Probability in the Engineering and Informational Sciences, **16**, 2002, 167–184. Printed in the U.S.A.

HITTING TIME IN AN ERLANG LOSS SYSTEM

SHELDON M. ROSS
Department of Industrial Engineering and Operations Research
University of California
Berkeley, CA 94720
E-mail: ross@newton.me.berkeley.edu

SRIDHAR SESHADRI
Operations Management Department
Leonard N. Stern School of Business
New York University
New York, NY 10012
E-mail: sseshadri@stern.nyu.edu

In this article, we develop methods for estimating the expected time to the first loss in an Erlang loss system. We are primarily interested in estimating this quantity under light traffic conditions. We propose and compare three simulation techniques as well as two Markov chain approximations. We show that the Markov chain approximations proposed by us are asymptotically exact when the load offered to the system goes to zero. The article also serves to highlight the fact that efficient estimation of transient quantities of stochastic systems often requires the use of techniques that combine analytical results with simulation.

1. INTRODUCTION

We consider an Erlang loss system with C servers, namely the $M/G/C/C$ queuing system. The arrival rate is λ. The service times are independent and identically distributed (i.i.d.). A typical service time is represented as the random variable S and is assumed to have finite first and second moments that are denoted as $E[S]$ and $E[S^2]$, respectively. The distribution function of S is $F(\cdot)$. Define $\mu = 1/E[S]$. The random variable S_e has the equilibrium distribution of S, namely $F_e(y) = \mu \int_0^y F^c(x)dx$, where

 0269-9648/02 $12.50

$F^c(y) = 1 - F(y)$. We are interested in estimating the expected time to the first loss of a customer in this system, $E[T]$, when the system starts empty at time zero and the traffic is light. Define the load offered to this system as $\rho = \lambda/C\mu$. The stationary probabilities, p_j, that there are j customers in this system are given by

$$p_j = \frac{1}{K(C)j!}\left(\frac{\lambda}{\mu}\right)^j, \qquad j = 0,1,2,\ldots,C,$$

where $K(C)$ is the normalization constant given by

$$1 + \frac{\lambda}{\mu} + \frac{1}{2!}\left(\frac{\lambda}{\mu}\right)^2 + \cdots + \frac{1}{C!}\left(\frac{\lambda}{\mu}\right)^C;$$

see Ross [1].

Using simulation to estimate the expected time to the first loss is straightforward under heavy traffic because a loss takes place relatively quickly. This, however, is not the case under light and moderately light traffic conditions, which are the cases we will discuss. In such cases, the efficiency of a simulation can be improved both by using stratified sampling and by continuing to simulate beyond the first loss, using, in the latter case, the additional information of the time to empty the system to estimate the time to first lose a customer. Utilizing these ideas, we propose and compare three simulation techniques as well as two Markov chain approximations for the mean time of the first loss.

The first Markov chain approximation is an analytical approximation shown to be asymptotically exact as traffic goes to zero. The second approximation, which is also asymptotically correct, requires some (but not extensive) use of simulation. In particular, the simulation effort required does not grow with the decrease in traffic intensity. Thus, the second approximation technique is a hybrid one that combines analytical methods with simulation. We show that even with suitable modification to estimate the time to first loss, the three simulation techniques require exponentially increasing time as the traffic intensity decreases. In the crossover region between heavy and light traffic, the second Markov chain approximation works best. In light traffic, both approximations work well. Thus, the article, apart from highlighting new approximation and simulation techniques, also serves to highlight the fact that efficient estimation of transient quantities of stochastic systems often requires the use of techniques that combine analytical results with simulation.

2. MARKOV CHAIN APPROXIMATION

We shall define the transition matrix of a Markov chain that plays an interesting role in determining $E[T]$. For $i = 0,1,\ldots,C-1$, let W_i be a random vector comprising $i+1$ independent components with the first component distributed according to $F(\cdot)$ and the rest according to the distribution $F_e(\cdot)$. Also, let W_C be a random vector having C i.i.d. components distributed according to F_e. Let the

random variable A be exponentially distributed with mean equal to $1/\lambda$. Denote the kth smallest component of W_i as $W_i^{[k]}$. Define the transition matrix

$$P_{i,j} = \Pr\{W_i^{[i-j+1]} \le A < W_i^{[i-j+2]}\}, \qquad j = 1,2,\ldots,i,\ 0 \le i < C, \tag{1}$$

$$P_{i,i+1} = \Pr\{A < W_i^{[1]}\}, \qquad 0 \le i < C, \tag{2}$$

$$P_{i,0} = \Pr\{W_i^{[i+1]} \le A\}, \qquad 0 \le i < C, \tag{3}$$

$$P_{C,j} = \Pr\{W_C^{[C-j]} \le A < W_C^{[C-j+1]}\}, \qquad j = 1,2,\ldots,C-1, \tag{4}$$

$$P_{C,0} = \Pr\{W_C^{[C]} \le A\}, \tag{5}$$

$$P_{C,C} = \Pr\{A < W_C^{[1]}\}. \tag{6}$$

In other words, given that an arrival finds i customers whose remaining service times are i.i.d. $F_e(\cdot)$, $P_{i,j}$ gives the probability that the next arrival finds j customers in the system. Therefore, this is the conditional probability, given that an arrival to a stationary $M/G/C/C$ queue finds i customers, that the next arrival finds j. Regard the time for one transition in this chain to be the interarrival time for the $M/G/C/C$ system.

We shall prove a property of this Markov chain that motivated us to use it in approximating $E[T]$. Define an empty-to-empty cycle as the time taken by the Erlang loss system starting empty of customers at time 0 to return to the empty state again. The length of this cycle has finite expectation. Let this expectation be denoted as $E[T_{0,0}^Q]$. Let $E[T_{0,0}]$ be the expected time for the Markov chain starting from state 0 to return to state 0.

THEOREM 2.1:

$$E[T_{0,0}^Q] = E[T_{0,0}].$$

PROOF: Call an Erlang arrival (whether or not it enters the system) an i arrival if there are i in the system when it arrives, $i = 0,\ldots,C$, and let $Q_{i,j}$ denote the long-run proportion of i arrivals that are followed by a j arrival. Also, let Q_j denote the proportion of arrivals that are j arrivals. Then,

$$Q_j = \sum_i Q_i Q_{i,j}. \tag{7}$$

However, using the fact that the long-run conditional distribution (given the number in system) of remaining service times are i.i.d. according to the equilibrium service time, along with the fact that Poisson arrivals see time averages, it follows from the ergodic theorem that $Q_{i,j} = P_{i,j}$, where the $P_{i,j}$ are the transition probabilities of the Markov chain. However, then (7) states that the Q_j's satisfy the stationary equations of the Markov chain and, by uniqueness, must therefore equal them; that is, if the π_i's are the stationary probabilities of the Markov chain, then

$$\pi_j = Q_j,$$

as

$$Q_0 = \frac{1/\lambda}{E[T_{0,0}^Q]}$$

and, for the Markov chain with an exponential time between transitions,

$$\pi_0 = \frac{1/\lambda}{E[T_{0,0}]}.$$

Thus,

$$E[T_{0,0}^Q] = E[T_{0,0}].$$

■

Thus, ignoring the exponential times between transitions, what we have are two stochastic processes: both regenerative, with one being a Markov chain with transition probabilities, $P_{i,j}$, and the other a non-Markovian process with the same states and satisfying the property that the proportion of transitions out of i that are into j is also $P_{i,j}$. The preceding argument then shows that their limiting state probabilities are equal.

However, although the preceding is a nice motivating argument for using the Markov chain to approximate the first passage time, T, it does not really say much about whether expected first passage times (except from the regenerative state back to itself) are roughly equal for the two systems. Thus, it is not clear a priori see why the Markov chain approximation should be particularly good in very low traffic. To motivate this, consider the interval $[0, T]$. (Theorem 2.2 will also provide insight into the rate of convergence of the approximation to the true value.) Let $E[T_{0,C}]$ be the expected time to reach state C given that the Markov chain started in state 0. Let n_0 be the expected number of visits to state 0 during $T_{0,C}$. Let $E[T_{C,0}]$ be the expected time to reach state 0 given that the Markov chain is in state C. Let T_{HZ} be the time for the queue to empty after the first loss of a customer. Let N_j be the number of arrivals to the queue that saw j customers during $[0, T]$ and M_j be the number of arrivals that saw j customers during $[0, T + T_{\mathrm{HZ}}]$. Also, assume that $\sup_x E[S - x | S > x] = k$, where k is finite.

THEOREM 2.2:

$$\lim_{\lambda \to 0} \frac{E[T]}{E[T_{0,C}]} = 1.$$

PROOF: The proof of this theorem is in several steps and indirect. Because T and $T + T_{\mathrm{HZ}}$ are stopping times, by Wald's equation we can equate

$$\lambda E[T] = E[N_0] + E[N_1] + \cdots + E[N_C]. \quad \textbf{(8)}$$

Similarly,

$$\lambda(E[T] + E[T_{\mathrm{HZ}}]) = E[M_0] + E[M_1] + \cdots + E[M_C]. \tag{9}$$

We proceed to bound the difference between these two quantities (i.e., bound $\lambda E[T_{\mathrm{HZ}}]$). Assume that $C\rho < 1$. Modify the queuing system once a customer is lost, such that every customer that arrives after the loss is routed to the server that had the maximum remaining work at the time of loss. Assume that customers are no longer lost—there is infinite waiting space. It can be shown that the time to return to the empty state with this modification will be greater than the time to return to the empty state in the original system. In fact, to prove this, note that work for this server is at every instant larger than the work for any server in the unmodified system. The expected remaining work at this server after the loss of a customer is less than or equal to Ck, because the maximum of C nonnegative random variables is surely less than or equal to their sum. The utilization of this server is $C\rho$. By using the expected length of a busy period in an $M/G/1$ queue with exceptional first service time (see [1]), the expected time to return to the empty state for this server is less than $Ck/(1 - C\rho)$. Thus, the expected time for the system to empty is less than $Ck/(1 - C\rho)$. Therefore,

$$\lambda E[T_{\mathrm{HZ}}] \le \lambda \frac{Ck}{1 - C\rho}. \tag{10}$$

Using (8)–(10), we obtain that when $C\rho < 1$,

$$\left| \sum_{i=0}^{C} E(M_i - N_i) \right| \le \lambda \frac{Ck}{1 - C\rho}. \tag{11}$$

We next derive the equations that the $E[M_j]$'s should satisfy, namely

$$E[M_j] = \sum_{i=0}^{C-1} P_{i,j} E[M_i] + P_{C,j} E[M_C](1 - P_{C,C}), \quad j = 1,2,\ldots,C-1, \tag{12}$$

$$E[M_C](1 - P_{C,C}) = \sum_{i=0}^{C-1} P_{i,C} E[M_i]. \tag{13}$$

These equations can be justified as follows. We shall denote an arrival that finds i customers as an i arrival and an arrival immediately after an i arrival that finds j customers as an ij arrival. Let $N_{i,j}$ be the number of ij arrivals during $[0, T]$. Let R_i and $R_{i,j}$ be the number of i arrivals and ij arrivals during the interval $(T, T + T_{\mathrm{HZ}}]$, for $i = 0,1,2,\ldots,C-1$. Define R_C to be the number of C arrivals that are immediately followed by an arrival that sees less than C customers and define $R_{C,j}$ analogously. Let $\hat{M}_C$ be the number of C arrivals that are immediately followed by a customer that sees less than C customers. Clearly,

$$N_j = \sum_{i=0}^{C-1} N_{i,j}, \qquad j = 1,2,\ldots,C,$$

$$N_0 = 1 + \sum_{i=0}^{C-1} N_{i,0},$$

$$R_j = \sum_{i=1}^{C-1} R_{i,j}, \qquad j = 1,2,\ldots,C,$$

$$R_C = \sum_{i=1}^{C-1} R_{i,C} + N_C,$$

$$R_0 = 1.$$

In the equation for R_C, we have equated transitions into state C and transitions out of state C; thus, transitions from state C to state C (lost customers) are not counted. Define the cycle for the $M/G/C/C$ queuing system that starts with an empty queue, loses a customer, and ends with the subsequent return to the empty queue as a "Zero to Hit to Zero" or simply a ZHZ cycle. The last return to the empty state in a ZHZ cycle can be written as

$$R_0 = \sum_{i=1}^{C} R_{i,0}.$$

Now, notice that either we can count the last visit to state 0 (i.e., R_0) or the first visit, but not both visits in M_0. Thus, adding the above equations and using the facts that $R_i + N_i = M_i$ and that $R_{i,j} + N_{i,j} = M_{i,j}$, $0 \le i \le C-1$, we obtain

$$M_j = \sum_{i=0}^{C-1} M_{i,j} + R_{C,j}, \qquad j = 0,1,2,\ldots,C-1,$$

$$\hat{M}_C = \sum_{i=0}^{C-1} M_{i,C}.$$

The $M/G/C/C$ system regenerates after every ZHZ cycle. The proof of Theorem 2.1 showed that the fraction of arrivals that see i customers that are followed by an arrival that sees j customers is given by $P_{i,j}$. By the renewal reward theorem [1], the rate of i arrivals is given by $E[M_i]/(E[T] + E[T_{\mathrm{HZ}}])$ (i.e., the expected number of i arrivals during a ZHZ cycle divided by the expected length of a ZHZ cycle). Similarly, the rate of ij arrivals is equal to $E[M_{i,j}]/(E[T] + E[T_{\mathrm{HZ}}])$. The ratio of these two quantities is the fraction of arrivals that see i customers that are followed by an arrival that sees j customers. Thus, $E[M_{i,j}] = P_{i,j}E[M_i]$, $0 \le i < C$. By a similar argument, $E[\hat{M}_C] = E[M_C](1 - P_{C,C})$, and $E[R_{C,j}] = E[M_C](1 - P_{C,C})P_{C,j}$, $0 \le j < C$. These identities yield (12) and (13).

Now, we proceed to examine the Markov chain approximation. Consider the equations

$$\begin{aligned} x_j &= \sum_{i=0}^{C-1} x_i P_{i,j}, \qquad j = 1,2,\ldots,C-1, \\ x_0 &= 1 + \sum_{i=0}^{C-1} P_{i,0} x_i. \end{aligned} \tag{14}$$

We note that (a) when the Markov chain starts out in state 0, x_j is the expected number of time periods spent in state j before entering C, and (b) that (14) can be proven by use of Wald's equation. Thus, the sum $x_0 + x_1 + \cdots + x_{C-1}$ represents the expected number of transitions minus 1, until absorption in state C, given that the Markov chain is started in state 0. (The minus 1 accounts for the fact that the final transition to state C is not counted in the sum $x_0 + x_1 + \cdots + x_{C-1}$. The additional one in the expression for x_0 accounts for the fact that the chain starts in state 0.) For the final step in the proof, let e denote a column vector of dimension C whose components are equal to 1, let e_j denote the unit vector again of dimension C that has a 1 as its jth element and the rest of its elements equal to zero, let $(\ \ )_i$ denote the ith component of a vector, and let Q denote the $C \times C$ matrix whose elements are $P_{i,j}$, $i = 0,1,2,\ldots,C-1, j = 0,1,2,\ldots,C-1$. Let e^T and Q^T denote the transposes of the corresponding matrices. Let $(P_{C,j})$ denote a column vector whose components are $P_{C,0}, P_{C,1},\ldots,P_{C,C-1}$. Let $(I - Q^T)^{-1}$ denote the inverse of the matrix $(I - Q^T)$. Observe that this matrix is nonnegative due to its expansion as $I + Q^T + (Q^T)^2 + \cdots$. This expansion converges because the matrix Q does not include the last row and column of the matrix P. From (14),

$$\sum_{j=0}^{C-1} x_j = e^T (I - Q^T)^{-1} e_1. \tag{15}$$

We also observe by referring to (14) that $e^T(I - Q^T)^{-1} e_j$ is the expected number of transitions minus 1 until absorption in state C when the chain is initially in state j. Therefore, from the lower triangular structure of the transition matrix, we infer that $e^T(I - Q^T)^{-1} e_1$ is greater than $e^T(I - Q^T)^{-1} e_j$ for $j > 0$. Thus,

$$Ce^T (I - Q^T)^{-1} e_1 > e^T (I - Q^T)^{-1} e. \tag{16}$$

From (12),

$$\begin{aligned} \left| \sum_{j=0}^{C-1} E[M_j] - e^T (I - Q^T)^{-1} e_1 \right| &\le e^T (I - Q^T)^{-1} \left| (e_1 - (P_{C,j})(1 - P_{C,C}) E[M_C]) \right| \\ &\le e^T (I - Q^T)^{-1} \left| (e_1 - (P_{C,j})(1 - P_{C,C})) \right| \\ &\quad + e^T (I - Q^T)^{-1} (P_{C,j})(1 - P_{C,C}) \frac{Ck}{1 - C\rho}, \end{aligned} \tag{17}$$

where we have used (10) to bound $E[M_C]$ with $1 + \lambda(Ck/(1 - C\rho))$. The first term on the right-hand side of (17) is less than or equal to $e^T(I - Q^T)^{-1}e \max\{1 - P_{C,0}, P_{C,1}, \ldots, P_{C,C-1}\}$. The second term is less than $e^T(I - Q^T)^{-1}e(Ck/(1 - C\rho))$.

From these observations, using (16) as well as the facts that $P_{C,0}$ and M_C go to 1 when λ goes to 0, we obtain that $e^T(I - Q^T)^{-1}|(e_1 - (P_{C,j})(1 - P_{C,C}))| + e^T(I - Q^T)^{-1}(P_{C,j})(1 - P_{C,C})(Ck/(1 - C\rho))$ divided by $E[T_{0,C}]$ also goes to 0. Thus using (11), (15), and (17), we obtain the theorem. ■

It is worth noting that when the service times have the New Better than Used (New Worse than Used) property, then, due to the structure of the Markov chain, the approximation gives a lower bound (upper bound) for the hitting time when the S_e's are replaced by S's in computing the transition matrix $(P_{i,j})$. (A nonnegative random variable H is said to be New Better than Used (New Worse than Used) if $(H - x|H > x) \leq_{\text{st}} H$ $(H - x|H > x) \geq_{\text{st}} H$); see [1].) To see this, assume that the arrival times are generated first and stored as $\{a_1, a_2, \ldots,\}$. Given this sequence of arrivals, the residual service times found by arriving customers are independent random variables. Assume that the service times have the New Better than Used property. When the first arrival takes place, replace the residual service times by independent service times drawn from $F(\cdot)$. Thereafter, let the system evolve as usual. The hitting time will be stochastically smaller in this system compared to the original system (given the sequence of arrivals); see, for example, [4]. Therefore, the expected hitting time will be smaller with this modification. By repeating this construction (i.e., induction over the a_i's), it follows that the expected hitting time will be smaller when the service times are replaced each time by S's. The proof for the New Worse than Used case is similar. It is also interesting that these bounds continue to hold even when the arrival process is not Poisson (e.g., see [3] for a method of establishing bounds under a "lack of anticipation" assumption).

The Markov chain approximation proves to be accurate for light traffic. The Markov chain approximation is, however, not quite as useful under other traffic conditions.

From (17), we can visualize that the rate of convergence to the true expected hitting time is rather slow. There are several ways in which the approximation $E[T] \approx x_0 + x_1 + \cdots + x_{C-1} + 1$ can be improved. We suggest a particular modification that proves to be robust under moderate traffic conditions. For the Markov chain, let n_0 be the expected number of times state 0 is visited during $[0, T_{0,C}]$ and let $E[T_{C,0}]$ be the expected time to reach state 0 starting from state C. Then, it can be shown that

$$\frac{n_0}{\lambda(E[T_{0,C} + T_{C,0}])} = p_0 \tag{18}$$

(i.e., the fraction of time the $M/G/C/C$ system is empty). In other words, when we look at the "zero to reach state C to return to zero state" cycle in the Markov chain, the fraction of transitions into state 0 during this cycle equals the fraction of time that the queue is empty. (This cycle is not identical to the ZHZ cycle in the $M/G/C/C$ system, but the expected fraction of time spent in each state during the two cycles are

identical—a proof of this can be constructed following the lines of Theorems 2.1 and 2.2.) We suggest the following approximation based on this insight:

$$\frac{E[T_{C,0}]}{E[T_{0,C}]} \approx \frac{E[T_{\mathrm{HZ}}]}{E[T]}, \tag{19}$$

namely that the ratio of the expected time to return to the empty state to the expected time to lose a customer are approximately equal for the queue and the Markov chain. Based on (19), we obtain

$$\frac{E[T_{0,C}] + E[T_{C,0}]}{E[T_{0,C}]} \approx \frac{E[T] + E[T_{\mathrm{HZ}}]}{E[T]}. \tag{20}$$

We know that the expected length of an empty-to-empty cycle for the $M/G/C/C$ queue is equal to $1/\lambda p_0$, namely the reciprocal of the rate at which customers see an empty queue. By an application of Wald's equation, the expected length of a ZHZ cycle should equal this quantity divided by the probability of losing a customer in an empty-to-empty cycle (denoted by p_{loss}), or

$$E[T] + E[T_{\mathrm{HZ}}] = \frac{1}{\lambda p_0 p_{\mathrm{loss}}}. \tag{21}$$

Using (18) and (21), we obtain

$$\frac{n_0}{(E[T_{0,C} + T_{C,0}])} = \frac{1}{(E[T] + E[T_{\mathrm{HZ}}])p_{\mathrm{loss}}}. \tag{22}$$

Finally, combining (20) and (22), we get

$$E[T] \approx E[T_{0,C}] \frac{1}{n_0 p_{\mathrm{loss}}}. \tag{23}$$

Note that as λ goes to zero, $n_0 p_{\mathrm{loss}}$ tends to 1 because at most one customer is lost before the queue empties. To see this, let n_0^q be the expected number of arrivals to the queue that see an empty system in a ZHZ cycle. From (21),

$$n_0^q = \frac{1}{p_{\mathrm{loss}}}. \tag{24}$$

We can show that $E[T_{C,0}]/E[T_{0,C}]$ goes to zero as λ becomes small. By an application of Wald's equation,

$$E[T_{0,C} + T_{C,0}] = n_0 E[T_{0,0}]. \tag{25}$$

Similarly,

$$E[T + T_{\mathrm{HZ}}] = n_0^q E[T_{0,0}^Q]. \tag{26}$$

Therefore, from Theorems 2.1 and 2.2 and (24)–(26),

$$\lim_{\lambda \to 0} \frac{n_0^q}{n_0} = 1$$

$$\Rightarrow \lim_{\lambda \to 0} n_0 p_{\text{loss}} = 1.$$

When the queue is lightly or moderately loaded, we can estimate p_{loss} quite quickly and accurately by simulation [i.e., estimate the probability of losing a customer in an empty-to-empty cycle (using, e.g., Method 1 given in Sect. 3)]. Intuitively, the product $n_0 p_{\text{loss}}$ should be greater than 1 when $E[T_{0,C}]$ overestimates the expected hitting time and should be less than 1 when it underestimates the expected hitting time. Thus, this correction.

3. SIMULATION

3.1. Method 1

We let B denote the total amount of time that the system has been at capacity by time T. Then, B is exponential with rate λ and can be used as a control variable (because it is clearly positively correlated with T and with N). Perhaps the easiest way to see that B is exponential is to consider its hazard rate function. Think of B as the lifetime of some item and suppose that the item has just reached age x (i.e., the total time at capacity is now equal to x). Then, the probability that the item dies before an additional small time h elapses (i.e., that the total time at capacity does not reach $x + h$) is just the probability that a new arrival comes within time h, namely λh. Therefore, the hazard rate function of B is λ, implying the result; see [2].

Let T_b be the time point at which the cumulative time at capacity is equal to b. The raw simulation estimator is

$$E_1 = T_B.$$

Method 1(a). A variance reduction possibility can also be obtained if we first generate the value of B, and, then, if $B = x$, we stop the simulation when the total time at capacity is equal to x. We can then make use of antithetic variables by generating B from a random number U and using $1 - U$ to obtain a second value of B. We can then run the simulation until the total time at capacity is the maximum of the two generated values of C and thus obtain two values of T.

3.2. Method 2

We know that

$$E[T] = \int_0^\infty E[T|B = b]\lambda e^{-\lambda b}\, db.$$

Now, if we shut off the arrivals when at capacity and let T_b be the time point at which the cumulative time at capacity is equal to b, we can estimate $E[T]$ by

$$E_2 = \int_0^\infty T_b \lambda e^{-\lambda b}\, db.$$

In essence, this is stratified sampling and it should work nicely in moderately heavy to heavy traffic. (In light traffic, it will probably take too long because accumulating a sufficiently large value of B in one simulation will take a long time.) The efficiency of this method can be improved by cutting off the above integral and switching to the first method. Therefore, choose a value b^* which need not be that large—say, $b^* = 2/\lambda$. Then, write

$$E[T] = \int_0^{b^*} E[T|B = b]\, db + E[T|B > b^*]e^{-\lambda b^*}.$$

Now, generate X, an exponentially distributed random variable with rate λ; set

$$b_0 = b^* + X$$

and estimate

$$E[T|B > b^*]$$

by T_{b_0}.

3.3. Method 3

Let T_e denote the time from the moment of the first lost customer until the system becomes empty, and let B_e be the amount of that time in which all servers are busy. We can write

$$\lambda p_C = \frac{1 + \lambda E[B_e]}{E[T_e] + E[T]}.$$

In this expression we have used the fact that the expected number of lost customers is the first lost customer plus the expected number of subsequent arrivals (until the system becomes empty) that find all servers busy. Thus, $E[T]$ can be obtained by a simulation in which a run does not stop when a customer is lost, but continues on until the system is again empty (in light traffic, this will not take much additional time). The resulting estimators of $E[B_e]$ and $E[T_e]$ will then yield, through the preceding equation, an estimator of $E[T]$; that is, if $\bar{B}_e$ and $\bar{T}_e$ are the averages of the generated B_e's and T_e's, then the estimator, call it E_3, of $E[T]$ is

$$E_3 = \frac{1 + \lambda \bar{B}_e}{\lambda p_C} - \bar{T}_e.$$

The new estimator is

$$E_3 = \frac{1}{\lambda p_C} + \frac{\bar{B}_e}{p_C} - \bar{T}_e.$$

Hence,

$$\begin{aligned}\mathrm{Var}(E_3) &= \frac{\mathrm{Var}(\bar{B}_e)}{p_C^2} + \mathrm{Var}(\bar{T}_e) - 2\,\mathrm{Cov}\left(\frac{\bar{B}_e}{p_C}, \bar{T}_e\right) \\ &< \frac{\mathrm{Var}(\bar{B}_e)}{p_C^2} + \mathrm{Var}(\bar{T}_e)\end{aligned}$$

since B_e and T_e from the same run are (intuitively) positively correlated. Hence, if the averages are based on n simulation runs, then

$$\mathrm{Var}(E_3) < \frac{\mathrm{Var}(B_e)}{np_C^2} + \frac{\mathrm{Var}(T_e)}{n} \approx \frac{\mathrm{Var}(B_e)}{np_C^2}.$$

Note, on the other hand, that if T is the time of the first lost customer, then because T is approximately exponential in light traffic,

$$\begin{aligned}\mathrm{Var}(T) &\approx E^2[T] \\ &\approx \frac{(1 + \lambda E[B_e])^2}{\lambda^2 p_C^2}.\end{aligned}$$

If $\mathrm{Var}(T_e)$ is not large, then $\mathrm{Var}(E_3)$ should be smaller than $\mathrm{Var}(T)/n$.

Method 3(a). If $\mathrm{Var}(E_3)$ is not smaller than $\mathrm{Var}(T)/n$, then we can use both E_3 and T, since a single simulation run gives T, B_e, and T_e.

4. NUMERICAL RESULTS

In this section, we present numerical results for the 10-server loss system. The first set of results (see Tables 1–3) were obtained using Erlang-8 service times. The second set of results were obtained using hyperexponential service times, generated using a mixture of exponentials with rates 0.9 and $\frac{1}{30}$ with probabilities 0.9 and 0.1, respectively. All of the results are scaled such that they correspond to an arrival rate of 1. In Method 2, we have cut off the integral at $b^* = 2/\lambda$. The computations were performed using the SUN SPARC stations network at the Stern School of Business.

From the results shown in Tables 1, 2, 4, and 5, we observe that direct simulation of the hitting time is efficient when the traffic is relatively heavy. Although it is true that when compared to directly simulating the hitting time, Methods 1 and 2 yield lower standard errors for a given number of replications, a corresponding price (higher CPU times) is paid for this gain in efficiency. (As the standard error decreases with the square root of the number of replications, the correct quantity to be used to

TABLE 1. Erlang-8 Service Time: Methods 1 and 2

	Direct Simulation			Method 1			Method 2		
Util. %	T	Std. Error[a]	CPU Time (s)	E_1	Std. Error[a]	CPU Time (s)	E_2	Std. Error[a]	CPU Time (s)
20	33,031.34	335.44	4,351.30	33,667.21	341.42	4,755.54	33,512.24	150.39	12,448.30
30	1,708.39	17.04	225.45	1,738.64	17.41	245.92	1,741.85	9.40	604.50
40	297.67	2.91	39.21	297.65	2.91	42.06	299.90	1.86	96.00
50	99.18	0.96	12.96	98.83	0.93	13.95	98.72	0.65	29.00
60	49.38	0.44	6.33	48.95	0.44	6.82	48.84	0.33	12.90
70	31.09	0.26	3.93	31.16	0.26	4.29	30.74	0.20	7.40
80	22.86	0.17	2.83	22.76	0.17	3.14	22.81	0.14	5.10
90	18.60	0.13	2.26	18.43	0.13	2.52	18.53	0.11	3.80
100	16.09	0.10	1.95	15.86	0.10	2.15	16.05	0.08	3.00

[a]Based on 10,000 replications.

TABLE 2. Erlang-8 Service Times: Method 3 and Markov Chain Approximations

	Method 3			Markov Chain Approximations				
Util. %	E_3	Std. Error[a]	CPU Time (s)	$E[T] \approx E[T_{0,C}]$	% Error wrt Method 2	$E[T] \approx$ Eq. (23)	% Error wrt Method 2	CPU Time (s)
20	33,238.51	60.92	4,524.18	32,892.69	1.85	33,336.86	0.52	10.73
30	1,745.70	4.64	235.47	1,772.76	1.77	1,757.48	0.90	29.08
40	296.63	1.49	47.06	320.75	6.95	298.17	0.57	82.46
50	98.68	1.85	33.08	113.11	14.58	105.24	6.60	237.17
60	51.40	4.06	57.67	59.41	21.65	56.79	16.28	390.26
70	31.30	9.94	128.64	39.63	28.88	38.704	25.87	308.46
80	22.34	24.79	303.43	30.3	32.83	29.71	30.26	215.00
90	12.47	59.13	691.42	25.13	35.60	25.02	34.98	155.11
100	2.00	132.14	1,477.70	22.10	37.73	22.05	37.42	119.24

[a]Based on 10,000 replications.

TABLE 3. Erlang-8 Service Times: Intermediate Results for Method 3

Util. %	p_C	B_e	T_e	Var(B)	Var(T_e)	Cov(B, T_e)
20	3.79586E−05	0.1310	3.6349	1.34E−02	7.50	1.44E−02
30	0.000810388	0.1437	6.7783	1.57E−02	34.76	5.51E−02
40	0.005307549	0.2172	13.8726	3.96E−02	173.47	1.053
50	0.018384570	0.7087	29.6919	0.474458	833.64	16.04
60	0.043141838	3.0137	65.1534	8.708772	4,042.05	178.98
70	0.078740883	11.2667	140.4286	126.526716	19,492.34	1,553.23
80	0.121661064	37.6837	307.9779	1,423.543923	95,348.38	11,617.66
90	0.167963226	110.2529	655.6870	12,188.343940	430,465.56	72,374.06
100	0.214582343	284.0324	1,323.9179	80,403.295580	1,744,879.30	374,454.33

compare the efficiency of different methods is the product of the standard error and the square root of the CPU times.)

Method 2 appears to be efficient under moderate loads. In relatively light traffic conditions, Method 3 is the most efficient. However, it is not as efficient when the traffic is moderate or heavy. This is because, as can be seen from Table 3, the variance of the time it takes the system to empty after the first loss takes place (i.e., T_e) grows rapidly as the offered load increases. Moreover, the simulation of T_e consumes a large amount of CPU time when the traffic is moderate to heavy.

We do not report results for Methods 1(a) and 3(a). For the examples studied by us, these methods offer 5–10% reduction in the standard error without additional effort (over Methods 1 and 3, respectively). This does not change the relative attractiveness of the three methods, namely direct simulation is preferred under heavy traffic, Method 2 under moderate traffic, and Method 3 or 3(a) under light traffic.

Both Markov chain approximations are tabulated in Tables 2 and 5. Simulation is necessary to compute the transition matrix (1)–(6). In our examples, it is relatively simple to generate a random variable having the equilibrium distribution because they have either the Erlang or the hyperexponential distribution. It is a harder task when an analytical expression is unavailable for the equilibrium distribution. In the case when the service time distribution has finite support, the use of the rejection technique is suggested; see Ross [2]. If care is taken to scale the service time such that its mean is not too small, then the simulation effort which is proportional to the maximum of $1/E[S]$ can be kept small.

The relatively simple approximation, $E[T_{0,C}]$, is surprisingly accurate for very light traffic. When this approximation is corrected by use of the estimate of the probability of losing a customer in an empty-to-empty cycle [see (23)], it proves to be robust, even for moderate traffic conditions. The CPU time required to compute $E[T_{0,C}]$ ranges from 12 to 16 s. The additional time required to estimate p_{loss} [i.e., required for the approximation; (23)] is reported in Tables 2 and 5. Thus, when the

TABLE 4. Hyperexponential Service Times: Methods 1 and 2

	Direct Simulation			Method 1			Method 2		
Util. %	T	Std. Error[a]	CPU Time (s)	E_1	Std. Error[a]	CPU Time (s)	E_2	Std. Error[a]	CPU Time (s)
20	39,485.71	392.60	1,654.04	38,271.43	379.83	1,886.71	39,387.15	202.79	4,493.50
30	2,623.15	25.69	110.78	2,580.69	25.62	128.06	2,605.80	16.55	270.89
40	617.78	5.65	25.93	625.38	5.65	31.05	618.05	4.17	55.07
50	282.64	2.30	11.94	280.79	2.30	13.95	279.34	1.80	22.26
60	174.86	1.29	7.35	173.71	1.25	8.62	174.23	1.04	12.78
70	130.24	0.87	5.43	128.58	0.86	6.38	127.74	0.70	9.06
80	105.41	0.65	4.39	104.75	0.65	5.22	103.59	0.55	7.15
90	88.73	0.52	3.68	88.61	0.52	4.41	88.58	0.45	6.03
100	77.75	0.44	3.22	76.55	0.44	3.81	77.06	0.38	5.26

[a]Based on 10,000 replications.

TABLE 5. Hyperexponential Service Times: Method 3 and Markov Chain Approximations

	Method 3			Markov Chain Approximations				
Util. %	E_3	Std. Error[a]	CPU Time (s)	$E[T] \approx E[T_{0,C}]$	% Error wrt Method 2	$E[T] \approx$ Eq. (23)	% Error wrt Method 2	CPU Time (s)
20	39,716.22	230.27	2,032.81	35,689.70	9.39	38,015.23	3.48	4.09
30	2,658.73	23.66	150.88	2,024.22	22.32	2,620.11	0.55	11.47
40	620.07	9.60	52.12	384.27	37.83	707.86	14.53	34.42
50	288.90	10.04	52.80	140.88	49.57	320.45	14.72	62.70
60	175.55	16.51	86.90	75.71	56.55	129.47	25.69	64.34
70	142.17	32.64	169.29	50.71	60.30	98.40	22.97	59.08
80	124.91	65.51	331.82	38.72	62.63	68.38	33.99	54.15
90	124.13	137.58	667.20	31.82	64.08	51.02	42.40	50.86
100	113.93	273.73	1,297.03	27.70	64.06	41.07	46.71	49.04

[4]Based on 10,000 replications.

CPU time required to estimate this probability is factored in, the Markov chain approximations are seen to be both efficient as well as accurate for light traffic. (We have cut off the simulation when the probability of loss is within 10^{-7} of 1. This is the reason that the CPU times, reported in the last column of Tables 2 and 5, first increase and then decrease.) Similar findings were observed for 20- and 30-server loss systems.

In conclusion, we see that simulation Method 2 is the best to use at medium to high utilization levels. The second Markov chain approximation is recommended for use at low utilization levels.

Acknowledgment

The research of Sheldon M. Ross was supported by National Science Foundation grant DMI-9901053 from the University of California.

References

1. Ross, S.M. (1983). *Stochastic processes*. New York: Wiley.
2. Ross, S.M. (1990). *Simulation*, 2nd ed. Burlington, MA: Academic Press.
3. Shanthikumar, J.G. & Zazanis, M.A. (1999). Inequalities between event and time averages. *Probability in the Engineering and Informational Sciences* 13: 293–308.
4. Stoyan, D. (1983). In D.J. Daley (ed.) (with revisions), *Comparison methods for queues and other stochastic models*. New York: Wiley.