The Transient Response of the $M/E_k/2$ Queue and Steady-State Simulation

Joseph R. Murray
W. David Kelton

Department of Industrial and Operations Engineering
The University of Michigan
Ann Arbor, Michigan 48109-2117

Technical Report 85-29

The Transient Response of the $M/E_k/2$ Queue and Steady-State Simulation*

Joseph R. Murray
W. David Kelton
Department of Industrial and Operations Engineering
The University of Michigan
Ann Arbor, Michigan 48109-2117

The probabilistic structure for the transient $M/E_k/2$ queue is derived in discrete time, where E_k denotes a k-stage Erlang distribution. This queue has a two-dimensional state space. Expressions in terms of transition probabilities are formulated for the expected delay in queue. Results are numerically evaluated for one case. The convergence behavior is similar to that seen in previous work on queues with one-dimensional state spaces. The implications for bivariate initialization of steady-state simulations are discussed.

Key Words: Queueing systems, Transient results, Simulation, Initialization of simulation models.

* This research was supported by a contract from Electronic Data Systems Decision Technologies Division to the University of Michigan.

1. INTRODUCTION

A review of the literature in queueing theory reveals an abundance of results for steady-state conditions and relatively few results for the transient phase of a queueing system. One reason for this is the complexity and intractability of the mathematics involved in solving the transient problem, and it is not uncommon to see results stated in terms of transforms which are very difficult, if not impossible, to invert. Results, when left as transforms, seem somewhat less satisfactory in terms of ease of interpretation. Analytical solutions for the transient characteristics of queueing models are useful for studying the finite-time properties of systems that are accurately represented by such models. Even if exact analytical transient results are not known, it would be useful to know, in some fashion, the rate and manner (e.g., monotonic or oscillatory) of convergence of the system to steady-state.

Analytical transient results are also valuable in the evaluation of alternative start-up strategies for simulations aimed at estimating steady-state parameters. Kelton and Law [6] and Kelton [5] present transient results for $M/M/s$, $M/E_k/1$, and $E_k/M/1$ queues and use them as benchmarks to evaluate alternative initialization methods for simulation of similar systems. Enlarging the range of benchmark models to include systems with multivariate state spaces and multiple servers with non-exponential service times served as the main motivation for this paper.

Another simulation-related area where transient results can be applied is the external control variates technique for variance reduction. When examining the transient behavior of a complex, analytically intractable system in a terminating simulation, a simpler system with known transient behavior is simulated alongside the system of interest. The results from the two systems, assuming the use of common random numbers, would be expected to be correlated, leading to a variance reduction. The larger the class of systems for which transient results are known analytically, the greater the similarity possible, leading to stronger variance reductions.

Known transient results can be classified according to whether or not the time measure is continuous (real time) or discrete (indexing by customer number). Continuous time results for various queues can be found in Saaty [15], Kleinrock [9], Odoni and Roth [12], Grassman [3], Pegden and Rosenshine [13], etc. Continuous time results describe the behavior of the parameters of the system, e.g., the number of customers in the system, at every point in time. Discrete time results, on the other hand, focus on the state of the

system at certain transition points, e.g., at the point of arrival of the n^{th} customer, or at the point of departure of the n^{th} customer.

The treatment of queues in discrete time is especially relevant from the standpoint of simulation. Standard measures of performance of general $GI/G/s$ queueing systems include the expected delay in queue (excluding service time), denoted by d, the expected wait in system (delay in queue plus service time), w, the expected number of customers in the queue, Q, and the number of customers in the system, L. Estimates for all of these quantities can be made directly during a simulation, but Carson and Law [2] have shown that it is preferable (in terms of achieving a reduction in variance) to estimate w , Q, and L indirectly from a direct estimate of d using the conservation equations: $w = d + E(S)$, $Q = \lambda d$ and $L = \lambda\big(d + E(S)\big)$ where $E(S)$ is the expected service time and λ is the arrival rate. For this reason, most simulation application and methodological research has focused on the delay-in-queue process, which clearly is in discrete time. Hence, analytical results for queueing systems in discrete time can be more easily related to simulation methodology. Morisaku [11], Kelton and Law [6], Kelton [5], Moore [10], Heathcote and Wiener [4], Stanford et.al. [16] and Bhat and Sahin [1] present discrete-time results for various queueing systems.

In this paper, we present discrete-time transient results for an $M/E_k/2$ queue, where E_k denotes the k-Erlang distribution. The state variable for this system has to be expressed as a tuple. The method of analysis used here admits arbitrary (deterministic) initial states for the queue; this allows a numerical evaluation of the effect of alternative initial states on the nature of convergence to steady-state, a general question of interest in simulation aimed at estimating steady-state parameters.

In Section 2 the analytical results for the $M/E_k/2$ system are derived. Section 3 reports some of the results based on the numerical evaluation of these results. Section 4 contains some conclusions, and the Appendix contains proofs of the results presented in Section 2.

2. THE $M/E_k/2$ SYSTEM

The arrival rate to the queue is denoted by λ, and the service times have a k-Erlang distribution with mean $1/\mu$. The service time distribution at both servers is assumed to be identical. It is also assumed that a customer professes no preference for one or the other of the servers, but enters service at the first available server. An arriving customer who finds both servers idle chooses one of the servers at random.

For purposes of analysis, each complete service period is modeled as k consecutive independent exponential stages, each at rate $k\mu$. The traffic intensity is $\rho = \lambda/(2\mu)$. Let $T_n, n = 1, 2, 3 \ldots$ be a random variable that represents the time of arrival of the n^{th} customer to the system. Let A_t be the number of service stages yet to be completed at server 1 at each point of continuous time $t, t \geq 0$, and let B_t be the total number of service stages present in the system at time t. (Server 1 is idle at time t if and only if $A_t = 0$.) The pair $\mathbf{X}(t) = (A_t, B_t)$ is sufficient to describe the state of the system at time t, since other quantities, such as the number of customers in queue at time t or the number of service stages yet to be completed at server 2 at time t, are functions of $\mathbf{X}(t)$. Note that $0 \leq A_t \leq k$ and $A_t \leq B_t$. The process $\mathbf{X}(t)$ renews at each point of continuous time t (i.e., the evolution of (A_s, B_s) for $s \geq t$ is a function only of (A_t, B_t), and is independent of all that happened in $[0, t)$), because both the interarrival time and the service stage distributions are exponential. $\mathbf{X}(t)$ is, in fact, a continuous time Markov chain. It is easily seen that the T_n's are stopping times for the process $\mathbf{X}(t)$, and $\mathbf{X}(t)$ therefore renews at each time T_n. In other words, $\mathbf{X}_n \equiv \mathbf{X}(T_n) = (A_{T_n}, B_{T_n})$ is a Markov chain. Similarly, the process $\mathbf{X}(t)$ also renews at each (random) epoch in time when a service stage is completed.

In the next Section, the transition probabilities for $\mathbf{X}_n$ are presented. In Section 2.2 various quantities of interest are derived in terms of these transition probabilities.

2.1 Transition probabilities

The process being analyzed is defined as $\mathbf{X}_n$ = system state at T_n, including the k stages of the n^{th} arriving customer, for $n \geq 1$. Let $\mathbf{x}_0 = (a_0, b_0)$ be the system state at time 0. The probabilities of interest are:

$$P_{\mathbf{x}_0}(\mathbf{x}; n) = P(\mathbf{X}_n = \mathbf{x} \mid \mathbf{X}_0 = \mathbf{x}_0), \qquad n = 1, 2, 3, \ldots$$

Note that the range of values $\mathbf{x}$ can take is determined by $\mathbf{x}_0$ and n. The first arrival occurs at T_1, which is exponentially distributed with mean $1/\lambda$.

The following Propostions, the proofs for which are given in the Appendix, are sufficient for the computation of the $P_{\mathbf{x}_0}(\mathbf{x}; n)'s$. For convenience, let $\alpha_i = \lambda/(\lambda + ik\mu)$ and $\beta_i = 1 - \alpha_i$, for $i = 1, 2$. α_1 is the probability that, given only one server is busy, the next customer arrives before the next service stage completion. Similarly, α_2 is the probability that, given both servers are busy, the next customer arrives before the next service stage completion. β_1 and β_2 are the probabilities of the complementary events.

(Note: $\mathbf{x}_0 = (a_0, b_0)$ and $\mathbf{x} = (a, b)$ below.)

Proposition 1: $P_{(0,0)}\big((k, k); 1\big) = 1$

Proposition 2:

(a) for $1 \leq b_0 \leq k$

$$P_{(0,b_0)}\big((k, k); 1\big) = {\beta_1}^{b_0}$$

(b) for $1 \leq b_0 \leq k$ and $k + 1 \leq b_1 \leq b_0 + k$

$$P_{(0,b_0)}\big((k, b_1); 1\big) = \alpha_1 {\beta_1}^{b_0 - b_1 + k}$$

This propostition represents the initial condition where server 1 is idle, server 2 is busy, and there is no queue.

Proposition 3:

(a) for $1 \leq b_0 \leq k$

$$P_{(b_0,b_0)}\big((k, k); 1\big) = {\beta_1}^{b_0}$$

(b) for $1 \leq j \leq b_0 \leq k$

$$P_{(b_0,b_0)}\big((j, j + k); 1\big) = \alpha_1 {\beta_1}^{b_0 - j}$$

This proposition represents the initial condition where server 1 is busy, server 2 is idle and there is no queue.

Proposition 4: For $1 \leq a_0 < b_0 \leq k+1$,

(a)

$$P_{(a_0,b_0)}\big((k,k);1\big) = \beta_1{}^{a_0} + \beta_1{}^{b_0-a_0} - 1 + \sum_{n_1=0}^{a_0-1}\sum_{n_2=0}^{b_0-1} \alpha_2(\beta_2/2)^{n_1+n_2}\binom{n_1+n_2}{n_1}$$

(b) for $1 \leq j \leq a_0$,

$$P_{(a_0,b_0)}\big((j,k+j);1\big) = \alpha_1\beta_1{}^{a_0-j} - \alpha_2(\beta_2/2)^{a_0-j}\sum_{n_1=0}^{b_0-a_0-1}(\beta_2/2)^{n_1}\binom{n_1+a_0-j}{n_1}$$

(c) for $1 \leq j \leq (b_0 - a_0)$,

$$P_{(a_0,b_0)}\big((k,k+j);1\big) = \alpha_1\beta_1{}^{b_0-a_0-j} - \alpha_2(\beta_2/2)^{b_0-a_0-j}\sum_{n_1=0}^{a_0-1}(\beta_2/2)^{n_1}\binom{n_1+b_0-a_0-j}{n_1}$$

(d) for $1 \leq a \leq a_0$ and $a+k+1 \leq b_0+k$,

$$P_{(a_0,b_0)}\big((a,b);1)\big) = \alpha_2(\beta_2/2)^{b_0-b+k}\binom{b_0-b+k}{a_0-a}$$

This Proposition holds for the initial conditions where server 1 and server 2 are both busy and there is no queue waiting for service. The Proposition also holds for the case when $b_0 > k+1$ as long as $b_0 - a_0 \leq k$.

Proposition 5: For $1 \leq a_0 \leq k$ and $b_0 > 2k$,

(a) for $b > 2k+1$,

$$P_{(a_0,b_0)}\big((a,b);1\big) = \alpha_2(\beta_2/2)^{b_0-b+k}\sum_{n=0}^{n'}\binom{b_0-b+k}{nk+c}$$

where:

$c = a_0 - a$ if $a_0 \geq a$ and $c = a_0 - a + k$ if $a_0 < a$

$n' = \lfloor (b_0 - b + k - c)/k \rfloor$ ($\lfloor x \rfloor$)denotes the greatest integer $\leq x$.)

and, if $b \leq b_0$, then $1 \leq a \leq k$,

else, if $b > b_0$, then $a \in \{a_j, j = 0,1,2,...,j'\}$, where

$j' = b_0 - b + k$

$a_j = a_0 - j$ if $a_0 > j$

$a_j = a_0 - j + k$ if $a_0 \leq j$.

(b) for $b \leq 2k+1$,

$$P_{(a_0,b_0)}\big((a,b);1\big) = (\beta_2/2)^{b_0-k-1} \sum_{a_I=1}^{k} \sum_{n=0}^{n'} \binom{b_0-k-1}{nk+c} P_{(a_I,k+1)}\big((a,b);1\big)$$

where:

$P_{(a_I,k+1)}\big((a,b);1\big)$ is found from Proposition 4.

$c = a_0 - a_I$ if $a_0 \geq a_I$ and $c = a_0 - a_I + k$ if $a_0 < a_I$, and,

$n' = \lfloor (b_0 - k - 1 - c)/k \rfloor$.

This proposition represents the initial conditions where both servers are initially busy, and there is at least one customer in the queue. The proposition also holds for the case with $b_0 \leq 2k$ as long as $b_0 - a_0 > k$.

Propositions 1 through 5 have covered the cases where $b_0 \leq k+1$ or $b_0 > 2k$. There are two subcases when $k+2 \leq b_0 \leq 2k$: (a) if $b_0 - a_0 \leq k$, then both servers are busy and the queue is empty initially, hence Proposition 4 applies, and (b) if $b_0 - a_0 > k$, then both servers are busy and there is one customer in queue initially, hence Proposition 5 applies.

Proposition 6:

$$P_{(a_0,b_0)}\big((a,b);n\big) = \sum_{j=k}^{b_0+(n-1)k} \sum_{i=1}^{k} P_{(a_0,b_0)}\big((i,j);n-1\big) P_{(i,j)}\big((a,b);1\big)$$

This follows directly from the fact that the process $\mathbf{X}(t)$ renews at every T. The quantities $P_{(i,j)}\big((a,b);1\big)$ can be found using Propositions 1 through 5.

2.2 Applications

If the probability mass function $P_{\mathbf{x}_0}(\mathbf{x}_n; n)$ of $\mathbf{X}_n$ is known, then formaulae for several standard measures of queueing performance can be developed easily. Some examples of these measures are : the expected total number of stages present in the system just after T_n, the expected number of customers present in the system just after T_n, the expected number of customers in queue just after T_n and the expected delay in queue for the n^{th} customer. Only the last performance measure is considered in detail below.

Let D_n denote the delay in queue of the n^{th} arriving customer. Then,

$$E_{\mathbf{x}_0}(D_n) = \sum_{\mathbf{x}_n} E(D_n \mid \mathbf{X}_n = \mathbf{x}_n) P_{\mathbf{x}_0}(\mathbf{x}_n; n)$$

The quantity of interest is therefore $E(D_n \mid \mathbf{X}_n = \mathbf{x}_n)$. Obviously, if $b_n \leq k+1$, then $E\big(D_n \mid (a_n, b_n)\big) = 0$. Also, if $(b_n - a_n) \leq k$, then $E\big(D_n \mid (a_n, b_n)\big) = 0$.

Suppose $\mathbf{x}_n$ is such that the n^{th} customer has to wait in queue before being served. The earliest time when this customer can enter service is after $b_n - 2k$ service stage completions, and the latest time when this customer can enter service is after $b_n - (k+1)$ service stage completions. Let $Q(a, b)$ be the probability that there will be exactly a service stages remaining at server 1 and $(b-a)$ service stages remaining at server 2 when the n^{th} customer enters service. It is clear that either a or $(b-a)$ or both will be equal to k.

Let EZ_j = Expected time for j service stage completions, when both servers remain busy throughout the period required for these j stage completions. The rate of service stage completions when both servers are busy is $2k\mu$, so $EZ_j = j/(2k\mu)$.

Conditioning on the total number of remaining service stages at server 1 and server 2 when the n^{th} customer just enters service, it can be seen that

$$\begin{aligned} E\big(D_n \mid (a_n, b_n)\big) &= \sum_{b=k+1}^{2k-1} EZ_{b_n - b}\big(Q(k, b) + Q(b-k, b)\big) + EZ_{b_n - 2k} Q(k, 2k) \\ &= \sum_{b=k+1}^{2k-1} \frac{(b_n - b)}{2k\mu}\big(Q(k, b) + Q(b-k, b)\big) + \frac{(b_n - 2k)}{2k\mu} Q(k, 2k) \end{aligned}$$

The formulae for $Q(a, b)$ are given and derived in the appendix. Finally,

$$\begin{aligned} E_{\mathbf{x}_0}(D_n) = \sum_{\mathbf{x}_n : (b_n - a_n) > k} P_{\mathbf{x}_0}(\mathbf{x}_n; n) \Bigg\{ & \sum_{b=k+1}^{2k-1} \big(Q(k, b) + Q(b-k, b)\big) \frac{b_n - b}{2k\mu} \\ & + Q(k, 2k) \frac{b_n - 2k}{2k\mu} \Bigg\} \end{aligned}$$

The cumulative distribution function of D_n can be calculated in a similar manner, using the $Q(a, b)$'s.

We empirically confirmed our results by simulating an $M/E_2/2$ queue, initially empty and idle, with $\rho = 0.7$ and $\lambda = 1$. The delay in queue of the first 25 customers was computed by averaging over 100,000 replications and the difference between the observed and theoretical values was less than 0.7%.

3. IMPLICATIONS FOR INITIALIZATION OF SIMULATIONS

In general, the goal of steady-state simulation is to estimate properties of the steady-state distribution. A judicious choice of the initial state can result in a reduction in the time required to reach steady-state. Clearly, the best choice for the initial state would be the steady-state value(s), but lack of knowledge precludes this choice.

Kelton [6] and Kelton and Law [7] present results for systems with a one-dimensional state-space (i.e., $\mathbf{X}_n$ is a scalar), and show that it is better to choose an initial state other than the popular empty-and-idle state to promote convergence to steady-state. The results in this paper extend the work to include two-dimensional state space systems, which require a bivariate initial state. The assumption of Erlang service time distributions lends realism to the model, since it appears that for many processes an Erlang-shaped histogram arises from the service time data.

Figure 1 shows a plot of $E_{\mathbf{x}_0}(D_n)$ as a function of n, for an $M/E_2/2$ system with traffic intensity $\rho = 0.7$ and $\lambda = 1.$, and $\mathbf{x}_0 = \{(0,0),(2,4),(2,6),(2,8),(2,10),(2,14),(2,20)\ \}$. Kelton and Law [7], Kelton [6] and Bhat and Sahin [1] took advantage of the special structure of the transition matrix for systems with a one-dimensional state-space and efficiently computed results for large values of n. We could not find a similar efficient computational algorithm for our results. Obtaining the numerical results, therefore, was computationally very expensive. * The value for the expected steady-state delay in queue for this system (dashed line) was found from tables in Hillier and Yu [5].

The convergence of $E_{\mathbf{x}_0}(D_n)$ to steady-state is highly dependent on $\mathbf{x}_0$ and is non-monotonic in some cases. Similar behavior was observed using discrete time analysis by Kelton [6], Kelton and Law [7] and Stanford et.al., [16], and using continuous time analysis by Grassman [3] and Odoni and Roth [12]. Odoni and Roth identified four types of

* We were limited to maximum values of $n \approx 30$. A program written in Fortran 77 took $\approx$ 200 mins. of CPU time to evaluate the delays of 30 customers on a Harris 800 computer and the VOS operating system.

behaviour: i) monotonic convergence from below, the function being concave in time, ii) initial decrease in the function, followed by a monotonic increase to the steady state value, iii) monotonic convergence from above, the function being convex in time, and, iv) monotonic convergence from above, the function being convex in time, but with a linear decrease initially. No claim was made that these types of behavior were exhaustive, and the basis for the characterization was empirical observations. The four types of behavior were observed for: i) an empty and idle or near-empty initial state, ii) initial state near steady-state value, iii) initial state > steady-state value, and iv) initial state $\gg$ steady-state value, respectively, as illustrated in Fig. 1.

It is clear, as in Kelton [5] and Kelton and Law [6] that the time for the expected delays to fall within a specified tolerance zone around the steady-state value is greatly influenced by the initial state. It is therefore advisable to investigate alternative initializations for simulation of systems such as these, in order to reduce bias and to shorten the length of the non-productive warm-up periods. A method similar to the one discussed in Kelton [5] that uses a series of preliminary runs, can be employed to find reasonable values for initialization of actual production runs.

4. CONCLUSIONS

Results for the transition probabilities for the transient, discrete time $M/E_k/2$ system have been presented in this paper. Further results using these transition probabilities have also been derived. It is shown that the results are in close agreement with previously published results, which were solely for single dimensional state-space systems.

A significant conclusion is that the method of analysis used here, though efficient for a scalar state-space appears computationally inefficient for higher dimensional state-spaces.

Simulation of models like the one investigated is greatly influenced by the choice of the initial state, thus warranting some experimentation to identify good starting conditions. Good starting conditions would result in a quicker approach to steady state and consequent reduction in computer run-time for the simulations.

The same method of analysis can be used for the $M/E_k/s$ system for $s > 2$, but the benefit of such an analysis is questionable, unless some method for reducing the computation time is found. Other multi-dimensional state-space models, e.g., $M/M/1/M/1$ queue can

possibly be studied in a similar manner, thereby providing more analytical test models for multivariate initialization heuristics for simulations.

APPENDIX

In the following proofs, when the initial state $\mathbf{x}_0$ is implicitly known, the subscript in $P_{\mathbf{x}_0}(\mathbf{x}; n)$ is dropped for notational convenience.

Proposition 1 is trivially true.

Proof of Proposition 2:

(a) Let A_1 denote the exponential arrival time of the first customer. Let $S_n, n = 1, 2, \ldots$ be a random variable that denotes the time of the n^{th} service stage completion. $P\big((k,k);1\big)$ is the probability that there will be b_0 service stage completions before the first arrival. Since $b_0 \geq 1$, there is a service stage in progress at time 0. Because this process is memoryless, it renews at time 0. The probability that the completion of the service stage in progress is before the first arrival is $k\mu/(\lambda + k\mu)$, since the two competing processes have exponential distributions. At the instant of completion of the first service stage, i.e., at S_1, the arrival process renews, and therefore $P(S_2 < A_1) = k\mu/(\lambda + k\mu)$. Continuing in this fashion,

$$P(S_1 < A_1, S_2 < A_1, \ldots, S_{b_0} < A_1) = \left(\frac{k\mu}{\lambda + k\mu}\right)^{b_0}$$

(b) $P\big((k,b_1);1\big)$ is the probability that there will be exactly $b_0 - (b_1 - k)$ service stage completions before the first arrival. At the instant of the $\big(b_0 - (b_1 - k)\big)^{th}$ service stage completion, the arrival process renews, and the probability that the arrival will be before the next service stage completion is $\lambda/(\lambda + k\mu)$. As in (a), the probability that there will be $b_0 - (b_1 - k)$ service stage completions before the first arrival is $\big(k\mu/(\lambda + k\mu)\big)^{b_0-(b_1-k)}$. The result follows immediately.

Proof of Propositon 3:

(a) Same as the proof of Proposition 2(a).

(b) Note that if all of the b_0 service stages in the system at time 0 are completed before the arrival of the first customer, then (a) applies. If, on the other hand, the number of service stage completions before the first arrival is i $(0 \leq i \leq b_0 - 1)$, then $\mathbf{X}_1 = (b_0 - i, b_0 - i + k)$. Define $j = b_0 - i$. Since $0 \leq i \leq b_0 - 1$, we have $1 \leq j \leq b_0$ and the result follows using Proposition 2(b).

Proof of Proposition 4:

(a) $P\big((k,k);1\big)$ is the probability that the first arriving customer finds the system empty. Define random variables C_1 , C_2 as follows:

$$C_i = \text{no. of service stage completions from server } i \text{ in } [0, T_1].$$

Note that as long as server i is busy, C_i is a Poisson counting process with rate $k\mu$. Further, C_1 and C_2 are independent. Initially, there are a_0 service stages remaining at server 1, and since $b_0 \leq k+1$, there are $b_0 - a_0$ service stages remaining at server 2. If the first arriving customer finds the system empty and idle, then $C_1 = a_0$ and $C_2 = b_0 - a_0$. Conditioning on T_1, the arrival time of the first customer, we get:

$$P(C_1 = a_0, C_2 = b_0 - a_0) = E_{T_1}\big(P(C_1 = a_0, C_2 = b_0 - a_0 \mid T_1)\big)$$

$$= \int_0^\infty P(C_1 = a_0, C_2 = b_0 - a_0 \mid T_1 = x)\ \lambda\ e^{-\lambda x} dx$$

Since C_1 and C_2 are independent processes,

$$P(C_1 = a_0, C_2 = b_0 - a_0 \mid T_1 = x) = P(C_1 = a_0 \mid T_1 = x)\, P(C_2 = b_0 - a_0 \mid T_1 = x).$$

Since C_1 cannot exceed a_0,

$$P(C_1 = a_0 \mid T_1 = x) = 1 - P(C_1 < a_0 \mid T_1 = x)$$

$$= 1 - \sum_{n_1=0}^{a_0-1} \exp(-k\mu x)(k\mu x)^{n_1}/n_1!$$

Similarly,

$$P(C_2 = b_0 - a_0 \mid T_1 = x) = 1 - P(C_2 < b_0 - a_0 \mid T_1 = x)$$

$$= 1 - \sum_{n_2=0}^{b_0-a_0-1} \exp(-k\mu x)(k\mu x)^{n_2}/n_2!$$

Finally,

$$P(C_1 = a_0, C_2 = b_0 - a_0) =$$

$$\int_0^\infty \left[1 - \sum_{n_1=0}^{a_0-1} e^{-k\mu x}(k\mu x)^{n_1}/n_1!\right]\left[1 - \sum_{n_2=0}^{b_0-a_0-1} e^{-k\mu x}(k\mu x)^{n_2}/n_2!\right]\lambda e^{-\lambda x}\, dx$$

The result follows after integration and simplification.

(b) $P\big((j, j+k); 1\big)$ is the probability that the first arriving customer finds j service stages remaining at server 1, and server 2 idle. This translates to $C_1 = a_0 - j$ and $C_2 = b_0 - a_0$. The rest of the proof follows the same lines as that of (a).

(c) $P\big((k, k+j); 1\big)$ is the probability that the first arriving customer finds server 1 idle and j service stages remaining at server 2. This translates to $C_1 = a_0$ and $C_2 = b_0 - a_0 - j$. The rest of the proof follows the same lines as that of (a).

(d) $P\big((a, b); 1)\big)$ is the probability that the first arriving customer finds both servers busy. This translates to $C_1 = a_0 - a$ and $C_2 = (b_0 - a_0) - (b - a - k)$. The rest of the proof follows the same lines as that of (a).

Proof of Proposition 5:

The initial state for this proposition implies that at time 0 both the servers are busy and there is at least one customer in the queue.

(a) The case $b > 2k + 1$ is examined here. Let C_1 and C_2 be as defined in the proof of Proposition 4(a). With c and n' as defined in the Proposition statement, it is easily seen that $\mathbf{X}_1 = (a, b)$ if and only if $C_1 + C_2 = b_0 - (b - k)$ and $C_1 \in \{c, c+k, c+2k, \ldots, c+n'k\}$. n' can be physically interpreted as the maximum possible number of customer exits from server 1 during the first interarrival time. The rest of the proof follows the same lines as that of Propostion 4(a). It should be noted that if $2k + 2 \leq b \leq b_0$, any value of $a \in \{1, 2, \ldots, k\}$ leads to a feasible $\mathbf{X}_1 = (a, b)$. If, however, $b > b_0$, then let $j' = b_0 + k - b$. Clearly, $0 \leq j' \leq k - 1$. The feasible values for a are $\{a_j : j = 0, 1, 2, \ldots, j' - 1\}$ where $a_j = a_0 - j$ if $a_0 > j$ and $a_j = a_0 - j + k$ if $a_0 \leq j$.

(b) If $b \leq 2k + 1$, then the process has to have passed through the state $(a_I, k + 1)$ at some $t \in [0, T_1)$, and $a_I \in \{1, 2, \ldots, k\}$. Designate this intermediate state as $\mathbf{X}_I$; $\mathbf{X}_I$ is entered at some random time $T_I < T_1$. T_I is a stopping time for $\mathbf{X}(t)$. Hence the process renews at time T_I. To reach the state $\mathbf{X}_I$ requires that $(b_0 - (k+1))$ service stage completions occur before the first arrival. Defining c as in the proposition statement it can be seen that in order to reach the state $(a_I, k+1)$ two conditions need to be satisfied: (a) $b_0 - (k+1)$ service stage completions occur before the first arrival and (b) the number of service stage completions from server 1 be c plus a multiple of k. Let $c_{tot} = b_0 - (k+1)$ and $n' = \lfloor (c_{tot} - c)/k \rfloor$. n' can again be physically interpreted as the maximum possible number of customer departures from server 1 during the first interarrival time. Then the number of service stage completions at server 1 must be in the set $\{c, k + c, \ldots, n'k + c\}$. Given that

there were c_{tot} service stage completions before the first arrival, the probability that any of these service stage completions is from server 1 is 1/2. Therefore,

$$P(c_1 \text{ service stage completions at server } 1 \mid c_{tot}) = \binom{c_{tot}}{c_1}(1/2)^{c_{tot}}$$

which is the symmetric binomial probability mass function.

Using this, we obtain

$$P\{b_0 - (k+1) \text{ completions before first arrival, } \mathbf{X}_I = (a_I, k+1)\}$$

$$= \left(\frac{2k\mu}{\lambda + 2k\mu}\right)^{c_{tot}} \sum_{n=0}^{n'} \binom{c_{tot}}{nk + c} (1/2)^{c_{tot}}$$

$$= \left(\frac{k\mu}{\lambda + 2k\mu}\right)^{c_{tot}} \sum_{n=0}^{n'} \binom{c_{tot}}{nk + c}$$

Now,

$$P(\mathbf{X}_1 = (a,b) \mid \mathbf{X}_0 = \mathbf{x}_0) = \sum_{a_I} P\{b_0 - (k+1) \text{ comp. before first arr., } \mathbf{X}_I = (a_I, k+1)\}.$$
$$P\{\mathbf{X}_1 = (a,b) \mid \mathbf{X}_0 = \mathbf{x}_0, \mathbf{X}_I = (a_I, k+1)\}$$

Since $T_I < T_1$ and because $\mathbf{X}(t)$ renews at T_I,

$$P\{\mathbf{X}_1 = (a,b) \mid \mathbf{X}_0 = \mathbf{x}_0, \mathbf{X}_I = (a_I, k+1)\} = P\{\mathbf{X}_1 = (a,b) \mid \mathbf{X}_0 = (a_I, k+1)\}.$$

The expression on the right hand side of the above equation can be evaluated using Proposition 4, and the result follows immediately.

Derivation of $Q(a,b)$'s

$Q(a,b)$ is defined as the probability that, given the n^{th} customer does not go into service immediately on arrival, there will be a and $b-a$ service stages remaining at server 1 and server 2 respectively when the n^{th} customer finally enters service. Though not explicit in the definition, $Q(a,b)$ clearly depends on $\mathbf{X}_n$. All arguments below are conditioned on the event that the n^{th} customer has to wait in queue. Also, service stage completions, unless otherwise qualified, mean service stage completions since the arrival of the n^{th} customer.

Case 1: $b_n > 2k$.

At the instant when a total of $b_n - 2k$ service stage completions have occured, the n^{th} customer is either the first in queue or has just entered service. Let $Q'(i, 2k)$ be the probability

that there will be i remaining sevice stages at server 1 when a total of $b_n - 2k$ service stage completions have occured. Both servers are busy during these $b_n - 2k$ service stage completions, and the rate of service stage completions is $2k\mu$. The probability that any of these completions is from a particular server is $1/2$. If $a_n \geq i$, then let $c = a_n - i$; else, if $a_n < i$, then let $c = a_n - i + k$. Let $n' = \lfloor (b_n - 2k - c) \rfloor$.

Then, using the same arguments as in the previous proofs,

$$Q'(i, 2k) = \sum_{n=0}^{n'} \binom{b_n - 2k}{nk + c} (1/2)^{b_n - 2k}.$$

Clearly, $Q(k, 2k) = Q'(k, 2k)$.

Let $Q'(i, j)$ be the probability that after $b_n - j$ service stage completions there are i remaining service stages at server 1, conditioned on the event that the n^{th} customer did not enter service after $b_n - j - 1$ service stage completions. Because both servers are busy during the $(b_n - j)^{th}$ service stage completion, the following equations hold:

For $j = 2k - 1$,

$$\begin{aligned}
Q'(1, 2k-1) &= (1/2)Q'(1, 2k) + (1/2)Q'(2, 2k) \\
&\vdots \\
Q'(k-2, 2k-1) &= (1/2)Q'(k-2, 2k) + (1/2)Q'(k-1, 2k) \\
Q'(k-1, 2k-1) &= (1/2)Q'(k-1, 2k) \\
Q'(k, 2k-1) &= (1/2)Q'(1, 2k)
\end{aligned}$$

Similarly, for $j = 2k - 2$,

$$\begin{aligned}
Q'(1, 2k-2) &= (1/2)Q'(1, 2k-1) + (1/2)Q'(2, 2k-2) \\
&\vdots \\
Q'(k-3, 2k-2) &= (1/2)Q'(k-3, 2k-1) + (1/2)Q'(k-2, 2k-1) \\
Q'(k-2, 2k-2) &= (1/2)Q'(k-2, 2k-1) \\
Q'(k, 2k-2) &= (1/2)Q'(1, 2k-1)
\end{aligned}$$

And in general, for j such that $k + 1 \leq j \leq 2k - 1$,

$$Q'(1, j) = (1/2)Q'(1, j+1) + (1/2)Q'(2, j+1)$$

$$\vdots$$

$$Q'(j-k-1,j) = (1/2)Q'(j-k-1,j+1) + (1/2)Q'(j-k,j+1)$$

$$Q'(j-k,j) = (1/2)Q'(j-k,j+1)$$

$$Q'(k,j) = (1/2)Q'(1,j+1).$$

It should be clear that for $k+1 \leq j \leq 2k-1$, $Q(k,j) = Q'(k,j)$ and $Q(j-k,j) = Q'(j-k,j)$. Since the $Q'(i,2k)$'s are known, the $Q(i,j)$'s can be calculated.

Case 2: $b_n \leq 2k$.

In this case, the n^{th} customer is the first in queue upon arrival. Therefore, $Q(i,j) = 0$ for $b_n \leq j \leq 2k$. Using the Q''s as defined in Case 1, set $Q'(a_n, b_n) = 1$, and $Q'(i, b_n) = 0$ for $i \neq a_n, i \in \{1, 2, \cdots, b_n - k, k\}$. Then, $Q'(i,j)$, for $k+1 \leq j \leq b_n - 1$, can be calculated recursively using the formulae in Case 1.

Finally we have as in Case 1, for $k+1 \leq j \leq b_n - 1$, $Q(k,j) = Q'(k,j)$, and $Q(j-k,j) = Q'(j-k,j)$.

ACKNOWLEDGMENTS

We would like to thank Electronic Data Systems Decision Technology Division for their support, and in particular express our appreciation to Ziv Barlach, Sam MacMillan and Michael Moore.

REFERENCES

1. Bhat, U.N., and Sahin, I. Transient behavior of queueing systems. $M/D/1$, $M/E_k/1$, $D/M/1$ and $E_k/M/1$. Tech. Memo. 135, Dept. of Operations Research, Case Western Reserve Univ., 1969.

2. Carson, J.S., and Law, A.M. Conservation equations and variance reduction in queueing simulations. *Oper. Res.* 28, 1980, 535 – 546.

3. Grassman, W.K. Transient and steady-state results for two parallel queues. *Omega* 8, 1980, 105 – 112.

4. Heathcote, C.R., and Winer, P. An approximation for the moments of waiting times. *Oper. Res.* 17, 1969, 175 – 186.

5. Hillier, F.S., and Yu, O.S. *Queueing Tables and Graphs.* North Holland. New York, 1981.

6. Kelton, W.D. Transient Exponential-Erlang Queues and Steady State Simulation. *CACM* 28, 1985, 741 – 749.

7. Kelton, W.D., and Law, A.M. The transient behavior of the $M/M/s$ queue, with implications for steady-state simulation. *Oper. Res.* 33, 1985, 378 – 395.

8. Kelton, W.D., and Law, A.M. A new approach for dealing with the startup problem in discrete event simulation. *Naval Res. Logist. Q.* 30, 1983, 641 – 658.

9. Kleinrock, L. *Queueing Systems, Vol. 1: Theory*, Wiley, New York, 1975.

10. Moore, S.C. Approximating the behavior of nonstationary single-server queues. *Oper. Res.* 23, 1975, 1011 – 1032.

11. Morisaku, T. Techniques for data truncation in digital computer simulation. Ph.D. dissertation, Dept. of Industrial and Systems Engineering, Univ. of Southern Calif., Los Angeles, 1976.

12. Odoni, A.R., and Roth, E. An empirical investigation of the transient behavior of stationary queueing systems. *Oper. Res.* 31, 1983, 432 – 455.

13. Pegden, C.D., and Rosenshine, M. Some new results for the $M/M/1$ queue. *Management Science* 28, 1982, 821 – 828.

14. Rothkopf, M.H., and Oren, S.S. A closure approximation for the non-stationary $M/M/s$ queue. *Management Science,* 25, 1979, 522 – 534.

15. Saaty, T.L. *Elements of Queueing Theory with Applications*, McGraw-Hill, New York, 1961.

16. Stanford, D.A., Pagurek, B., and Woodside, C.M. Optimal Prediction of Times and Queue Lengths in the $GI/M/1$ Queue. *Oper. Res.* 31, 1983, 322 – 337.

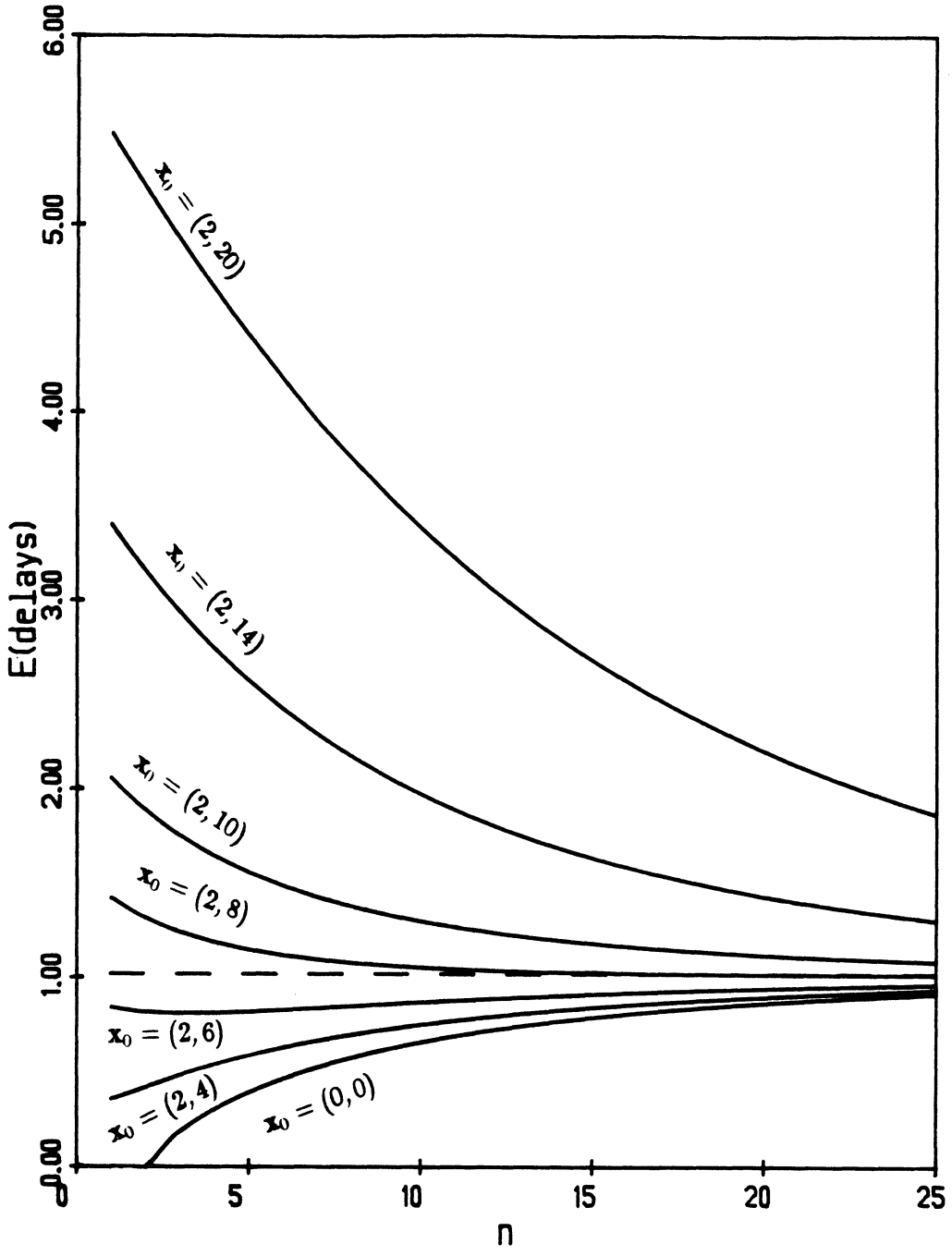

Figure 1. $E_{x_0}(D_n)$ for the $M/E_2/2$ Queue with $\rho = 0.7$