Simulation Methods for Response Times in Networks of Queues

Donald L. Iglehart
Department of Operations Research
Stanford University
Stanford, California 94305

Gerald S. Shedler
IBM Research Laboratory
San Jose, California 95193

ABSTRACT

We describe theoretically sound and computationally efficient estimation methods for "passage times" in certain closed networks of queues. Informally, a passage time is the time for a job to traverse a portion of the network. Such quantities are important in computer and communication system models where they represent job response times, and in this context, quantities other than mean values are of interest. From a single simulation run, the methods described here provide both point estimates and confidence intervals for general characteristics of response times.

1. INTRODUCTION

Networks of queues occur frequently in diverse applications. In particular, they are widely used in studies of computer and communication system performance as models for the interactions among system resources. This paper deals with mathematical and statistical methods for discrete event simulation of networks of queues. The emphasis is on methods for the estimation of general characteristics of "passage times" in closed networks. Informally, a passage time is the time for a job to traverse a portion of a network. Such quantities, calculated as random sums of queueing times, are important in computer and communication system models where they represent job response times.

The most obvious methodological advantage of simulation is that in principle it is applicable to stochastic systems of arbitrary complexity. In practice, however, it is often a decidedly nontrivial matter to obtain from a simulation information which is both useful and accurate, and to obtain it in an efficient manner. These difficulties arise primarily from the inherent variability in a stochastic system, and it is necessary to seek theoretically sound and computationally efficient methods for carrying out the simulation. It is fundamental for simulation, since results are based on observation of a stochastic system, that some assessment of the precision of results be provided. Assessing the statistical precision of a point estimate requires careful design of the simulation experiments and analysis of the simulation output. In general, the desired statistical precision takes the form of a confidence interval.

The estimation methods described here for passage times in networks of queues use the regenerative method [2], [3], [5] for analysis of simulation output. Based on a single simulation run, they provide (strongly consistent) point estimates and (asymptotically) valid confidence intervals for general characteristics of "steady state" (limiting) passage times.

The notion of a distinguished "marked job" is fundamental to the method for estimation of passage times described in Section 3. We arbitrarily select a job to serve as the marked job, and measure its passage times in the course of the simulation. In developing the marked job method, we take as the state vector of the network of queues the usual numbers-in-queue, stages-of-service state vector augmented by information sufficient to track the marked job. A key step in the development of the method is the definition of a related continuous time stochastic process. The starts of passage times for the marked job are the successive entrances of this process to a fixed set of states, and the terminations of such passage times are the successive entrances to another fixed set of states. We develop in this setting a ratio formula from which point estimates and confidence

CH1437-3/79/0221-0230$00.75
1979 Winter Simulation Conference

intervals can be obtained for quantities associated with the limiting passage time.

In Section 4 we give a second estimation method for passage times. With this decomposition method, observed passage times for all of the jobs enter into the construction of the point and interval estimates. This estimation method applies to the restricted but practically important class of passage times which are not complete circuits in a network.

2. NETWORKS OF QUEUES AND PASSAGE TIMES

We deal with closed networks of queues having a finite number of jobs (customers), N, a finite number of service centers, s, and a finite number of (mutually exclusive) job classes, c. At every epoch of continuous time each job is in exactly one job class, but jobs may change class as they traverse the network. Upon completion of service at center i a job of class j goes to center k and changes to class ℓ with probability $p_{ij,k\ell}$, where $P = \{p_{ij,k\ell}\}$ is a given irreducible Markov matrix. At each service center jobs queue and receive service according to a fixed priority scheme among classes; the priority scheme may differ from center to center. Within a class at a center, jobs receive service according to a fixed queue service discipline, e.g., first-come, first-served (FCFS). Note that in accordance with the matrix P, some centers may never see jobs of certain classes. Only one job can receive service at a center at a time, i.e., the centers are single servers. According to a fixed procedure for each center, a job in service may or may not be preempted if another job of higher priority joins the queue at the center.

Probabilistic assumptions

The marked job method discussed in Section 3 for passage time simulation applies to networks of this kind in which all service times in the network are mutually independent, and at a center have a distribution with a Cox-phase representation [1], i.e., consisting of a sequence of exponential stages; see Figure 1. We permit parameters of the service time distribution to depend on the service center, the class of job being served, and the "state" of the entire network.

Each service time distribution has its own finite number of stages, say n. Realization of a service time is as a sum of a random number (≥ 1) of exponentially distributed times. Within the jth stage $(1 \leq j \leq n-1)$, an amount of service exponentially distributed with parameter λ_j accrues; with probability $1 - b_j$ service is complete, and with probability b_j additional service exponentially distributed with parameter λ_{j+1} accrues. The density function of the resulting service time has rational Laplace transform

$$f^*(s) = \sum_{j=1}^{n} (1-b_j)a_j \prod_{k=1}^{j} \lambda_k/(\lambda_k+s) ,$$

where $a_1 = 1$ and for $1 < j \leq n$, $a_j = b_1 \ldots b_{j-1}$. The class of density functions having Laplace transforms of this form includes hyperexponential and mixtures of Erlang densities; see the Appendix of [4]. Note that we exclude the case of zero service times occurring with positive probability.

State vector definition

In order to characterize the state of the network at time t, we let $S_i(t)$ denote the class of the job receiving service at center i at time t, where $i = 1,2,\ldots,s$; by convention $S_i(t) = 0$ if at time t there are no jobs at center i. The classes of jobs serviced at center i ordered by decreasing priority are $j_1(i), j_2(i),\ldots, j_{k(i)}(i)$, elements of the set $\{1,2,\ldots,c\}$. Let $c^{(i)}_{j_1}(t),\ldots,c^{(i)}_{j_{k(i)}}(t)$ denote the number of jobs in queue at time t of the various classes of jobs serviced at center i, $i = 1,2,\ldots s$. For the queue length of jobs of various classes at the several centers, these state variables (together with the stages-of-service) would suffice. They are not adequate, however, to deal with general characteristics of passage times. An apparently minimal state vector augmentation is based on the concept of a marked job. The idea is to keep track of the position of an arbitrarily chosen marked job, and to measure its passage times during the simulation. It is convenient to think of the N jobs being completely ordered in a linear stack according to the following scheme. For $t \geq 0$, we define the vector $Z(t)$ by

$$Z(t) = (c^{(1)}_{jk(1)}(t),\ldots,c^{(1)}_{j_1}(t),S_1(t);\ldots;$$

$$c^{(s)}_{j_{k(s)}}(t),\ldots,c^{(s)}_{j_1}(t),S_s(t)). \qquad (1)$$

The linear job stack then corresponds to the order of components in the vector $Z(t)$ after ignoring any zero components. Within a class at a particular service center, jobs waiting appear in the job stack in FCFS order, i.e., in order of their arrival at the center, the latest to arrive being closest to the top of the stack. We denote by $N(t)$ the position (from the top) of the marked job in this job stack at time t.

With each job in the network, we associate a stage of service as follows. A job in service is in a particular stage of its service time distribution at that center; for such a job, this is the associated stage. For a job in queue at a center, the associated stage is that stage of service which is to be provided when the job next receives service; typically this is the first stage of service, but may be a subsequent stage if the job has been preempted. For $t \geq 0$, we define the vector $U(t)$ by

$$U(t) = (U_1(t),\ldots,U_N(t)) ;$$

where $U_j(t)$ is the stage of service associated with the jth job in the linear job stack, and take as the state vector of the network of queues the vector

$$X(t) = (Z(t),N(t),U(t)) . \qquad (2)$$

For any service center i that sees only one class of job ($k(i) = 1$), it is possible to simplify the state vector by replacing $C^{(i)}_{j\,k(i)}(t)$, $S_i(t)$ by $Q_i(t)$, the total number of jobs at center i.

Definition of passage times

Given a particular closed network of queues, we must specify the passage time of interest. This can be done in terms of the marked job by means of four subsets (A_1,A_2,B_1,B_2) of the state space, E, of the stochastic process $\underset{\sim}{X} = \{X(t):t \geq 0\}$. The sets A_1, A_2 [respectively B_1, B_2] jointly define the random times at which passage times for the marked job start [respectively terminate]. The sets A_1, A_2, B_1, B_2 in effect determine when to start and stop the clock measuring a particular passage time of the marked job.

It is convenient to introduce the jump times $\{\tau_n:n \geq 0\}$ of the process $\underset{\sim}{X}$. For $k,n \geq 1$, we require that the sets A_1, A_2, B_1 and B_2 satisfy:

if $X(\tau_{n-1}) \in A_1, X(\tau_n) \in A_2, X(\tau_{n-1+k}) \in A_1$ and $X(\tau_{n+k}) \in A_2$,

then $X(\tau_{n-1+m}) \in B_1$ and $X(\tau_{n+m}) \in B_2$ for some $0 < m \leq k$;

and

if $X(\tau_{n-1}) \in B_1, X(\tau_n) \in B_2, X(\tau_{n-1+k}) \in B_1$ and $X(\tau_{n+k}) \in B_2$,

then $X(\tau_{n-1+m}) \in A_1$ and $X(\tau_{n+m}) \in A_2$ for some $0 \leq m < k$;

These conditions ensure that the start and termination times for the specified passage time strictly alternate.

In terms of the sets A_1, A_2, B_1, and B_2, we define two sequences of random times, $\{S_j:j \geq 0\}$ and $\{T_j:j \geq 1\}$, where S_{j-1} is the start time of the jth passage time for the marked job and T_j is the termination time of this jth passage time. Assuming that the initial state of the process $\underset{\sim}{X}$ is such that a passage time for the marked job begins at $t = 0$, let

$$S_0 = 0$$

$$S_j = \inf\{\tau_n \geq T_j : X(\tau_n) \in A_2, X(\tau_{n-1}) \in A_1\}, \; j \geq 1$$

and

$$T_j = \inf\{\tau_n > S_{j-1} : X(\tau_n) \in B_2, X(\tau_{n-1}) \in B_1\}, \; j \geq 1 .$$

Then the jth passage time for the marked job is simply $P_j = T_j - S_{j-1}$, $j \geq 1$. For passage times that are complete circuits in the networks, $A_1 = B_1$ and $A_2 = B_2$; consequently $S_j = T_j$ for all $j \geq 1$.

3. THE MARKED JOB METHOD

Estimation of general characteristics of passage times by simulation of a closed network of queues can be based on the measurement of passage times for a typical job, the marked job discussed above. It is intuitively clear (and is shown in the Appendix of [6] that the sequence of passage times for any other job (as well as the sequence of passage times irrespective of job identity, in order of start or termination) converges in distribution to the same random variable as the sequence of passage times for the marked job. It follows that we can estimate general characteristics of passage times in the network by simulation of the process $\underset{\sim}{X}$. We use the regenerative method (applied to a stochastic process defined in terms of $\underset{\sim}{X}$) to obtain from a single simulation run point and interval estimates for passage time characteristics.

We denote by X_n, $n \geq 0$, the state of the system when the $(n + 1)$st passage time of the marked job starts. For $j \geq 1$, let P_j be the jth passage time for the marked job and take the quantity of interest in the simulation to be

$$r(f) = E\{f(P)\} , \qquad (3)$$

where f is a real-valued function. The quantity P is the limiting passage time, i.e.,

$$\lim_{n\to\infty} P\{P_n \leq x\} = P\{P \leq x\} ,$$

for all points x at which the right hand side of the equa-

tion is continuous. For example, to estimate $E\{P\}$, we take $f(p) = p$; to estimate $P\{P \leq t\}$, we take $f(p) = 1_{[0,t]}(p)$, where $1_{[0,t]}$ is the indicator function of the set $[0,t]$.

Algorithm 1. Marked job method

1. To serve as a return state, select a state of the system, i_0, at which a passage time of the marked job starts. Begin the simulation with $X(0) = i_0$.

2. Carry out the simulation of $\underset{\sim}{X}$ for a fixed number, n, of cycles (having random length) defined by the successive returns to the state i_0.

3. In each cycle measure all the passage times of the marked job and record these along with the number of passage times for the marked job in the cycle.

4. For $k \geq 1$, denote the number of passage times observed in the kth cycle by M_k and compute $Y_k(f)$, the sum of the quantities $f(P_j)$ for the passage times P_j in the kth cycle.

5. Take as a point estimate (based on n cycles) of $r(f)$ the quantity

$$\hat{r}_n(f) = \bar{Y}_n(f)/\bar{M}_n ,$$

where

$$\bar{Y}_n(f) = n^{-1} \sum_{k=1}^{n} Y_k(f)$$

and

$$\bar{M}_n = n^{-1} \sum_{k=1}^{n} M_k .$$

6. Take as a 100 $(1-2\gamma)$% confidence interval (based on n cycles) for $r(f)$ the interval

$$\hat{I}_n(f) = [\hat{r}_n(f) - z_{1-\gamma} s_n/(\bar{M}_n n^{1/2}),\ \hat{r}_n(f) + z_{1-\gamma} s_n/(\bar{M}_n n^{1/2})] .$$

Here $z_{1-\gamma} = \Phi^{-1}(1-\gamma)$, where $\Phi(\cdot)$ is the distribution function of a standardized (mean zero, variance one) normal random variable, and s_n is the quantity

$$s_n = [s_{11} - 2\hat{r}_n(f) s_{12} + (\hat{r}_n(f))^2 s_{22}]^{1/2}$$

where s_{11}, s_{22}, and s_{12} are the usual unbiased estimates for $\mathrm{Var}\{Y_k(f)\}$, $\mathrm{Var}\{M_k\}$, and $\mathrm{Cov}\{Y_k(f), M_k\}$, respectively.

The underlying stochastic structure

The force of the assumptions of Cox-phase service time distributions and Markovian routing in the closed networks of queues is that the process $\underset{\sim}{X}$ is an irreducible, positive recurrent continuous time Markov chain with a finite state space, E. We let X_n denote the state of the Markov chain X when the $(n+1)$st passage time of the marked job begins: $X_n = X(S_n)$, $n \geq 0$. The process $\{X_n : n \geq 0\}$ is a discrete time Markov chain (with state space A_2) which we assume to be irreducible and aperiodic. Next we observe that the process

$$\{(X_n, P_{n+1}) : n \geq 0\}$$

is a regenerative process in discrete time. The regenerative property guarantees ([11]) that as $n \to \infty$,

$$(X_n, P_{n+1}) \Rightarrow (X, P) ,$$

where $\Rightarrow$ denotes weak convergence (convergence in distribution). The random variables X and P are, respectively the limiting state vector for the network and the limiting passage time for the marked job. The goal of the simulation is estimation of $r(f) = E\{f(P)\}$, where f is a real-valued measurable function with domain $R_+ = [0, +\infty)$.

We denote by $L(t)$ the last state visited by the Markov chain $\underset{\sim}{X}$ before jumping to $X(t)$, and for $t \geq 0$ define

$$V(t) = (L(t), X(t)) . \qquad (4)$$

The process $\underset{\sim}{V} = \{V(t) : t \geq 0\}$ is the fundamental stochastic process of the passage time simulation. This process $\underset{\sim}{V}$ has a state space F consisting of all pairs of states (i,j), $i,j \in E$, for which a transition in $\underset{\sim}{X}$ from state i to state j can occur with positive probability. Since $\underset{\sim}{X}$ is an irreducible, positive recurrent Markov chain, so is $\underset{\sim}{V}$. We define subsets S and T of F according to

$$S = \{(k,m) \in F : k \in A_1,\ m \in A_2\}$$

and

$$T = \{(k,m) \in F : k \in B_1,\ m \in B_2\} \qquad (5)$$

and observe that the entrances of $\underset{\sim}{V}$ to S [resp. T] correspond to the starts [resp. terminations] of passage

times for the marked job.

The next step is to select a fixed element of S, which for convenience we designate state 0. We set $V(0) = 0$ and let $\{V_n : n \geq 0\}$ denote the embedded jump chain associated with the continuous time process $\underset{\sim}{V}$. We denote by $\{\gamma_n : n \geq 1\}$ the lengths in discrete time units of the successive 0-cycles (returns to the fixed state 0) for $\{V_n : n \geq 0\}$, and define $\delta_0 = 0$ and $\delta_m = \gamma_1 + \ldots + \gamma_m$, $m \geq 1$. Then the number of passage times for the marked job in the first 0-cycle of $\underset{\sim}{V}$ is

$$M_1 = \sum_{j=0}^{\delta_1 - 1} 1_{\{V_j \in S\}},$$

and the sum of the values of the function f for the passage times for the marked job in this cycle is

$$Y_1(f) = \sum_{j=1}^{M_1} f(P_j) .$$

We denote the analogous quantities in the kth 0-cycle by M_k and $Y_k(f)$. Since V is an irreducible, positive recurrent Markov chain, it is a regenerative process, and the pairs of random variables

$$\{(Y_k(f), M_k) : k \geq 1\} \qquad (6)$$

are independent and identically distributed (i.i.d). Then, provided that $P\{P \in D(f)\} = 0$, where $D(f)$ is the set of discontinuities of the function f in the definition of $r(f)$, and $E\{|f(P)|\} < \infty$, it follows that

$$r(f) = E\{Y_1(f)\}/E\{M_1\} . \qquad (7)$$

Given that the random variables in (6) are i.i.d. and the ratio formula of (7), the regenerative method applies and (from a fixed number n of cycles) provides the strongly consistent point estimate

$$\hat{r}_n(f) = \bar{Y}_n(f)/\bar{M}_n$$

for $r(f)$. The associated confidence interval is based on the central limit theorem

$$n^{1/2}\{\hat{r}_n(f) - r(f)\}/(\sigma(f)/E\{M_1\}) \Rightarrow N(0,1) , \qquad (8)$$

where $\sigma(f)$ is the variance of $Y_1(f) - r(f)M_1$ and $N(0,1)$ is a standardized (mean 0, variance 1) normal random variable. For an alternative derivation of the marked job method based on Markov renewal processes, see [7].

An Example

There are two service centers in the network of queues shown in Figure 2. Upon completion of exponentially distributed α service at center 1, in accordance with a binary random variable ψ, a job joins the tail of the queue in center 1 (when $\psi = 1$) or joins the tail of the queue in center 2 (when $\psi = 0$). After completion of exponentially distributed β service at center 2, the job joins the tail of the queue in center 1. Neither center 1 nor center 2 service is subject to interruption. Assume that both queues are served according to a first-come, first-served (FCFS) discipline, and consider the limiting passage time P defined as the time measured from entrance into the center 1 queue upon completion of a center 2 service until the job next enters the center 2 queue.

In this network there are two classes of jobs: class 1 jobs at center 1 and class 2 jobs at center 2. Since each center sees only one job class, by taking into account the fixed number of jobs in the network, we can define $Z(t)$ to be the number of jobs waiting or in service at center 1 at time t. Then the process $\underset{\sim}{X} = \{(Z(t), N(t)) : t \geq 0\}$, where $N(t)$ is the position of the marked job in the job stack at time t, has state space

$$E = \{(i,j) : 0 \leq i \leq N,\ 1 \leq j \leq N\} .$$

For the passage time P, the sets A_1 and A_2 defining the starts of passage times for the marked job are

$$A_1 = \{(i,N) : 0 \leq i < N\}$$

and

$$A_2 = \{(i,1) : 0 < i \leq N\} .$$

Similarly, the sets B_1 and B_2 defining the terminations of the passage time P are

$$B_1 = \{(i,i) : 0 < i \leq N\}$$

and

$$B_2 = \{(i-1,i) : 0 < i \leq N\}.$$

The process $\underset{\sim}{V} = \{(L(t), X(t)) : t \geq 0\}$, where $L(t)$ is the last state visited by the Markov chain $\underset{\sim}{X}$ before jumping to $X(t)$, has state space

$$F = \{(i,j,i+1,j+1):0\le i<N,\ 1\le j<N\} \cup \{(i,N,i+1,1):0\le i<N\}$$

$$\cup \{(i,j,i-1,j):0<i\le N,\ 1\le j<N\} \cup \{(i,i,i,1):1<i\le N\} .$$

The subsets of F defining the starts and terminations of passage times for the marked job are

$$S = \{(i,N,i+1,1):0 \le i < N\}$$

and

$$T = \{(i,i,i-1,i):0 < i \le N\} .$$

4. THE DECOMPOSITION METHOD

The marked job method of Section 3 is applicable to passage times in the general sense, i.e., whether or not the passage time is a complete circuit in the network. For general characteristics of passage times which are complete circuits, no other method for obtaining confidence intervals from a single simulation run is known to the authors. Note that since only the observed passage times for the marked job enter into the construction of point and interval estimates, there may be some loss of statistical efficiency as the price for obtaining confidence intervals.

In this section we concentrate on passage times through a subnetwork of a given closed network of queues, i.e., passage times which are not complete circuits in the network. For this class of passage times we develop an estimation method in which observed passage times for all the jobs enter into the construction of point estimates and confidence intervals. We consider closed networks of queues and passage times as in Section 2, but we make the additional assumption with respect to the sets S and T of (5) that $S \cap T = \phi$; this effectively rules out passage times which are complete circuits.

The basis of the decomposition method for estimation of passage times through a subnetwork is simulation of the network in random blocks defined by the terminations of certain passage times. The distinguished passage times are those that (i) terminate when no other passage times are underway, and (ii) leave a fixed configuration of the job stack defined by (1). These terminations serve to decompose the sequence of passage times for all of the jobs into independent and identically distributed blocks.

We denote by $\{P_n^0 : n \ge 1\}$ the sequence of passage times (irrespective of job identity) enumerated in order of passage time start. As before, we let f be a real-valued function with domain R_+, and the goal of the simulation is the estimation of

$$r^0(f) = E\{f(P^0)\} , \qquad (9)$$

where $P_n^0 \Rightarrow P^0$. Note that $P^0 = P$, the limiting passage time for (any) marked job.

Algorithm 2. Decomposition method

1. Select a configuration z^0 of the job stack at which a passage time terminates and there are no other passage times underway. Begin the simulation with this configuration of the job stack at time 0.

2. Carry out the simulation for a fixed number n of blocks defined by the successive terminations of passage times irrespective of job identity which leave the job stack in the (fixed) configuration z^0.

3. In each block, measure the passage times for all of the jobs and record these along with the number of passage times observed in the block.

4. Denote the number of passage times observed in the mth block by K_m^0 and compute $Y_m^0(f)$, the sum of the quantities $f(P_j^0)$ for the passage times P_j^0 in the mth block.

5. Based on n blocks, form point and interval estimates for $r^0(f)$ from the quantities $\{(Y_m^0(f),K_m^0):1 \le m \le n\}$ as in the standard regenerative method.

The underlying stochastic structure

We begin by labelling the jobs from 1 to N, and for $i = 1,2,\ldots,N$, denote by $N^i(t)$ the position of job i in the linear job stack at time t. Then in terms of the vector $Z(t)$ of (1), for $t \ge 0$ we set

$$X^0(t) = (Z(t),N^1(t),\ldots,N^N(t)) . \qquad (10)$$

The process $\underset{\sim}{X}^0 = \{X^0(t):t \ge 0\}$ is an irreducible, positive recurrent continuous time Markov chain with finite state space E^0. Next we let $L^0(t)$ denote the last state visited by the Markov chain $\underset{\sim}{X}^0$ before jumping to $X^0(t)$, and for $t \ge 0$ define

$$V^0(t) = (L^0(t),X^0(t)) . \qquad (11)$$

The process $\underset{\sim}{V}^0 = \{V^0(t):t \ge 0\}$ is the fundamental stochastic process of the simulation. Since $\underset{\sim}{X}^0$ is an irreducible, positive recurrent continuous time Markov chain, so is $\underset{\sim}{V}^0$. We denote the state space of $\underset{\sim}{V}^0$ by F^0 and define two subsets S^0 and T^0 of F^0 according to

$$S^0 = \{(z,n_1,\ldots,n_N,z',n_1',\ldots,n_N') \in F^0: \text{ for some } k,$$

$$(z,n_k) \in A_1 \text{ and } (z',n_k') \in A_2\}$$

and

$$T^0 = \{(z,n_1,\ldots,n_N,z',n_1',\ldots,n_N') \in F^0: \text{ for some } k,$$

$$(z,n_k) \in B_1 \text{ and } (z',n_k') \in B_2\} . \qquad (12)$$

The entrances of $\underset{\sim}{V}^0$ to the set S^0 correspond to the starts of passage times (irrespective of job identity) and the entrances of $\underset{\sim}{V}^0$ to the set T^0 correspond to the terminations. Thus from a simulation of the process $\underset{\sim}{V}^0$, it is possible to measure the passage times for all of the jobs.

Now consider $\{P_n^0: n \geq 1\}$ the sequence of passage times (irrespective of job identity), enumerated in order of passage time start. Recall that the goal of the simulation is estimation of

$$r^0(f) = E\{f(P^0)\} .$$

where $P_n^0 \Rightarrow P^0$, and f is a real-valued (measurable) function with domain R_+.

We carry out the simulation in random blocks of the process V^0 defined by the successive entrances of $\underset{\sim}{V}^0$ to a fixed set of states U^0. Entrances of $\underset{\sim}{V}^0$ to the set U^0 correspond to the terminations of passage times (irrespective of job identity) which occur when no other passage times are underway, and which leave a fixed configuration (z^0) of the job stack; see [10] for a formal definition of the set U^0. For convenience, we assume that $V^0(0) \in U^0$. Denoting by $\{\gamma_m^0: m \geq 1\}$ the lengths in discrete time units of the successive blocks (returns to the set U^0) for $\{V_n^0: n \geq 0\}$ the embedded jump chain associated with $\underset{\sim}{V}^0$, we define $\delta_0^0 = 0$ and $\delta_m^0 = \gamma_1^0 + \ldots + \gamma_m^0$, $m \geq 1$. (Note that the successive entrances to the set U^0 are not regeneration points for the process V^0). The number of passage times K_1^0 in the first block of the process $\underset{\sim}{V}^0$ is

$$K_1^0 = \sum_{j=0}^{\delta_1^0 - 1} 1_{\{V_j^0 \in S^0\}}$$

and we denote the analogous quantity in the mth block of $\underset{\sim}{V}^0$ by K_m^0. Note that within each block of $\underset{\sim}{V}^0$ defined by entrances to the set U^0, at least one passage time starts and terminates. Next, we let $Y_m^0(f)$ be the sum of the quantities $f(P_j^0)$ over the passage times P_j^0 in the mth block of $\underset{\sim}{V}^0$, e.g.,

$$Y_m^0(f) = \sum_{j=0}^{K_1 - 1} f(P_j^0) .$$

The key observation is that the sequence of pairs of random variables

$$\{(Y_m^0(f), K_m^0): m \geq 1\} \qquad (13)$$

are i.i.d. For the function f appearing in (9), let $D(f)$ denote the set of discontinuities of f. Assuming that $P\{P^0 \in D(f)\} = 0$, it follows that as $n \to \infty$,

$$f(P_n^0) \Rightarrow f(P^0) .$$

Finally, provided that $P\{P \in D(f)\} = 0$ and $E\{|f(P^0)|\} < \infty$, standard arguments (cf. [8]) establish the ratio formula

$$r^0(f) = E\{Y_1^0(f)\}/E\{K_1^0\} .$$

Given this ratio formula and the fact that the random variables in (13) are i.i.d., the regenerative method provides from a fixed number of blocks a strongly consistent point estimate and an associated confidence interval.

5. EFFICIENCY OF SIMULATION

For mean passage times, theoretical values can be obtained [10] for variance constants entering into the central limit theorems used in previous sections to obtain confidence intervals for passage time characteristics. These calculations provide a firm basis for an assessment of the statistical efficiency of the two methods.

For a comparison of the statistical efficiency of the marked job and decomposition methods where both apply, it is convenient to have a central limit theorem comparable to (8) but in terms of simulation time, t, rather than number of cycles, n. Let $m(t)$ be the number of passage times completed by time t, i.e., in the interval $(0,t]$. Then as $t \to \infty$,

$$t^{1/2}[(m(t))^{-1} \sum_{i=1}^{m(t)} f(P_i) - r(f)]/[(E\{\alpha_1\})^{1/2}\sigma/E\{M_1\}] \Rightarrow N(0,1) ,$$

where $E\{\alpha_1\}$ is the expected length of a 0-cycle in the continuous time process $\underset{\sim}{V}$. It follows that the quantity

$$e = (E\{\alpha_1\})^{1/2}\sigma/E\{M_1\}$$

is the appropriate measure of statistical efficiency for the marked job method and is independent of the state $0 \in S$

selected to form cycles. (For mean passage times, the function f is the identity function and σ^2 denotes the corresponding variance constant.) Similarly, the quantity

$$e^0 = (E\{\alpha_1^0\})^{1/2}\sigma^0/E\{K_1^0\},$$

where α_1^0 is the length of a block in V^0, is the appropriate measure of statistical efficiency for the decomposition method.

Table 1 gives theoretical values for simulation by the decomposition method for $r^0 = E\{P^0\}$ in the network of queues of Figure 2. Results are displayed for $N = 1$ to 6 jobs; λ_1 and λ_2 are the rate parameters for the exponential service times at centers 1 and 2, respectively, and p is the probability of feedback to center 1. We hold the value of $\lambda_1 = 1.0$ and $p = 0.75$ fixed, but vary λ_2. For $N = 2$ jobs (with $p = 0.75$, $\lambda_1 = 1.0$ and $\lambda_2 = 0.5$), the value of e^0 which measures the statistical efficiency of the decomposition method is 16.546. The corresponding value for the marked job method is 20.890. Thus, for these parameter values the decomposition method is approximately 21 percent more efficient than the marked job method. For $N = 4$ jobs, the decomposition method is 41 percent more efficient. Table 2 gives a comparison of the relative efficiency (e/e^0) of the marked job and decomposition methods for the same sets of parameter values as in Table 1.

6. CONCLUDING REMARKS

We have limited the discussion of estimation method for passage times to networks of queues having service centers which operate as single servers. The generalization to networks having multiple server service centers, however, is straightforward. To estimate passage times it is sufficient to incorporate into the linear job stack (and state vector definition) information carrying the class of job being serviced by each of the servers at a multiple server service center.

Estimation methods for passage times in finite capacity, open networks of queues are available. In [8], two formulations of the finite capacity constraint are considered and particular stochastic point processes associated with a Markov renewal process generate arrivals to the networks. These so-called Markov arrival processes facilitate the incorporation into computer and communication system models of departures from a Poisson arrival pattern. The basis for the estimation of passage times is the measurement of passage times for a sequence of marked jobs; these are typical jobs in the sense that the sequence of passage times for the marked jobs converges in distribution to the same random variable as do the passage times for all jobs.

The marked job method of Section 3 applies equally well to networks with stochastically nonidentical jobs. In such networks, jobs of each type have their own routing structure and service requirements, and in the case of finite capacity, open networks, (independent) arrival processes. Methods for the estimation of joint characteristics of passage times over the several job types are discussed in [9].

TABLE 1

Statistical efficiency of the decomposition method

	p=0.75 λ_1=1.0 λ_2=0.125		p=0.75 λ_1=1.0 λ_2=0.25		p=0.75 λ_1=1.0 λ_2=0.5	
N	E{P}	e^0	E{P}	e^0	E{P}	e^0
1	4.000	13.856	4.000	11.314	4.000	9.798
2	5.333	19.956	6.000	17.664	6.667	16.546
3	6.286	27.380	8.000	26.128	9.714	25.035
4	6.933	35.189	10.000	36.606	13.067	34.491
5	7.355	42.597	12:000	49.107	16.645	44.296
6	7.619	49.068	14.000	63.645	20.381	54.021

TABLE 2

Relative efficiency of marked job and decomposition methods

N	p=0.75 λ_1=1.0 λ_2=0.125	p=0.75 λ_1=1.0 λ_2=0.25	p=0.75 λ_1=1.0 λ_2=0.5
1	1.000	1.000	1.000
2	1.190	1.189	1.211
3	1.207	1.224	1.319
4	1.190	1.209	1.408
5	1.176	1.186	1.499
6	1.394	1.162	1.597

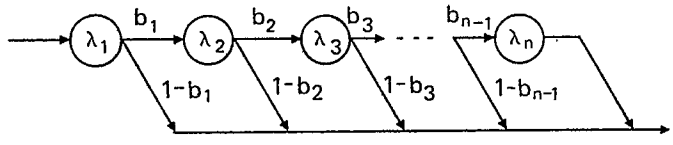

Figure 1. Exponential stage representation

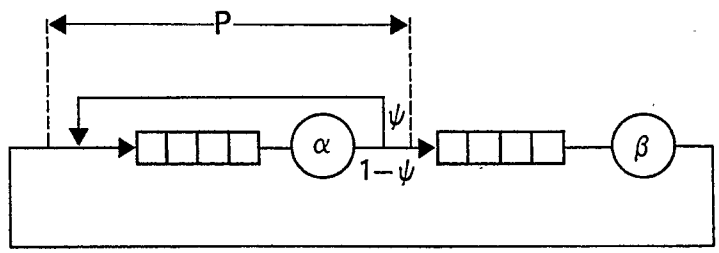

Figure 2. Closed network of queues

REFERENCES

1. Cox, D. R. A use of complex probabilities in the theory of stochastic processes. Proc. Cambridge Philos. Soc. 51, 313-319 (1955).

2. Crane, M. A. and Iglehart, D. L. Simulating stable stochastic systems: III, Regenerative processes and discrete event simulation. Operations Res. 23, 33-45, 1975.

3. Crane, M. A. and Lemoine, A. J. An Introduction to the Regenerative Method for Simulation Analysis. Lecture Notes in Control and Information Sciences, Vol. 4. Springer-Verlag, Berlin, Heidelberg, New York, 1977.

4. Gelenbe, E. and Muntz, R. R. Probabilistic models of computer systems - Part I (Exact results). Acta Informat. 7, 35-60 (1976).

5. Iglehart, D. L. The regenerative method for simulation analysis. In Current Trends in Programming Methodology Vol. III: Software Engineering. K. M. Chandy and R. T. Yeh (eds.), 52-71. Prentice-Hall. Englewood Cliffs, New Jersey.

6. Iglehart, D. L. and Shedler, G. S. Estimation via regenerative simulation of response times in networks of queues. IBM Research Report RJ 1740. San Jose, California, 1976 (an earlier version of 7.).

7. Iglehart, D. L. and Shedler, G.S. Regenerative simulation of response times in networks of queues. J. Assoc. Comput. Mach. 25, 449-461 (1978).

8. Iglehart, D. L. and Shedler, G. S. Simulation of response times in finite capacity open networks of queues. Operations Res. 26, 896-914 (1978).

9. Iglehart, D. L. and Shedler, G. S. Regenerative simulation of response times in networks of queues with multiple job types. Acta Informat. 12, 159-175, (1979).

10. Iglehart, D. L. and Shedler, G. S. Regenerative simulation of response times in networks of queues: Statistical efficiency. IBM Research Report RJ 2587. San Jose, California, 1979. Submitted for publication.

11. Miller, D. R. Existence of limits in regenerative processes. Ann. Math. Statist. 43, 1275-1282 (1972).

12. Shedler, G. S. Regenerative simulation of response times in networks of queues, III: Passage through subnetworks. IBM Research Report RJ 2466. San Jose, California, 1979. Submitted for publication.