Further Transient Analysis of the *BMAP/G/1* Queue

David Lucantoni†
IsoQuantic Technologies

ABSTRACT

Previously, we derived the two-dimensional transforms of the emptiness function, 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.

We continue the transient analysis of this model in this paper by deriving explicit expressions for the transforms of the queue length at the n-th departure (assuming a departure at time $t = 0$), and the delay of the n-th arrival (keeping track of the appropriate phase changes). Also, 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.

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. In order to obtain tractable results, we assume that the arrival process is a batch Markovian arrival process (*BMAP*), as in Lucantoni [1] and [2]. The *BMAP* is a convenient representation of the *versatile Markovian point process* (Neuts [3]). The *BMAP* generalizes the *Markovian arrival process (MAP)*, which was introduced by Lucantoni, Meier-Hellstern and Neuts [4] and the *Markov-modulated Poisson process* (see, e.g., Heffes and Lucantoni [5]). Indeed, stationary *MAP*'s are dense in the family of all stationary point processes; see Asmussen and Koole [6].

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 [7] to study the effect of statistically multiplexing a large number of bursty sources.

In a previous paper [8] we derived the two-dimensional transforms of the emptiness function, the transient workload and queue-length distributions in the single-server queue with general service times and a batch Markovian arrival process. In this paper we continue the transient analysis of this model by deriving explicit expressions for the transforms 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). We also provide a partial characterization of the departure process by the double transform of the probability that the n-th departure occurs at time less than or equal to x. (This is similar to that derived by Saito [9]).

†IsoQuantic Technologies, 10 Oak Tree Lane, Wayside, NJ 07712; davidl@isoquantic.com; www.isoquantic.com

These transient results along with those in [8] can be regarded as matrix generalizations of transient results for the *M/G/1* queue, which can be found in Takács [10], Abate and Whitt [11] and references cited there. A distinctive feature of this paper and [8] in relation to previous papers on transient behavior for these *M/G/1*-type queues, is that *we demonstrated that our formulas are computable*. In particular, we calculated the time-dependent probability distributions by *numerically inverting the two-dimensional transforms*. For this purpose, we applied the two-dimensional transform inversion algorithms in Choudhury, Lucantoni and Whitt [12]. These algorithms are based on the Fourier-series method [13], and are enhancements and generalizations of the Euler and Lattice-Poisson algorithms described there. We note that the same multidimensional transform inversion algorithms can be used to obtain numerical results from the expressions in this paper.

The remainder of this paper is organized as follows. In Section 2 we review the definition and basic properties of the Batch Markovian Arrival Process and the single server queue with this arrival process. In Section 3, we review the transform of the duration of a busy period and the number of customers served during a busy period which plays a fundamental role in the transient solution of this model. Sections 4, 5 and 6 contain the main results on the transient distributions discussed in this paper. Some numerical examples are presented in Section 7. All of the proofs are given in Section 8.

2. The *BMAP/G/1* Queue

The Batch Markovian Arrival Process

The *BMAP* is a natural generalization of the Poisson process (see Lucantoni [1], [2]). 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

$$Q = \begin{bmatrix} D_0 & D_1 & D_2 & D_3 \\ 0 & D_0 & D_1 & D_2 \\ 0 & 0 & D_0 & D_1 \\ 0 & 0 & 0 & D_0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{1}$$

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*,*n*), $k \geq 1$, $1 \leq j, n \leq m$, correspond to batch arrivals of size *k*, and thus the batch size can depend on *j* and *n*. The matrix D_0 is a stable matrix (see e.g., pg. 251 of Bellman [14]), 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 [15]. This implies that the arrival process does not terminate.

Let π be the stationary probability vector of the Markov process with generator *D*, i.e., π satisfies

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

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

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

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

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 [2].

A key quantity used in the analysis of the *BMAP/G/1* queue is the matrix generating function

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

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

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

is given explicitly by

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

where $e^{D(z)t}$ is an exponential matrix (see e.g., pg. 169 of Bellman, [14]). Note that for Poisson arrivals, we have *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. Let $X(t)$ and $J(t)$ be the number of customers in the system (including customers 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 $\tau_0 = 0$). Define $\xi_k \equiv X(\tau_k^+)$ and $J_k \equiv J(\tau_k^+)$ to be the number in the system and the phase of the arrival process immediately following the *n*-th departure. Then the triple $(\xi_k, J_k, \tau_{k+1} - \tau_k)$, for $k \geq 0$, 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 $\rho < 1$; however, the transient results derived here are valid for any value of ρ. The transition probability matrix of the semi-Markov process is given by

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

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

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.

$\left[\hat{A}_n(x)\right]_{ij}$ = 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),

$\left[\hat{B}_n(x)\right]_{ij}$ = 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).

We introduce the transform matrices

$$\tilde{A}_n(s) = \int_0^\infty e^{-sx} d\hat{A}_n(x) \ , \quad A(z,s) = \sum_{n=0}^\infty \tilde{A}_n(s) z^n \ , \quad A(z) = A(z,0) \ ,$$

$$\tilde{B}_n(s) = \int_0^\infty e^{-sx} d\hat{B}_n(x) \ , \quad B(z,s) = \sum_{n=0}^\infty \tilde{B}_n(s) z^n \ , \quad B(z) = B(z,0) \ ,$$

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

$$A(z,s) = \int_0^\infty e^{-sx} e^{D(z)x} dH(x) = h(sI - D(z)) \tag{7}$$

and

$$B(z,s) = z^{-1}[sI - D_0]^{-1}[D(z) - D_0]A(z,s) \tag{8}$$

The definition in (7) above is consistent with the usual definition of a scalar function evaluated at a matrix argument (see Theorem 2, pg. 113 of Gantmacher, [16]). 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 (7) we see that $A(z,s)$ is a power series in $D(z)$. Thus, $A(z,s)$ and $D(z)$ commute. This property is often exploited in the proofs.

3. The Busy Period

Following the general treatment of Markov chains of *M/G/1*-type in [3], we define $\tilde{G}_{jj'}^{[r]}(k;x)$, $k \geq 1, x \geq 0$, 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 in exactly k transitions and no later than time x, and that (i, j') is the first state visited in the set $\{(i, j) 1 \leq j \leq m\}$. $\tilde{G}^{[r]}(k;x)$ is the matrix with elements $\tilde{G}_{jj'}^{[r]}(k;x)$.

By a first passage argument, it was shown in Neuts [17] that the joint transform matrix $G(z, s)$, defined by

$$G(z, s) = \sum_{k=1}^{\infty} \int_0^{\infty} e^{-sx} d\tilde{G}^{[1]}(k;x) z^k, \qquad \text{for } |z| \leq 1, \text{Re}(s) \geq 0 \quad ,$$

satisfies the nonlinear matrix equation

$$G(z, s) = z \sum_{n=0}^{\infty} A_n(s) G(z, s)^n \quad . \tag{9}$$

In the context of the *BMAP/G/1* queue, $G(z, s)$ governs the number served during, and the duration of, the busy period. It can be shown that the joint transform matrix governing the number served during and the duration of a busy period starting with r customers, is given by $G(z, s)^r$. Equation (9) is the key equation in the matrix analytic solution to queues of the *M/G/1* paradigm.

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

$$G(z, s) = z \int_0^{\infty} e^{-sx} e^{D[G(z, s)]x} dH(x) \equiv zh(sI - D[G(z, s)]) \quad , \tag{10}$$

where $D[G(z, s)] \equiv \sum_{k=0}^{\infty} D_k G(z, s)^k$. Equation (10) is the matrix analogue of Takács' equation for the busy period in the ordinary *M/G/1* queue [10].

We also define the transform of the number of customers served during a busy period, $G(z) \equiv G(z, 0)$.

4. The Queue Length at the *n*-th Departure

In this section, we derive the transform of the number of customers in the system at the n-th departure. We first derive it for the general class of *M/G/1*-type models, i.e., models which have an embedded Markov-renewal process with a transition matrix having the structure displayed in (6),

and then we present a simpler expression by particularizing the result for the *BMAP/G/1* queue.

The n-step transition probability matrices $P_{ij}^{(n)}$ are defined to have (k, l) elements

$$[P_{ij}^{(n)}]_{kl} = P(\xi_n = j, J_n = l \,|\, \xi_0 = i, J_0 = k)$$

If we define the transform matrix

$$P_i(z, w) \equiv \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} P_{ij}^n z^j w^n, \qquad |z| < 1, |w| < 1,$$

then we have the following

Theorem 1: For general models of *M/G/1*-type, the matrices $P_i(z, w)$ are given by

$$P_i(z, w) = [z^{i+1} I + wG(w)^i [I - K(w)]^{-1} [zB(z) - A(z)]][zI - wA(z)]^{-1}$$

where $K(w) \equiv w \sum_{n=0}^{\infty} B_n G(w)^n$.

For the *BMAP/G/1* queue, it has been shown (see, e.g., Lucantoni [1]) that

$$zB(z) - A(z) = -D_0^{-1} D(z) A(z) \tag{11}$$

and

$$K(w) = I - D_0^{-1} D[G(w)] \tag{12}$$

so we immediately have the following

Theorem 2: For the *BMAP/G/1* queue the matrices $P_i(z, w)$ are given by

$$P_i(z, w) = [z^{i+1} I - wG(w)^i D[G(w)]^{-1} D(z) A(z)][zI - wA(z)]^{-1} \tag{13}$$

Corollary 2.1: For the *M/G/1* queue, Equation (13) simplifies to

$$P_i(z, w) = \frac{1}{z - wA(z)} \left[z^{i+1} - \frac{wG(w)^i (1 - z) A(z)}{1 - G(w)} \right] \tag{14}$$

which is Equation (59) on page 70 of Takács [10].

5. The Delay of *n*-th Arrival in the *MAP/G/1* Queue

In this section we give the transform for the delay of the n-th arrival of the *MAP/G/1* queue. That is, for this section, we only allow single arrivals so that $D_k = 0$ for $k \geq 2$. (A similar result can be obtained for the general batch-arrival case but that will be left for future work). Let the matrix $W_n(x)$ have (i, j)-entry $[W_n(x)]_{ij}$ which is the probability that the delay of the n-th arrival is less than or equal to x, and that the phase immediately following that arrival is j, given that the phase at time $t = 0$ is i. We assume that the work in the system at the arrival of the first customer is $W_1(u)$. (Note that this is completely general; if an arbitrary time is picked as the origin, then W_1 would represent the total work in the system at the time of the first arrival after the origin. Let the LST of $W_n(x)$ be $w_n(s)$.

Define $U \equiv (-D_0)^{-1} D_1$. Then U is a stochastic matrix which keeps track of the phase at successive arrivals. We see that the probability that the next arrival after time 0 occurs at a time $\leq x$ and that the phase immediately following that arrival is j given that the phase was i at time $t = 0$, is given by the (i, j)-entry of the matrix

$$\int_0^x e^{D_0 u} D_1 du = (I - e^{D_0 x}) U .$$

We then have

Theorem 3: The joint transform $\tilde{w}(z, s) \equiv \sum_{n=1}^{\infty} w_n(s) z^n$, is given explicitly by

$$\tilde{w}(z, s) = z\left[w_1(s) - s\int_0^{\infty} dW_1(u) e^{D[G(z)]u} G(z) D[G(z)]^{-1} D_1 (sI + D_0)^{-1} U \right] \times [sI + D_0 + zh(s)D_1]^{-1}(sI + D_0) \qquad (15)$$

Corollary 3.1: For the *M/G/1* queue, Equation (15) reduces to

$$\tilde{w}(z, s) = z\left[\frac{(1 - G(z))(\lambda - s) w_1(s) - s\lambda G(z) w_1(\lambda(1 - G(z)))}{(1 - G(z))\lambda - s - \lambda z h(s)} \right] , \qquad (16)$$

which agrees with Equation (23) on page 57 of Takács [10].

6. The Departure Process

In this section we derive the transform of the distribution of the time till the n-th departure in the *BMAP/G/1* queue. Let the matrix $\tilde{U}_n(k, x)$ have (i, j)-entry

$$[\tilde{U}_n(k, x)]_{ij} = P(\xi_n = k, \tau_n \le x, J_n = j \,|\, J_0 = i) \quad .$$

Then

$$\tilde{U}_0(k, x) = \begin{cases} \Delta(\boldsymbol{r}_k), & x = 0, \\ 0, & x > 0, \end{cases}$$

where $\Delta(\boldsymbol{r}_k)$ is a diagonal matrix with the elements of $\boldsymbol{r}_k$ along the diagonal, and $r_{kj} = P(\xi_0 = k, J_0 = j)$ is the initial condition at the departure at time zero.

We next define the transforms

$$\hat{U}_n(z, x) = \sum_{k=0}^{\infty} z^k \tilde{U}_n(k, x), \qquad \text{for} \quad n \ge 0,$$

$$U_n(z, s) = \int_0^{\infty} e^{-sx} d_x \hat{U}_n(z, x) = \int_0^{\infty} s e^{-sx} \hat{U}_n(z, x) dx, \qquad \text{for} \quad n \ge 0,$$

(since $\hat{U}_n(z, 0) = 0$), for $n \ge 1$, and

$$U(z, s, w) = \sum_{n=0}^{\infty} U_n(z, s) w^n \quad .$$

We then have the following expression for the joint transform of the time of the n-th departure and the number in the system at that time.

Theorem 4: The matrix $U(z, s, w)$ is given explicitly by

$$U(z, s, w) = \Big[z U_0(z, s) - w U_0(G(w, s), s)(sI - D[G(w, s)])^{-1} (sI - D(z)) A(z, s) \Big] \times [zI - wA(z, s)]^{-1} \qquad (17)$$

where (with a slight abuse of notation) $U_0(G(w,s),s)$ is defined by

$$U_0(G(w,s),s) \equiv \sum_{j=0}^{\infty} \hat{U}_0(j,s)G(w,s)^j \quad .$$

Note that for the *M/G/1* queue, (17) reduces to

$$U(z,s,w) = \frac{1}{z - wA(z,s)}\left[zU_0(z,s) - \frac{wA(z,s)U_0(G(w,s),s)(s+\lambda(1-z))}{s+\lambda(1-G(w,s))}\right] \qquad (18)$$

which is Equation (69) in Takács [10]

Corollary 4.1: The joint transform of the probability that the n-th departure occurs at a time less than or equal to x, is

$$U(1,s,w) = \left[U_0(1,s) - wU_0(G(w,s),s)\left(sI - D[G(w,s)]\right)^{-1}(sI-D)A(1,s)\right] \times [zI - wA(1,s)]^{-1} \qquad (19)$$

Corollary 4.2:The Laplace transform of the expected number of departures up till time x is $U(1,s,1)$.

7. Examples

In this section we present several numerical examples demonstrating the computability of our results. The results are obtained by numerically inverting the multi-dimensional transforms directly using the algorithms presented in [12]. The details of those algorithms will not be described here. We consider N bursty sources as shown in Figure 1 where each source is an interrupted Poisson process (see, e.g., Kuczura [18]). That is, the process alternates between on and off periods where the durations of the on and off periods are exponentially distributed and there are Poisson arrivals during the on periods.

For the examples considered here, we fix $N = 8$. Each source has a mean arrival rate of 1/8 per unit time. The peak-to-mean ratio is 10:1 and the mean burst length is 0.1 time units; that is, the arrival rate in the on period is ten times the mean arrival rate and the mean on time is 0.1 time units. Thus the average arrival rate into the queue is one arrival per unit time. We approximate a deterministic service time by a high order Erlang distribution (E_{1024}). We adjust the mean service time to achieve the desired traffic intensity.

The first example has a traffic intensity of $\rho = 0.8$ and the density function of the queue length

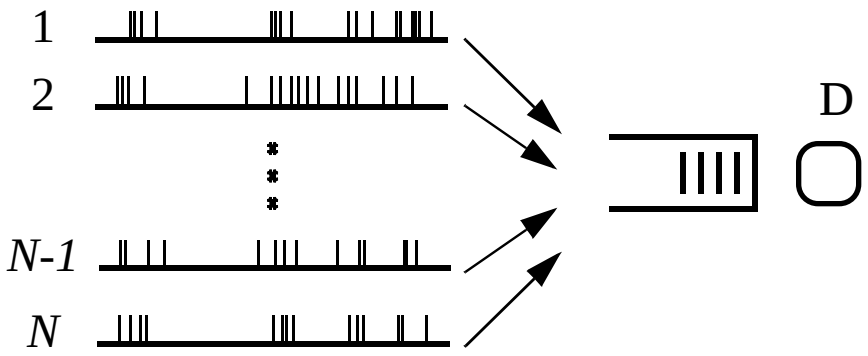

Figure 1: N Bursty Sources

at departures is shown in Figure 2. The solid line is the stationary density of the number in the system

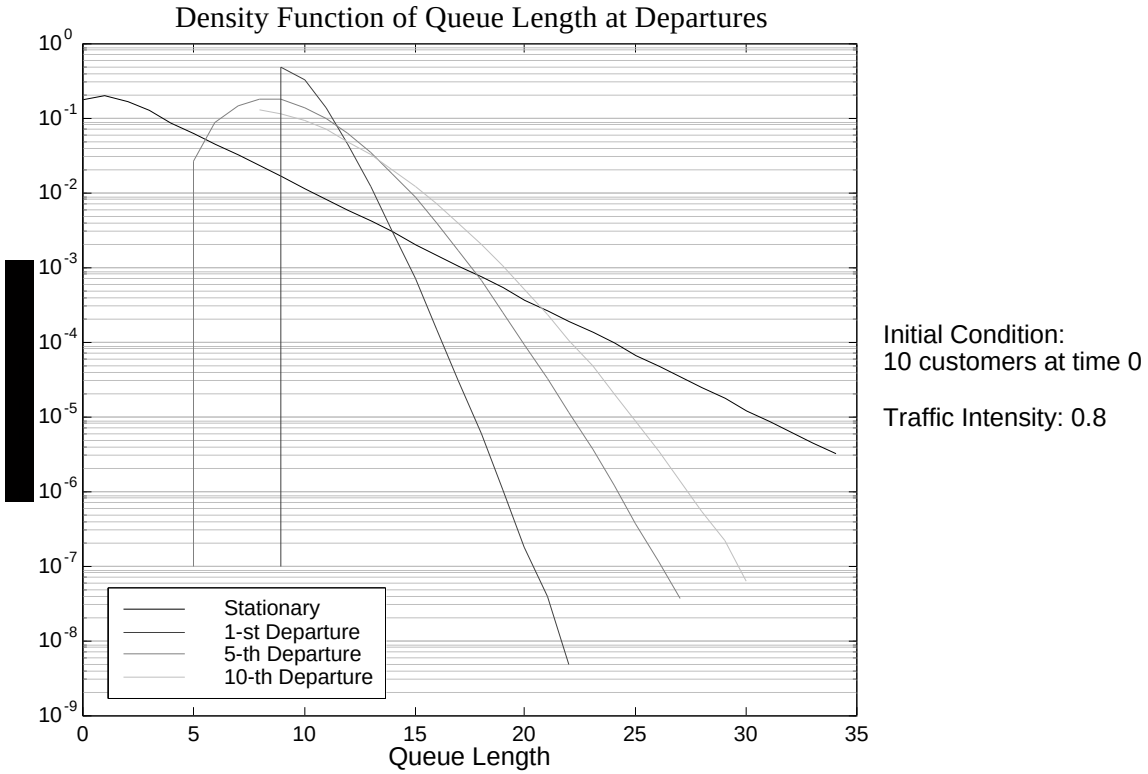

Figure 2: Queue Length at Departures for ρ = 0.8

immediately following departures. The other curves display the density of the number in the system following the 1-st, 5-th and 10-th departure, respectively, assuming that the number of customers in the system at time $t = 0$ was 10. We see that these distributions are converging to the stationary distribution but if we are interested in predicting performance on time scales on the order of tens of service times then the stationary distribution would be a poor predictor.

The next example is for a traffic intensity of $\rho = 1.5$. Note that this system is not stable and stationary distributions do not exist. The transient results are well defined, however, and this would be a useful model to determine how bad things can get during temporary overloads. The results are shown in Figure 3. Once again, we show the probability density function of the number of customers

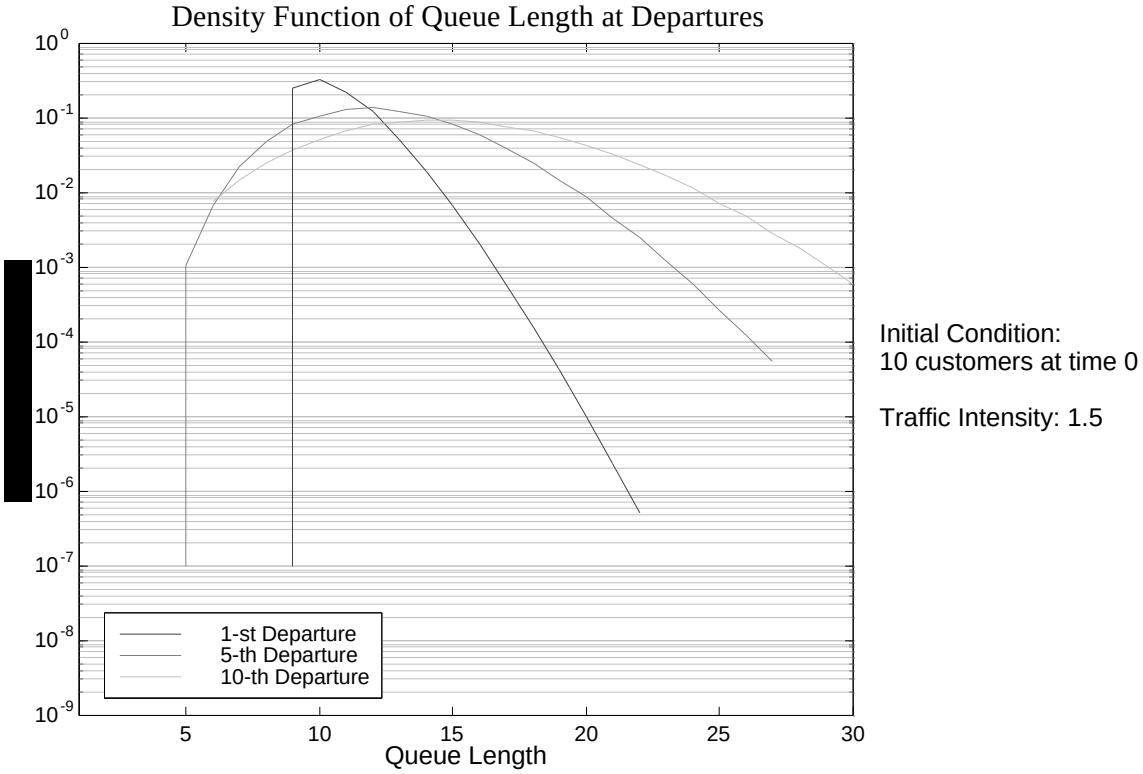

Figure 3: Queue Length at Departures for $\rho = 1.5$

left behind by the 1-st, 5-th and 10-th departure, respectively, assuming that the number of customers in the system at time $t = 0$ was 10. While these curves eventually converge to a straight line at probability equal to one, for small time scales the queue might be fairly well behaved. Such results will be useful in designing overload control algorithms.

8. Proofs of the Theorems

Proof of Theorem 1: Clearly,

$$P_{ij}^{(0)} = \begin{cases} I, & i = j, \\ 0, & i \neq j. \end{cases}$$

By conditioning on the number left behind by the $(n-1)$ -st departure, we have, for $n \geq 1$,

$$P_{ij}^{(n)} = P_{i0}^{(n-1)} B_j + \sum_{k=1}^{j+1} P_{ik}^{(n-1)} A_{j+1-k} \ , \tag{20}$$

where $B_j \equiv B_j(0)$ and $A_j \equiv A_j(0)$. Define the transform matrices

$$\hat{P}_{ij}(w) \equiv \sum_{n=0}^{\infty} P_{ij}^{(n)} w^n, \qquad P_i^{(n)}(z) \equiv \sum_{j=0}^{\infty} P_{ij}^{(n)} z^j .$$

Then from (20) we have

$$P_i^{(n)}(z) = z^{-1} P_i^{(n-1)}(z) + \hat{P}_{i0}^{(n-1)}(B(z) - z^{-1} A(z)) \tag{21}$$

and

$$P_i(z, w) - P_i^{(0)}(z) = \frac{w}{z} P_i(z, w) A(z) + w \hat{P}_{i0}(w)(B(z) - z^{-1} A(z)) . \tag{22}$$

Now, $P_i^{(0)}(z) = z^i I$, and by conditioning on the end of the busy period starting with *i* customers, we have

$$P_{i0}^{(n)} = \sum_{k=1}^{n} \tilde{G}^{(i)}(k) P_{00}^{(n-k)} \tag{23}$$

and therefore,

$$\hat{P}_{i0}(w) = G(w)^i \hat{P}_{00}(w) \ . \tag{24}$$

The matrix generating function *K(z)* can be written as $K(z) = \sum_{j=0}^{\infty} \tilde{K}(j) z^j$, where $\tilde{K}(j)$ is the probability that there are *j* transitions (i.e., departures) between successive returns to an empty system (keeping track of the appropriate phase changes). We can then write

$$P_{00}^{(n)} = \sum_{j=1}^{n} \tilde{K}(j) z^j P_{00}^{(n-j)}$$

and

$$\hat{P}_{00}(w) = I + K(w)\hat{P}_{00}(w) = (I - K(w))^{-1}. \tag{25}$$

Substituting (24) and (25) into (22) and simplifying yields Theorem 1.

Proof of Theorem 3: To simplify our analysis we define

$$R_n(x) = \int_0^x dW_n(u)H(x-u),$$

so that the (i, j)-entry of $R_n(x)$ is the probability that the work in the system and the phase immediately following the n-th arrival are less than or equal to x and j, respectively, given an arrival phase of i at $t = 0$. Let $r_n(s)$ and $w_n(s)$ be the Laplace-Stieltjes transforms of $R_n(x)$ and $W_n(x)$, respectively. Then, $r_n(s) = w_n(s)h(s)$.

Clearly, the $(n + 1)$-st arrival will see an empty system if and only if there are no arrivals during the time it takes to work off the total amount of work present immediately following the n-th arrival. Therefore,

$$W_{n+1}(0) = \int_0^\infty dR_n(u)e^{D_0 u}.$$

Let the matrix $W_n^c(x)$ have (i, j)-entry $[W_n^c(x)]_{ij}$ which is the probability that the delay of n-th arrival is greater than x, and the arrival phase immediately following that arrival is j, given that the phase at time t=0 is i. Then we have

$$W_n(x) + W_n^c(x) = U^n;$$

that is, disregarding the time till the n-th arrival, $W + W^c$ keeps track of the phase change during the n arrivals. Therefore, we can write

$$w_{n+1}(s) = U^{n+1} - \int_0^\infty se^{-sx}W_{n+1}^c(x)dx. \tag{26}$$

But then

$$W_{n+1}^{C}(x) = \int_{x}^{\infty} dR_n(u)(I - e^{D_0(u-x)})U. \tag{27}$$

Substituting (27) into (26), simplifying, and using the fact that $\int_{0}^{\infty} dR_n(u) = U^n$, leads to

$$w_{n+1}(s) = sW_{n+1}(0)(sI + D_0)^{-1}U - w_n(s)h(s)(sI + D_0)^{-1}D_1. \tag{28}$$

Note that for the *M/G/1* queue, Equation (28) reduces to Equation (25) on page 57 of Takács [10].

From (28), we obtain the generating function $\tilde{w}(z, s) = \sum_{n=1}^{\infty} w_n(s)z^n$ explicitly as

$$\tilde{w}(z, s) = \left(zw_1(s) + s\sum_{n=2}^{\infty} W_n(0)z^n\right)(sI + D_0)^{-1}U\div(sI + D_0 + zh(s)D_1)^{-1}(sI + D_0). \tag{29}$$

Now, in order to obtain an expression for $\sum_{n=2}^{\infty} W_n(0)z^n$ where $W_n(0)$ is the probability that the n-th arrival finds the system empty, we note that for this to occur, the $(n-1)$-st arrival must leave the system empty at its *departure*. Therefore, given $W_1(t)$, let $R_1(t) = \int_{0}^{t} dW_1(u)H(t-u)$ be the total work in the system immediately following the first arrival. (Note that we may arbitrarily select a customer as the first arrival and that there may in fact be previous work in the system at the arrival of this *first* customer.) Then for $n \geq 2$, we have

$$W_n(0) = \int_{0}^{\infty} dR_1(u) \sum_{j=0}^{n-2} P(j, u)P_{j0}^{(n-2)}U. \tag{30}$$

Equation (30) is obtained by noting that there were j arrivals during time u at which point the *first* customer departs. Now, with j customers in the system, $0 \leq j \leq n-2$, the $(n-2)$-nd departure leaves the system empty (with probability $P_{j0}^{(n-2)}$). At this point, the n-th arrival will be the next arrival and the system will still be empty. The matrix U keeps track of the arrival phase during the time between the $(n-1)$-st departure and the n-th arrival.

Multiplying (30) by z^n, summing over n and simplifying leads to

$$\sum_{n=2}^{\infty} W_n(0)z^n = -z\int_0^{\infty} dW_1(u)e^{D[G(z)]u}G(z)D[G(z)]^{-1}D_1. \tag{31}$$

For the *M/G/1* queue, this reduces to Equation (27) on page 57 of Takács [10]. We use a purely probabilistic argument which avoids the discussion of the roots of a transcendental equation. Finally, substituting (31) into (29) leads to (15).

Proof of Theorem 4: By conditioning on the previous departure we have

$$\begin{aligned}\tilde{U}_{n+1}(k,x) = \int_0^x d_u\tilde{U}_n(0,u)\int_0^{x-u} e^{D_0 v}\sum_{l=1}^{k+1} D_l dv \int_0^{x-u-v} P(k+1-l,w)dH(w) \\ + \sum_{l=1}^{k+1}\int_0^x d_u\tilde{U}_n(l,u)\int_0^{x-u} P(k+1-l,w)dH(w)\end{aligned} \tag{32}$$

where the first term corresponds to the case where the n-th departure left the system empty and the second term corresponds to the case where the n-th departure left l customers in the system. Using the fact that

$$\sum_{k=0}^{\infty} z^k \sum_{l=1}^{k+1} D_l P(k+1-l,w) = z^{-1}(D(z)-D_0)e^{D(z)w} \quad ,$$

we have

$$\begin{aligned}\hat{U}_{n+1}(z,x) = z^{-1}\int_0^x d_u\tilde{U}_n(0,u)\int_0^{x-u} e^{D_0 v}(D(z)-D_0)dv \int_0^{x-u-v} e^{D(z)w}dH(w) \\ + z^{-1}\int_0^x d_u(\hat{U}_n(z,u)-\hat{U}_n(0,u))\int_0^{x-u} e^{D(z)w}dH(w)\end{aligned} \tag{33}$$

Since $\hat{U}_n(z,0) = 0$ for $n \geq 1$ and $|z| \leq 1$, we have

$$
\begin{aligned}
U_{n+1}(z,s) &= \int_0^\infty s e^{-sx} \hat{U}_{n+1}(z,x)\,dx \\
&= z^{-1} U_n(0,s)(sI - D_0)^{-1}(D(z) - D_0)A(z,s) + z^{-1}(U_n(z,s) - U_n(0,s))A(z,s) \\
&= z^{-1}\left(U_n(0,s)(sI - D_0)^{-1}(D(z) - sI) + U_n(z,s)\right)A(z,s)
\end{aligned}
\tag{34}
$$

Note that for *M/G/1*, Equation (34) reduces to the equation between Equations (73) and (74) on page 73 of Takács [10].

We then easily obtain

$$
\begin{aligned}
U(z,s,w) &= U_0(z,s) + \sum_{n=0}^{\infty} U_{n+1}(z,s) w^{n+1} \\
&= \left(zU_0(z,s) + U(0,s,w)(sI - D_0)^{-1}(D(z) - sI)A(z,s)\right)\left[zI - wA(z,s)\right]^{-1}.
\end{aligned}
\tag{35}
$$

where we still need to determine $U(0,s,w)$. To that end, we condition on the last departure before the *n*-th which left the system empty (if any) and write

$$
\begin{aligned}
\tilde{U}_n(0,x) = \sum_{k=1}^{n-1} \int_0^x d_u \tilde{U}_k(0,u) \int_0^{x-u} e^{D_0 v} \sum_{l=1}^{\infty} D_l\, dv\, \tilde{G}^{(l)}(n-k, x-u-v) \\
+ \sum_{j=0}^{n} \tilde{U}_0(j,0)\tilde{G}^{(j)}(n,x).
\end{aligned}
\tag{36}
$$

Taking the Laplace-Stieltjes transform in (36), multiplying by w^n, summing over n, and simplifying leads to

$$
\begin{aligned}
U(0,s,w) &= U(0,s,w)(sI - D_0)^{-1}[D[G(w,s)] - D_0] + U_0(G(w,s),s) \\
&= U_0(G(w,s),s)\left[sI - D[G(w,s)]\right]^{-1}(sI - D_0)
\end{aligned}
\tag{37}
$$

where $U_0(G(w,s),s) \equiv \sum_{j=0}^{\infty} \tilde{U}_0(j,s) G(w,s)^j$. Substituting (37) into (35) yields the result in (17).

9. References

1. Lucantoni, D. M., "New results for the single server queue with a batch Markovian arrival process," *Stoch. Mod.*, vol. 7, pp. 1-46, 1991.
2. Lucantoni, D. M., "The BMAP/G/1 queue: A tutorial," in *Models and Techniques for Performance Evaluation of Computer and Communications Systems*, L. Donatiello and R. Nelson, Eds.: Springer Verlag, 1993, pp. 330-58.
3. Neuts, M. F., *Structured Stochastic Matrices of M/G/1 Type and Their Applications*: Marcel Dekker, 1989.
4. Lucantoni, D. M., Meier-Hellstern, K. S., and Neuts, M. F., "The single server queue with server vacations and a class of non-renewal arrival processes," *Adv. Appl. Prob.*, vol. 22, pp. 676-705, 1990.
5. Heffes, H. and Lucantoni, D. M., "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. on Sel. Areas in Comm.*, vol. SAC-4, pp. 856-868, 1986.
6. Asmussen, S. and Koole, G., "Marked point processes as limits of Markovian arrival streams," *J. Appl. Prob.*, vol. 30, pp. 365-72, 1993.
7. Choudhury, G. L., Lucantoni, D. M., and Whitt, W., "Squeezing the most out of ATM," *IEEE Trans. on Comm.*, vol. 44, pp. 203-17, 1996.
8. Lucantoni, D. M., Choudhury, G. L., and Whitt, W., "The transient BMAP/G/1 queue," *Stoch. Mod.*, vol. 10, pp. 145-82, 1994.
9. Saito, H., "The departure process of an N/G/1 queue," *Perf. Eval.*, vol. 11, pp. 241-51, 1990.
10. Takács, L., *Introduction to the Theory of Queues*. New York: Oxford University Press, 1962.
11. Abate, J. and Whitt, W., "Transient behavior of the M/G/1 workload process," *Oper. Res.*, 1993.
12. Choudhury, G. L., Lucantoni, D. M., and Whitt, W., "Multidimensional transform inversion with applications to the transient M/G/1 queue," *Ann. Appl. Prob.*, vol. 4, pp. 719-740, 1994.
13. Abate, J. and Whitt, W., "The Fourier-series method for inverting transforms of probability functions," *Queueing Systems Theory Appl.*, vol. 10, pp. 5-88, 1992.
14. Bellman, R., *Introduction to Matrix Analysis*. New York: McGraw Hill, 1960.
15. Neuts, M. F., *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Baltimore: The Johns Hopkins University Press, 1981.
16. Gantmacher, F. R., *The Theory of Matrices*, vol. 1. New York: Chelsea, 1977.
17. Neuts, M. F., "Moment formulas for the Markov renewal branching process," *Adv. Appl. Prob.*, vol. 8, pp. 690-711, 1976.
18. Kuczura, A., "The interupted Poisson process as an overflow process," *Bell Syst. Tech. J.*, vol. 52, pp. 437-48, 1973.