note that Theorem 2 is a special case of Theorem 3 where $f(s) = h(s)^{i_0}$. Note that (29) is the direct analogue of Equation (8) on pg. 51 of [12].

4.2 The Limiting Distribution of the Waiting Time

Differentiating with respect to t in (29), we have

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

Therefore, using (23) 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

$$w(s) \equiv \lim_{t \to \infty} w(t,s) = \begin{cases} s(1-\rho)\boldsymbol{eg}[sI + D(h(s))]^{-1}, & \text{for } \rho < 1, \\ 0, & \text{for } \rho \geq 1, \end{cases}$$

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

4.3 The First Moment Function

Let the first moment function be the $m \times m$ matrix

$$m_1(t) \equiv -\frac{\partial}{\partial s} w(t,s)\bigg|_{s=0},$$

where the (i,j) component is $E[\ V(t)\ I_{\{J(t)=j\}} \mid J(0) = i\ ]$ with I_A being the indicator function of the set A. Let $\beta \equiv -f'(0)$ be the expected work in the system at time $t=0$, and let $D^{(1)} = \sum_{k=1}^{\infty} kD_k$. Recall from (4) that $\boldsymbol{d} = D^{(1)}\boldsymbol{e}$. Then we have the following theorem.

Theorem 4: Assume that $\alpha < \infty$ and $\beta < \infty$. Then the matrix $m_1(t)$ is explicitly given by

$$m_1(t) = \alpha \int_0^t e^{Du} D^{(1)} e^{D(t-u)}\, du + (\beta - t) e^{Dt} + \int_0^t P_0(u) e^{D(t-u)}\, du. \quad (30)$$

Equivalently, $m_1(t)$ satisfies the following differential equation

$$m_1'(t) = \alpha e^{Dt} D^{(1)} - e^{Dt} + P_0(t) + m_1(t) D, \quad (31)$$

with $m_1(0) = \beta I$. The row sums of $m_1(t)$, i.e., $m_1(t)\boldsymbol{e}$, satisfy the differential equations

$$m_1'(t)\boldsymbol{e} = \alpha e^{Dt}\boldsymbol{d} - \boldsymbol{e} + P_0(t)\boldsymbol{e}, \qquad m_1(0)\boldsymbol{e} = \beta\boldsymbol{e}. \quad (32)$$

Note that the expression for $m_1(t)$ in (31) Is more complex than the corresponding $M/G/1$ case since the matrices D and $D^{(1)}$ do not commute in general. Assuming that $m_1'(t) \to 0$ as $t \to \infty$, if we solve for $m_1(t)$ in (31) and let $t \to \infty$ we obtain expression (47) of [1] for the mean workload in the stationary $BMAP/G/1$ queue. Note that (31) is the matrix analogue of Equation (20) on pg. 55 of [12]. In the batch-Poisson case, the matrices become scalars and $D=0$, so that the last term in (31) does not appear.

If we pick the initial phase of the arrival process at time $t=0$ according to the stationary distribution $\boldsymbol{\pi}$, then we have

$$\boldsymbol{\pi} m_1'(t)\boldsymbol{e} = \rho - 1 + \boldsymbol{\pi} P_0(t)\boldsymbol{e}, \tag{33}$$

which is a generalization of Equation (17) in Abate and Whitt [13]. See Miyazawa [33] for more results related to (33).

4.4 Higher-Order Moment Functions

Along the lines of Abate and Whitt [13], we can derive differential equations for the higher order moments of the delay at time t. In particular, let $V(s) = D(h(s))$ and let the i^{th} derivatives be $V^{(i)} \equiv (-1)^i V^{(i)}(0)$, and $D^{(i)} = D^{(i)}(1)$ for $i \geq 1$. Then, by successively differentiating $V(s)$, we get

$$\begin{aligned} V^{(1)} &= \alpha D^{(1)}, \\ V^{(2)} &= \alpha^2 D^{(2)} + \alpha_2 D^{(1)}, \\ V^{(3)} &= \alpha^3 D^{(3)} + 3\alpha\alpha_2 D^{(2)} + \alpha_3 D^{(1)}, \end{aligned}$$

etc. The expression for the n^{th} moment can be obtained by Fáa di Bruno's formula for the n^{th} derivative of a composite function, e.g., see p.36 of Riordan [34], Ch.5 of Riordan [35], and Klimko and Neuts [36]. Let the k^{th} moment function be defined by

$$m_k(t) \equiv (-1)^k \frac{\partial^k}{\partial s^k} W(t,s) \bigg|_{s=0}, \tag{34}$$

and let β_k be the k^{th} moment of the workload at time 0.

Theorem 5: If $\alpha_k < \infty$ and $\beta_k < \infty$, then the k^{th} moment function in (34) can be expressed as

$$\begin{aligned} m_k(t) = &-k \int_0^t m_{k-1}(u) e^{D(t-u)}\, du \\ &+ \sum_{j=0}^{k-1} \binom{k}{j} \int_0^t m_j(u)\, V^{(k-j)}\, e^{D(t-u)}\, du + \beta_k e^{Dt}. \end{aligned} \tag{35}$$

Equivalently, $m_k(t)$ satisfies the system of differential equations

$$m'_k(t) = -km_{k-1}(t) + \sum_{j=0}^{k-1} \binom{k}{j} m_j(t) V^{(k-j)} + m_k(t) D, \tag{36}$$

and $m_k(0) = \beta_k I$.

Once again, simpler equations result if we are only interested in the marginal moments, i.e., the row sums of $m_k(t)$. Equation (36) is the matrix analogue of (19) in [13]. Also, assuming that $m'_k(t) \to 0$ as $t \to \infty$, if we solve (36) for $m_k(t)$ and let $t \to \infty$, we obtain expressions (A.1.3) and (A.1.4) in Lucantoni and Neuts [37] for the n^{th} moments of the workload in the stationary version of the $BMAP/G/1$ queue.

4.5 Simplifications for the $M^X/G/1$ Queue

For the $M^X/G/1$ queue, with the notation in §2.5, we have

$$\tilde{w}_{i_0}(\xi,s) = \frac{h(s)^{i_0} - s\hat{p}_{i_00}(\xi)}{\xi - s + \delta - \delta\gamma(h(s))}, \tag{37}$$

$$w_{i_0}(t,s) = e^{(s-\delta+\delta\gamma(h(s)))t} \left[h(s)^{i_0} - s\int_0^t \hat{P}_{i_00}(u) e^{-(s-\delta+\delta\gamma(h(s)))u} du \right], \tag{38}$$

$$\tilde{w}(\xi,s) = \frac{f(s) - sp_0(\xi)}{\xi - s + \delta - \delta\gamma(h(s))}, \tag{39}$$

$$w(t,s) = e^{(s-\delta+\delta\gamma(h(s)))t} \left[f(s) - s\int_0^t P_0(u) e^{-(s-\delta+\delta\gamma(h(s)))u} du \right], \tag{40}$$

where $\mathrm{Re}(s) \geq 0$, $\mathrm{Re}(\xi) > 0$ and $\hat{p}_{i_00}(\xi)$ and $p_0(\xi)$ are given in (17) and (20), respectively. Note that (39) and (40) generalize (15) on p.53 and (8) on p.51 of [12] to batch arrivals, respectively.

5. The Queue Length

5.1 The Transient Results

In this section, we compute the transient queue length distribution at time t given an initial number of customers present immediately after a departure at time $t=0$. Let $Y^{jk}_{i_0i}(t) = P(\ X(t)=i,\ J(t)=k \mid X(0)=i_0,\ J(0)=j,\ \tau_0=0\ )$, and let

$Y_{i_0 i}(t)$ have (j,k)-entry $Y_{i_0 i}^{jk}(t)$. 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

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

Theorem 6: The matrix $\tilde{y}_{i_0}(z,s)$ is given by

$$\tilde{y}_{i_0}(z,s) = \Big[z^{i_0+1}(I - \tilde{A}(z,s))(sI - D(z))^{-1} + (z-1)\hat{p}_{i_0 0}(s)\tilde{A}(z,s) \Big] [zI - \tilde{A}(z,s)]^{-1}, \tag{41}$$

for Re$(s)>0$ and $|z|<1$, where $\hat{p}_{i_0 0}(s)$ is given in (17) and $\tilde{A}(z,s)$ is given in (10).

Equation (41) is the matrix analogue of Equation (77) on pg. 74 in [12]. 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

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

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

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

$$\tilde{y}^*_{i_0}(z,s) = \frac{1}{1-z}\Big[(sI-D)^{-1} - \tilde{y}_{i_0}(z,s) \Big]. \tag{42}$$

5.2 Simplifications for the $M^X/G/1$ Queue

For the $M^X/G/1$ queue, with the notation in §2.5, we have

$$\tilde{y}_{i_0}(z,s) = \frac{z^{i_0+1}(1-\tilde{A}(z,s))}{(s+\delta-\delta\gamma(z))(z-\tilde{A}(z,s))} + \frac{(z-1)\hat{p}_{i_0 0}(s)\tilde{A}(z,s)}{z-\tilde{A}(z,s)}, \quad (43)$$

where $\mathrm{Re}(s)>0$, $|z|<1$ and $\tilde{A}(z,s) = h(s+\delta-\delta\gamma(z))$, and $\hat{p}_{i_0 0}(s)$ is given in Equation (24).

6. An Algorithm for Inverting Two-Dimensional Laplace Transforms

Recently, we have developed effective algorithms for multi-dimensional transform inversions Choudhury, Lucantoni, Whitt [18]. The algorithms are based on the multi-dimensional Poisson summation formula (of continuous, discrete, and mixed variety, see e.g., (5.47) of [19] for the continuous variety), and are generalizations of the *EULER* and *Lattice − Poisson* algorithms presented in [19]. We briefly describe the algorithms used here and refer to [18] for further discussion.

Let $F(t_1,t_2)$ represent a function of two non-negative real variables with Laplace transform

$$\tilde{f}(s_1,s_2) = \int_0^\infty\int_0^\infty F(t_1,t_2)e^{-(s_1t_1+s_2t_2)}\,dt_1\,dt_2. \quad (44)$$

The values of F may be obtained by inverting the transform via

$$F(t_1,t_2) = \frac{\exp(A_1/2l_1)}{2l_1t_1}\times$$

$$\sum_{j_1=1}^{l_1}\sum_{j=-\infty}^{\infty}(-1)^j e^{-\frac{ij_1\pi}{l_1}}\left\{\frac{\exp(A_2/2l_2)}{2l_2t_2}\sum_{k_1=1}^{l_2}\sum_{k=-\infty}^{\infty}(-1)^k\frac{e^{-ik_1\pi}}{l_2}\right.$$

$$\left.\times\tilde{f}\left[\frac{A_1}{2l_1t_1}-\frac{ij_1\pi}{l_1t_1}-\frac{ij\pi}{t_1},\ \frac{A_2}{2l_2t_2}-\frac{ik_1\pi}{l_2t_2}-\frac{ik\pi}{t_2}\right]\right\} - e_d. \quad (45)$$

where e_d is the error term, $i\equiv\sqrt{-1}$, and for j=1 and 2, A_j is a real constant and l_j is an integer constant (discussed below). When $|F(t_1,t_2)| \le 1$ for all t_1,t_2, as when $F(t_1,t_2)$ represents a probability, the error term e_d is bounded as follows:

$$|e_d| \le \frac{e^{-A_1}+e^{-A_2}-e^{-(A_1+A_2)}}{(1-e^{-A_1})(1-e^{-A_2})} \approx e^{-A_1}+e^{-A_2}. \quad (46)$$

The constants A_1 and A_2 are chosen appropriately to control the error term. For example, choosing $A_1 = A_2 = 23.7$ ensures that $|e_d| \leq 10^{-10}$. There is also a round-off error associated with computing (45) which may be controlled by increasing l_1 and l_2. Typically the choice of $l_1 = l_2 = 2$ is adequate.

Now let $F(t,n)$ be a function of two nonnegative variables where t is continuous and n is an integer. Its two-dimensional Laplace-z transform is defined by

$$\tilde{f}(s,z) = \sum_{n=0}^{\infty} \left[\int_0^{\infty} F(t,n) e^{-st} dt \right] z^n .$$

The inversion formula is given by

$$F(t,n) = \frac{e^{\left(\frac{A}{2l_1}\right)}}{2l_1 t} \sum_{j_1=1}^{l_1} \sum_{j=-\infty}^{\infty} (-1)^j e^{-\frac{ij_1\pi}{l_1}} \left\{ \frac{1}{2l_2 n r^n} \sum_{k_1=0}^{l_2-1} \sum_{k=-n}^{n-1} (-1)^k e^{-\frac{ik_1\pi}{l_2}} \right.$$

$$\left. \times \tilde{f}\left[\frac{A}{2l_1 t} - \frac{ij_1\pi}{l_1 t} - \frac{ij\pi}{t} , \; r e^{\frac{\pi i(k_1 + l_2 k)}{l_2 n}} \right] \right\} - \tilde{e}_d , \tag{47}$$

where $\tilde{e}_d$ is the error term, A and r are real constants and l_1, l_2 are integer constants. If $|F(t,n)| \leq 1$ for all t,n, as when $F(t,n)$ represents a probability, then we have

$$|\tilde{e}_d| \leq \frac{e^{-A} + r^m - e^{-A} r^m}{(1-e^{-A})(1-r^m)} \approx e^{-A} + r^m . \tag{48}$$

A and r are chosen to control $|\tilde{e}_d|$ and l_1, l_2 are chosen to control the round-off error.

Equations (45) and (47) hold even when $F(t_1,t_2)$ and $F(t,n)$ are complex. In this paper, however, they are always real and in that case it is possible to reduce the amount of computation by a factor of 2; see [18] for details).

Equations (45) and (47) contain infinite sums. Straightforward computation of those sums by truncation may in general require the computation of a large number of terms. However, since each infinite sum is nearly an alternating series, the sums are efficiently computed via the Euler summation technique using finite differences; see §6 of [19]. In particular, (see pg. 230 of Davis and Rabinowitz [38]), we have

$$\sum_{i=0}^{\infty}(-1)^i u_i = \sum_{i=0}^{n-1}(-1)^i u_i + (-1)^n \left[\tfrac{1}{2}u_n - \tfrac{1}{4}\Delta u_n \tfrac{1}{8}\Delta^2 u_n - \cdots\right], \quad (49)$$

where $\Delta u_n = u_{n+1} - u_n$, $\Delta^2 = \Delta(\Delta u_n) = u_{n+2} - 2u_{n+1} + u_n$, etc. In many cases, the series on the right hand side of (49) converges much more rapidly than the series on the left. Our experience shows that each infinite sum may be computed accurately by evaluating only about 50 terms for most cases of interest.

The inversion formula (45) will lose accuracy when t_1 and t_2 are very small. However, since $W_{i_0}(t,0)$ and $W(t,0)$ are given by $\hat{P}_{i_0 0}(t)$ and $P_0(t)$, respectively, these values are obtained by one-dimensional transform inversions of (17) and (20), respectively.

7. Implementation Details

Application of the inversion formulas (45) and (47) to the transient workload ((25) and (28)) and the transient queue length ((41) and (42)), respectively, requires numerous evaluations of the matrix $G(s)$ for complex s. It was shown in Choudhury, Lucantoni, and Whitt [20] that $G(s)$ can be computed by successively iterating in (14). (See the discussion below for an efficient method of evaluating the integral in (14) for a wide class of service time distributions). It was also shown in [20] that, when *rho* < 1, starting the iterations with $G_0(s) \equiv 0$ and $G_0(s) \equiv G$, respectively, produce the sequences of iterates, $\underline{G}_k(s)$ and $\overline{G}_k(s)$. The matrices $\underline{G}_k(s)$ and $\overline{G}_k(s)$ are themselves Laplace-Stieltjes transforms of distribution functions $\underline{F}_k(x)$ and $\overline{F}_k(x)$ where these distributions bound the true distribution in the sense that

$$\underline{F}_k(x) \le \hat{G}(x) \le \overline{F}_k(x),$$

where $\hat{G}(x)$ is the distribution function with *LST* $G(s)$. Thus, stopping the iteration at any point will give useful bounds on the true distributions, however, we usually carry out the computations until the bounds are within 10^{-12}. When $\rho \ge 1$, the iteration must be started with $G_0(s) \equiv 0$. It is clear that the numerous evaluations of $G(s)$ may constitute a significant component of the total computational cost of evaluating the transient distributions. In the examples we have run so far, in order to get successive iterates to within 10^{-13}, elementwise, the number of iterations has ranged from several tens to a few hundreds.

In this section, we show how to organize the computations in such a way as to compute the transient distributions for several values of time with a minimum number of evaluations of $G(s)$. Note that $G(s)$ enters into (25), (28) and (41) only through the transforms of the emptiness functions given in (17) and (20). In each case, $G(\cdot)$ needs to be evaluated at ξ, where ξ is the Laplace transform variable with respect to time. Note from (45) and (47), that the double transforms need only be computed at $\xi = \delta - j\gamma$, where for a fixed t, δ and γ are constants and j is an integer which typically ranges from 0 to 50 to achieve good accuracy in most cases.

Therefore, for each time-point, t, we compute and store $G(\delta - j\gamma)$ for j=0,1,...,50. Next, we evaluate the double transform inversion formulas for as many workload and queue length values as we desire, using the stored values of $G(\xi)$ as needed. This procedure is repeated for each value of t.

We illustrate the savings gained by computing and storing $G(\cdot)$ by a simple example. Suppose we are interested in computing the workload distributions for five different time values and we would like twenty points of the distribution in each case. The pre-computation and storage of $G(\cdot)$ would require its computation at (5)(50)=250 different values. By contrast, if we directly apply the inversion formulas (45) and (47) and compute $G(\cdot)$ whenever needed, this same example would require (5)(20)(50)(50)=250,000 evaluations. Furthermore, if we also wanted the transient queue length distributions but at the same time-points, we could use the existing stored values of $G(\cdot)$ with no further evaluations required.

In several steps of the general algorithm, we need to compute the integrals of exponential matrices of the form

$$M' = \int_0^\infty e^{-Mx} dH(x), \tag{50}$$

where M and M' are, in general, complex matrices; see, e.g., $\tilde{A}(z,s)$ and $G(s)$ in (10) and (14), respectively. When M is real, efficient algorithms based on the concept of uniformization have been proposed; see (61) and (62) in [1] for computing G. Here we present an efficient algorithm for the case of complex M when H is Coxian (i.e., H has a rational Laplace-Stieltjes transform).

Theorem 7: If H has a Coxian distribution with the Laplace-Stieltjes transform

$$h(s) = \int_0^\infty e^{-sx} dH(x) = \frac{\sum_{k=0}^{i} a_k s^k}{\sum_{k=0}^{j} b_k s^k}, \qquad i \le j, \tag{51}$$

then (50) may be evaluated as

$$M' = \left(\sum_{k=0}^{j} b_k M^k\right)^{-1}\left(\sum_{k=0}^{i} a_k M^k\right). \tag{52}$$

Proof: Expanding e^{-sx} in (51) in a power series and interchanging the integral and summation, we have

$$h(s) = \sum_{k=0}^{\infty} c_k s^k, \tag{53}$$

where $c_k = (-1)^k \int_0^{\infty} \frac{x^k}{k!} dH(x)$. From (51), we get

$$\sum_{n=0}^{j} b_n s^n \sum_{k=0}^{\infty} c_k s^k = \sum_{k=0}^{i} a_k s^k. \tag{54}$$

This equality is valid for all s in the positive half-plane and remains valid if s is replaced by a complex matrix, M, provided that all the eigenvalues of M are also in the positive half-plane. Therefore, we have

$$\sum_{n=0}^{j} b_n M^n \sum_{k=0}^{\infty} c_k M^k = \sum_{k=0}^{i} a_k M^k, \tag{55}$$

and the proposition follows by noting that

$$\sum_{k=0}^{\infty} c_k M^k = \int_0^{\infty} e^{-Mx} dH(x).$$

Note that (52) only requires the computation of two matrix polynomials and one matrix inversion. Often the matrix polynomials may be computed with just a few matrix multiplications. As an example, note that when H is an Erlang distribution of order n, i.e., E_n, with mean one, then $h(s) = (1+n^{-1}s)^{-n}$ and (50) is evaluated as

$$M' = (I+n^{-1}M)^{-n}. \tag{56}$$

If $n=2^m$, for an integer m, then only m matrix multiplications are needed. For example, evaluating (56) when h is E_{1024} requires one matrix inverse and ten matrix multiplications. This example also shows that we can approach very close to the practically important, non-Coxian, deterministic distribution with relatively few matrix operations. (Of course, the algorithms apply directly to general service-time distributions, but computing integrals like (50) even in the deterministic case is more computationally intensive then in the E_n, case, for large n; moreover, it is harder to get high accuracy in the transform inversion algorithms for deterministic service).

8. Numerical Results

In this section, we demonstrate the computability of our results. 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 1-5 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 except Figure 3, 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 the transient workload tail probabilities (i.e., the transient complementary cdf of the workload) at time $t=10$ with different initial queue lengths and $\rho=0.7$. We have summed over all auxiliary phase states at time $t=10$, so that we obtain a one-dimensional distribution. In particular, we display $P(\ V(10)>x \mid X(0)=i_0,\ J(0)=j_0\ )$ for the designated initial phase state $j_0=2$ corresponding to two of the four *MMPP*'s starting in the high-rate state. We consider four different initial queue lengths: $i_0 = 0$, 2, 8 and 32. It is interesting to note that for small x the transient complementary cdf may be higher or lower than the corresponding stationary values, depending on the initial conditions, but for larger x, the transient results alway decay faster then the stationary distribution. We elaborate on this point in [18].

Figure 2 shows how the transient workload distribution approaches the stationary distribution as t increases. In particular, Figure 2 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 behavior also depends on the initial phase of the *BMAP* as shown in Figure 3. Here we show the transient distribution for a fixed time $t=10$ and a fixed initial queue length of 2. We vary the number of sources in the high-rate state, considering the cases of 0, 2 and 4.

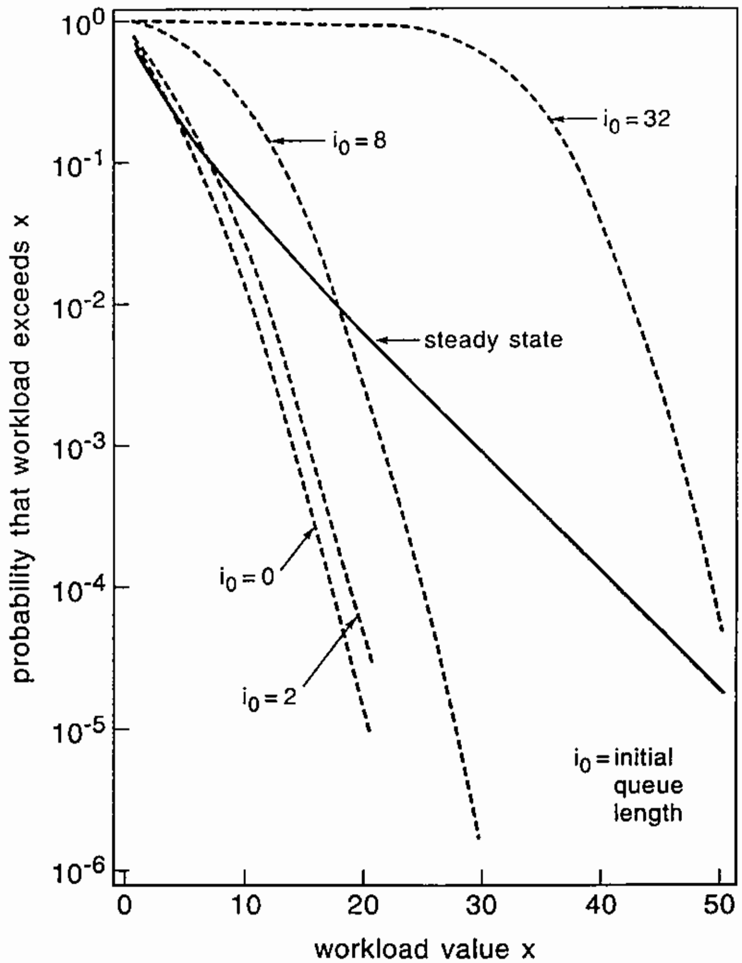

Figure 1. Numerical results for the workload tail probabilities as a function of the initial queue length i_0 in the $\sum^{4} MMPP_i/E_{16}/1$ queue with traffic intensity $\rho = 0.7$, time $t = 10$ and $j_0 = 2$ of the four MMPPs starting in the high-rate state.

The transient distributions are proper for $\rho \geq 1$, as well. This is demonstrated in Figure 4 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 $t \to \infty$, $V(t) \to \infty$ w.p. 1, so that $V(t) \to V(\infty)$, 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,

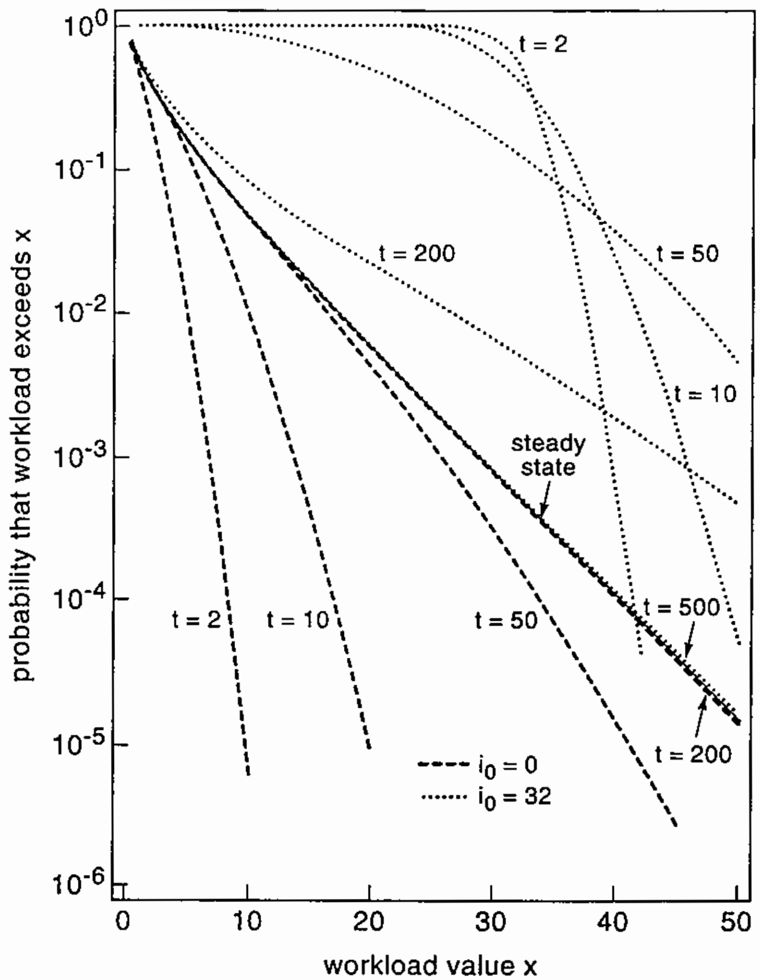

Figure 2. Numerical results for the workload tail probabilities as a function of the time t and initial queue length i_0 in the $\sum^4 MMPP_i/E_{16}/1$ queue with traffic intensity $\rho = 0.7$ and $j_0 = 2$ two of the four MMPPs starting in the high-rate state.

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 5, 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

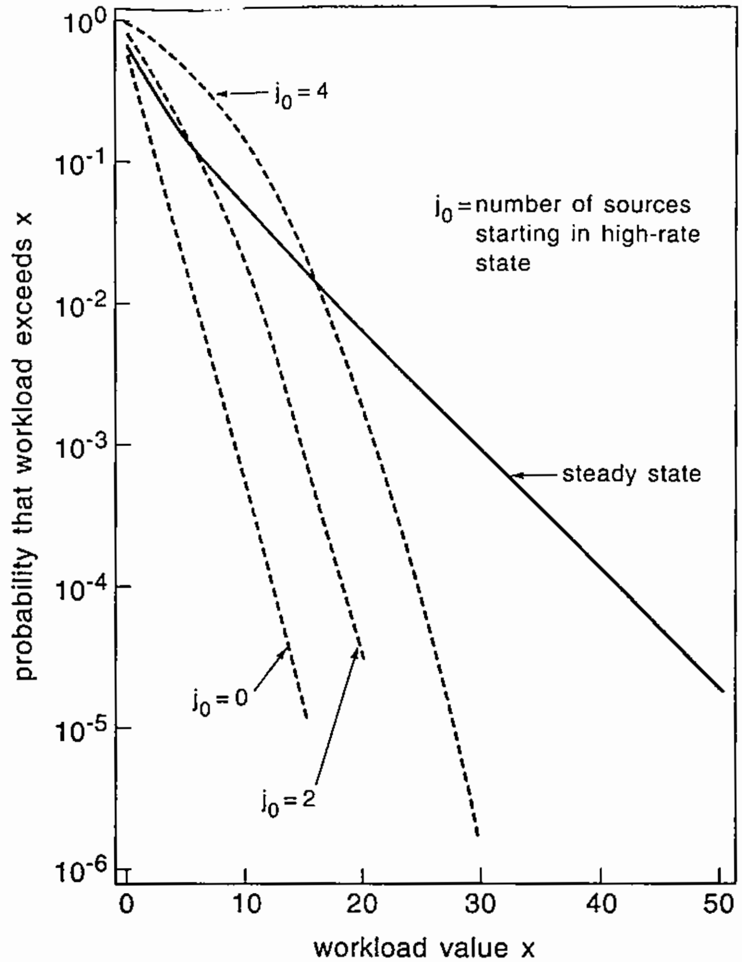

Figure 3. Numerical results for the workload tail probabilities as a function of the number j_0 of MMPPs initially in the high-rate state in the $\sum^{4} MMPP_i/E_{16}/1$ model with traffic intensity $\rho = 0.7$, time $t = 10$ and initial queue length $i_0 = 2$.

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.

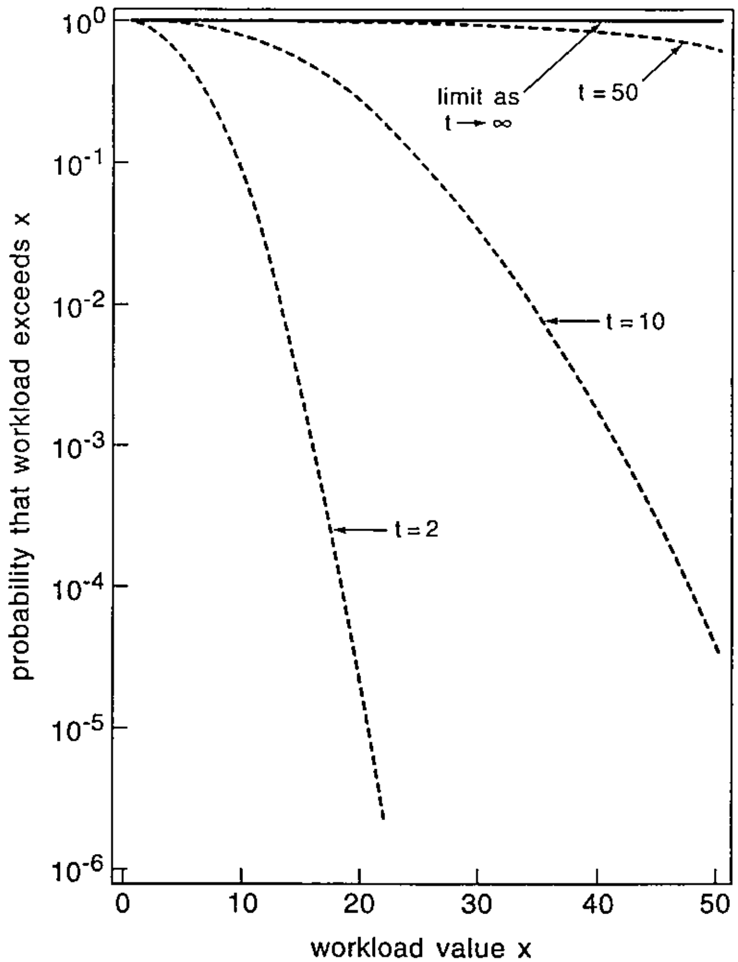

Figure 4. Numerical results for the workload tail probabilities as a function of time t in the unstable $\sum^{4} MMPP_i/E_{16}/1$ model with traffic intensity $\rho = 2.0$. In each case, the initial queue length is $i_0 = 2$ and the number of the four MMPPs starting off in the high-rate state is $j_0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

9. Proofs

In several of the following proofs, multiple interchanges of integrals are required. In all cases the integrands are either probabilities, generating functions or Laplace transforms so that the interchanges are justified by the Bounded Convergence Theorem (see, e.g., p.81 of Royden [39]).

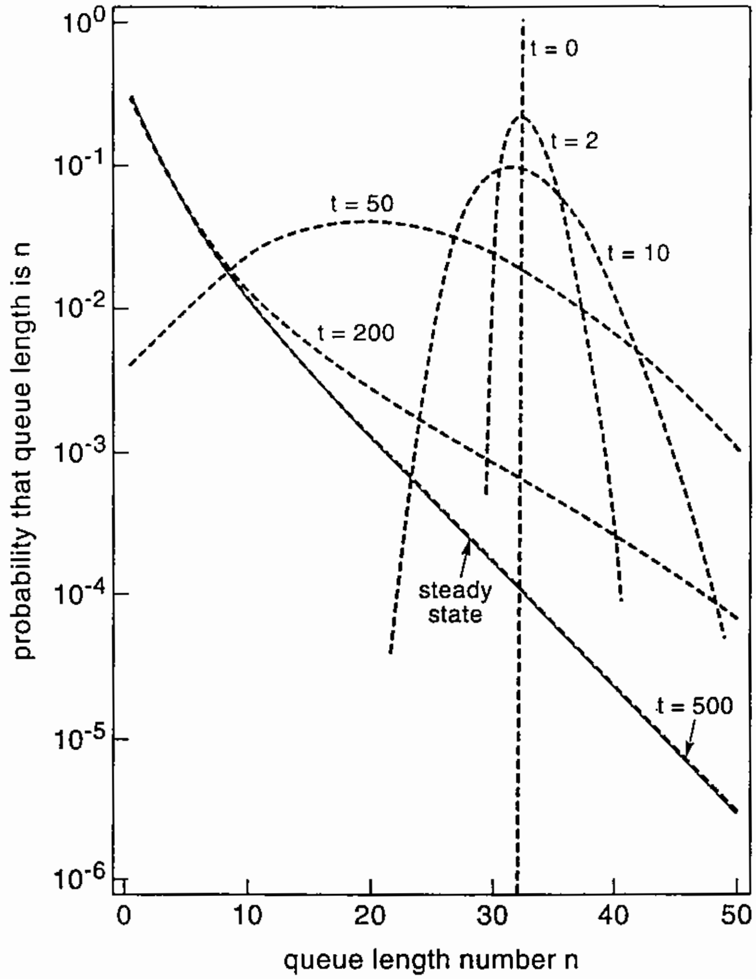

Figure 5. Numerical results for the transient queue-length probability mass function as a function of time t in the $\sum^{4} MMPP_i/E_{16}/1$ model with traffic intensity $\rho = 0.7$, initial queue length $i_0 = 32$ and $j_0 = 2$ of the four MMPPs starting in the high-rate state.

Proof of Theorem 1 We know from Lemma 2 in Lucantoni and Neuts [40] that the Laplace transform of the time required for the system to empty given an initial workload of x, and keeping track of the phase change, is given by $e^{-(sI-D[G(s)])x}$. Therefore,

$$p_{x0}(s) = e^{-(sI-D[G(s)])x} p_{00}(s). \tag{57}$$

Now, if we condition on the first arrival before t (if any), we get

$$P_{00}(t) = e^{D_0 t} + \int_0^t e^{D_0 u} \sum_{k=1}^{\infty} D_k du \int_0^{t-u} dG^{(k)}(v) P_{00}(t-u-v). \tag{58}$$

The first term corresponds to the case where there are no arrivals before t. The second term corresponds to the case where there is a batch arrival of size k at time u and the system next empties out v time units later (at time $u+v$). Taking Laplace transforms, exploiting the convolution in (58) and letting $y = t-u$, we obtain

$$\begin{aligned} p_{00}(s) &= (sI - D_0)^{-1} + \int_0^{\infty} e^{-su} e^{D_0 u} du \int_0^{\infty} e^{-sy} dy \sum_{k=1}^{\infty} D_k \int_0^{y} dG^{(k)}(v) P_{00}(y-v) \\ &= (sI - D_0)^{-1} + (sI - D_0)^{-1} \Big[D[G(s)] - D_0 \Big] p_{00}(s). \end{aligned}$$

Rearranging the terms gives $p_{00}(s) = (sI - D[G(s)])^{-1}$ which combined with (57) gives (16).

Proof of Theorem 2

We first prove the following lemma.

Lemma 1: The following integral is explicitly evaluated as

$$\begin{aligned} &\int_0^{\infty} e^{-(\xi I - D(h(s)))y} dy \int_0^{\infty} e^{-sw} d_w H(y+w) [\xi I - sI - D(h(s))] \\ &\qquad = h(s) I - \tilde{A}(h(s), \xi). \end{aligned} \tag{59}$$

Proof of Lemma 1: Using the change of variable, $v = y+w$, we have

$$\begin{aligned} &\int_0^{\infty} e^{-(\xi I - D(h(s)))y} dy \int_0^{\infty} e^{-sw} d_w H(y+w) [\xi I - sI - D(h(s))] \\ &\qquad = \int_0^{\infty} e^{-vs} dH(v) \int_0^{v} e^{-(\xi I - sI - D(h(s)))y} dy [\xi I - sI - D(h(s))] \\ &\qquad = [h(s) I - h(\xi I - D(h(s)))] \end{aligned}$$

which, with (10), proves the result. ■

We now prove Theorem 2. First note that the mass at the origin, $W_{i_0}(t,0) = \hat{P}_{i_0 0}(t)$, with Laplace transform $m_{i_0 0}(\xi)(\xi I - D_0)^{-1}$ (from (18)). Now, by conditioning on the last departure before time t, we can write

$$\begin{aligned} W_{i_0}(t,x) &- W_{i_0}(t,0) \\ &= \sum_{i=0}^{\infty} \int_0^t dM_{i_0 0}(u) \int_0^{t-u} e^{D_0 v} \sum_{k=1}^{\infty} D_k dv \\ &\quad \times \int_0^x P(i,t-u-v)\, d_w H(t+w-u-v)\, H^{(i+k-1)}(x-w) \qquad (60) \\ &\quad + \sum_{i=1}^{\infty} \sum_{k=1}^{i} \int_0^t dM_{i_0 k}(u) \int_0^x P(i-k,t-u)\, d_w H(t+w-u)\, H^{(i-1)}(x-w), \end{aligned}$$

where $H^{(i)}$ is the i-fold convolution of H with itself. The first term corresponds to the case where the last departure occurs at time u and leaves the system empty; there is a batch arrival of size k at time $u+v$; the service time of the first customer lasts until time $t+w$; there are i additional arrivals between $u+v$ and t and the total service time of all customers present at time t is less than or equal to x. The second term corresponds to the case where the last departure left the system at time u with $k \geq 1$ customers remaining, and there are $i-k$ additional arrivals by time t. Also, we have

$$w_{i_0}(t,s) = W_{i_0}(t,0) + \int_0^{\infty} se^{-sx}(W_{i_0}(t,x) - W_{i_0}(t,0))\,dx. \qquad (61)$$

Multiplying the first term in (60) by se^{-sx} and integrating with respect to x from 0 to ∞, gives

$$\begin{aligned} &\int_0^{\infty} se^{-sx} dx \sum_{i=0}^{\infty} \int_0^t dM_{i_0 0}(u) \int_0^{t-u} e^{D_0 v} \sum_{k=1}^{\infty} D_k dv \\ &\qquad \times \int_0^x P(i,t-u-v)\, d_w H(t+w-u-v)\, H^{(i+k-1)}(x-w). \end{aligned}$$

Changing the order of integration with respect to x and w and making the change of variables $y = x-w$ leads to

$$\begin{aligned} &\int_0^t dM_{i_0 0}(u) \int_0^{t-u} e^{D_0 v} dv [D(h(s)) - D_0] \\ &\qquad \times \int_0^{\infty} e^{-sw} e^{D(h(s))(t-u-v)}\, d_w H(t+w-u-v)\, h(s)^{-1}, \end{aligned}$$

by using (5). Forming the Laplace transform of this by multiplying by $e^{-\xi t}$ and integrating followed by several change of variables gives

$$h(s)^{-1} m_{i_0 0}(\xi)(\xi I - D_0)^{-1}[D(h(s)) - D_0] \times \int_0^\infty e^{-(\xi I - D(h(s)))y} dy \int_0^\infty e^{-sw} d_w H(y+w). \tag{62}$$

Now, multiplying the second term in (60) by se^{-sx}, integrating with respect to x from 0 to ∞, and performing similar manipulations leads to

$$\sum_{k=1}^{\infty} \int_0^t dM_{i_0 k}(u) \int_0^\infty e^{-sw} e^{D(h(s))(t-u)} d_w H(t+w-u) h(s)^{k-1}.$$

Forming the Laplace transform of this by multiplying by $e^{-\xi t}$ and integrating leads to

$$h(s)^{-1}[\tilde{m}_{i_0}(h(s),\xi) - m_{i_0 0}(\xi)] \times \int_0^\infty e^{-(\xi I - D(h(s)))y} dy \int_0^\infty e^{-sw} d_w H(y+w). \tag{63}$$

Next, adding (62) and (63), post-multiplying by $[\xi I - sI - D(h(s))]$ and using Lemma 1 and (12), we have

$$\begin{aligned}(\tilde{w}_{i_0}(\xi,s) - m_{i_0 0}(\xi))[\xi I - sI - D(h(s))] &= \Big\{ m_{i_0 0}(\xi)(\xi I - D_0)^{-1}[D(h(s)) - D_0] \\ &\quad + \tilde{m}_{i_0}(h(s),\xi) - m_{i_0 0}(\xi)\Big\} [h(s)I - \tilde{A}(h(s),\xi)] h(s)^{-1},\end{aligned}$$

so that

$$\tilde{w}_{i_0}(\xi,s)[\xi I - sI - D(h(s))] = h(s)^{i_0} I - s\hat{p}_{i_0 0}(\xi). \tag{64}$$

This yields (25). Taking the Laplace transform of $w_{i_0}(t,s)$ in (26) readily leads to (25).

Proof of Theorem 3

Conditioning on the amount of work at time $t=0$, we can write

$$W(t,x) = \sum_{i=0}^{\infty} \int_0^t P(i,y) W_i(t-y,x) dF(y) \tag{65}$$

$$+ \sum_{i=0}^{\infty} \int_{t}^{t+x} P(i,t) H^{(i)}(t+x-y)\, dF(y),$$

where the first term corresponds to the case where the amount of work at time $t=0$ is less than or equal to t and the second term corresponds to the case where the amount of work at $t=0$ is greater than t. (Note that it must be less than or equal to $t+x$ for the work in the system at time t to be less than x.) The Laplace-Stieltjes transform with respect to x is

$$\begin{aligned} w(t,s) &\equiv \int_0^{\infty} se^{-sx} W(t,x)\, dx \\ &= \sum_{i=0}^{\infty} \int_0^{t} P(i,y)\, w_i(t-y,s)\, dF(y) \qquad (66) \\ &\quad + \sum_{i=0}^{\infty} \int_0^{\infty} se^{-sx} dx \int_{t}^{t+x} P(i,t) H^{(i)}(t+x-y)\, dF(y). \end{aligned}$$

The second term becomes, after an interchange of integrals and a subsequent change of variables,

$$\begin{aligned} &\sum_{i=0}^{\infty} \int_t^{\infty} dF(y) \int_{y-t}^{\infty} se^{-sx} P(i,t) H^{(i)}(t+x-y)\, dx \\ &= \sum_{i=0}^{\infty} \int_t^{\infty} dF(y) \int_0^{\infty} se^{-s(y+u-t)}\, P(i,t) H^{(i)}(u)\, du \\ &= \sum_{i=0}^{\infty} h(s)^i e^{st} P(i,t) \int_t^{\infty} e^{-sy} dF(y) \;=\; e^{(sI+D(h(s)))t} \int_t^{\infty} e^{-sy} dF(y), \end{aligned}$$

by using (5). The Laplace transform of $w(t,s)$ in (66) with respect to t is given by

$$\begin{aligned} \tilde{w}(\xi,s) &= \sum_{i=0}^{\infty} \int_0^{\infty} dF(y) \int_y^{\infty} e^{-\xi t} P(i,y)\, w_i(t-y,s)\, dt \qquad (67) \\ &\quad + \int_0^{\infty} dF(y) \int_0^{y} e^{-(\xi I - sI - D[h(s)])t} e^{-sy} dt. \end{aligned}$$

Upon applying (5) and (25), we see that the first term in (67) becomes

$$\sum_{i=0}^{\infty} \int_0^{\infty} e^{-\xi y} P(i,y)\, \tilde{w}_i(\xi,s)\, dF(y)$$

$$= \left\{ f(\xi I - D(h(s))) - sf(\xi I - D[G(\xi)]) \left[\xi I - D[G(\xi)] \right]^{-1} \right\}$$

$$\times [\xi I - sI - D(h(s))]^{-1}, \tag{68}$$

where we only consider pairs (ξ, s) for which the inverse in (68) exists. The second term in (67) is simplified as follows.

$$\int_0^\infty dF(y) \int_0^y e^{-(\xi I - sI - D(h(s)))t} e^{-sy} dt$$

$$= \int_0^\infty e^{-sy} (I - e^{-[\xi I - sI - D(h(s))]y}) [\xi I - sI - D(h(s))]^{-1} dF(y)$$

$$= [f(s)I - f(\xi I - D(h(s)))][\xi I - sI - D(h(s))]^{-1}. \tag{69}$$

Adding (68) and (69) yields (28). Finally, taking the Laplace transform of (29) yields (28).

Proof of Theorem 4

Multiplying both sides of (28) by $[\xi I - sI - D(h(s))]$, differentiating with respect to s and setting $s = 0$ readily leads to

$$-\frac{\partial \tilde{w}(\xi, 0)}{\partial s} = [\beta I + p_0(\xi) + \tilde{w}(\xi, 0)(\alpha D^{(1)} - I)](\xi I - D)^{-1}. \tag{70}$$

Inverting (70) by inspection, noting that $w(t,0) = e^{Dt}$, we obtain (30). Note that D and $D^{(1)}$ do not commute in general. Equations (31) and (32) follow routinely from (30).

Proof of Theorem 5

By successively differentiating with respect to s in

$$\tilde{w}(\xi, s)[\xi I - sI - V(s)] = f(s)I + sp_0(\xi),$$

we obtain, for $k \geq 2$,

$$\frac{\partial^k}{\partial s^k} \tilde{w}(\xi, s)[\xi I - sI - V(s)] = f^{(k)}(s) + k \frac{\partial^{k-1}}{\partial s^{k-1}} \tilde{w}(\xi, s)$$

$$+ \sum_{j=0}^{k-1} \binom{k}{j} \frac{\partial^j}{\partial s^j} \tilde{w}(\xi, s) \frac{d^{k-j}}{ds^{k-j}} V(s).$$

Setting $s=0$, multiplying by $(-1)^k$ and inverting the transform by inspection, we obtain (35). Differentiating in (35) with respect to t yields (36).

Proof of Theorem 6

Once again, by conditioning on the last departure before time t we can write $Y_{i_0 i}(t)$ as

$$\int_0^t dM_{i_0 0}(u) \int_0^{t-u} e^{D_0 v} \sum_{k=1}^{i} D_k dv P(i-k,t-u-v)[1-H(t-u-v)]$$
$$+ \sum_{j=1}^{i} \int_0^t dM_{i_0 j}(u) P(i-j,t-u)[1-H(t-u)]. \quad (71)$$

The first term corresponds to the case where the last departure left the system empty and the second term corresponds to where the last departure left j customers in the system. Taking Laplace transforms leads successively to

$$\int_0^\infty e^{-st}dt \int_0^t dM_{i_0 0}(u) \int_0^{t-u} e^{D_0 v} \sum_{k=1}^{i} D_k dv P(i-k,t-u-v)[1-H(t-u-v)]$$
$$+ \sum_{j=1}^{i} \int_0^\infty e^{-st}dt \int_0^t dM_{i_0 j}(u) P(i-j,t-u)[1-H(t-u)]$$
$$= m_{i_0 0}(s) \int_0^\infty dv \int_v^\infty e^{-sx} e^{D_0 v} \sum_{k=1}^{i} D_k P(i-k,x-v)[1-H(x-v)]\,dx$$
$$+ \sum_{j=1}^{i} m_{i_0 j}(s) \int_0^\infty e^{-sx} P(i-j,x)[1-H(x)]\,dx$$
$$= m_{i_0 0}(s)(sI-D_0)^{-1} \int_0^\infty e^{-sw} \sum_{k=1}^{i} D_k P(i-k,w)[1-H(w)]\,dw$$
$$+ \sum_{j=1}^{i} m_{i_0 j}(s) \int_0^\infty e^{-sx} P(i-j,x)[1-H(x)]\,dx.$$

Taking probability generating functions yields

$$\tilde{y}_{i_0}(z,s) \equiv \sum_{i=0}^{\infty} y_{i_0 i}(s) z^i$$
$$= y_{i_0 0}(s) + m_{i_0 0}(s)(sI-D_0)^{-1}[D(z)-D_0] \int_0^\infty e^{-sw} P^*(z,w)[1-H(w)]\,dw$$

$$+[m_{i_0}(z,s) - m_{i_0 0}(s)]\int_0^{\infty} e^{-sw}P^*(z,w)[1-H(w)]\,dw. \tag{72}$$

Recall that $y_{i_0 0}(s) = G(s)^{i_0} p_{00}(s)$. The integral on the right side of each of the above terms is evaluated by applying (5), (10) and integration by parts to give

$$\int_0^{\infty} e^{-(sI-D(z))x}[1-H(x)]\,dx = (sI-D(z))^{-1}[I-\tilde{A}(z,s)]. \tag{73}$$

Substituting (73) into (72) and simplifying leads to (41).

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. Neuts, M. F., *Structured Stochastic Matrices of M/G/*1 *Type and Their Applications*, New York: Marcel Dekker, 1989.

3. Neuts, M. F., A versatile Markovian point process, *J. Appl. Prob.*, **16** (1979) 764-79.

4. Ramaswami, V., The *N/G/*1 queue and its detailed analysis, *Adv. Appl. Prob.*, **12** (1980) 222-61.

5. 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.

6. 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.

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., Tail probabilities in queues with many independent sources, in preparation.

11. Choudhury, G. L., Lucantoni, D. M., and Whitt, W., Squeezing the most out of ATM, submitted for publication, 1993.

12. Takács, L., *Introduction to the Theory of Queues*, New York: Oxford University Press, 1962.

13. Abate, J, and Whitt, W., Transient behavior of the *M/G/*1 workload process, to appear in *Oper. Res.*, 1993.

14. Cohen, J. W., *The Single Server Queue*, Amsterdam: North-Holland, 1969.

15. Benes, V., *General Stochastic Processes in the Theory of Queues*, Reading, MA: Addison-Wesley, 1963.

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

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

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

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

20. 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.

21. Lucantoni, D. M., Further transient analysis of the *BMAP/G/*1 queue, in preparation.

22. Saito, H., The departure process of an *N/G/*1 queue, *Perf. Eval.*, **11** , (1990) 241--251.

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

24. Çinlar, E, *Introduction to Stochastic Processes*, Englewood Cliffs, NJ: Prentice-Hall, 1973.

25. Gantmacher, F. R., *The Theory of Matrices, Vol. 1*, New York: Chelsea, 1977.

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

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

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

29. Gantmacher, F. R., *The Theory of Matrices, Vol. 2*, New York: Chelsea, 1977.

30. Neuts, M. F., The single server queue with Poisson input and semi-Markov service times, *J. Appl. Prob.*, **3** (1966) 202-230.

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

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

33. Miyazawa, M., Rate conservation laws: a survey, *Queueing Systems,* to appear.

34. Riordan, J., *An Introduction to Combinatorial Analysis,* New York: Wiley, 1958.

35. Riordan, J., *Combinatorial Identities,* New York: Wiley, 1968.

36. Klimko, E. M. and Neuts, M. F., The single server queue in discrete time - numerical analysis II., *Nav. Res. Log. Quart.,* **20** (1973) 305-19.

37. Lucantoni, D. M. and Neuts, M. F., The customer delay in a single server queue with a batch Markovian arrival process, submitted for publication.

38. Davis, P. J., and Rabinowitz, P., *Methods of Numerical Integration,* 2nd. ed. New York: Academic Press, 1984.

39. Royden, H. L., *Real Analysis,* 2nd edition, New York: Macmillan Publishing, 1968.

40. Lucantoni, D. M., and Neuts, M. F., Simpler proofs of some properties of the fundamental period of the *MAP/G/*1 queue, to appear in *J. Appl. Prob.,* 1994.

Received: 3/15/1993
Accepted: 7/24/1993

Recommended by Gennadi Falin, Editor