S. S. Lavenberg
D. R. Slutz

Regenerative Simulation of a Queuing Model of an Automated Tape Library

Abstract: Recently, techniques have been developed for estimating confidence intervals when simulating stochastic systems having a regenerative structure. These techniques are applied to the simulation of a queuing model of a computer system's automated tape library. Theoretical and practical issues related to the application of these techniques are addressed. An interesting feature of the automated tape library represented in the queuing model is that certain queues have finite capacity; when these queues are filled to capacity certain services are prevented from occurring. The regenerative techniques are used in conjunction with multiple comparison procedures to make statistically valid statements about the effect of the finite queue capacities on performance.

Introduction

Contention for the resources that comprise a computer system can have a significant impact on system performance. Networks of interconnected queues are commonly used to model this contention, but often a queuing model that represents the system in sufficient detail to be of interest in performance evaluation is not analytically tractable. One then faces the choice of developing a less detailed but analytically tractable model, applying analytic approximation techniques, or simulating the model. Simulation of a queuing model of the system is included in many computer performance studies, either as the main tool or to validate simpler models and approximation techniques.

The simulation of a stochastic system such as a queuing model is a statistical experiment. In order to draw meaningful conclusions from such an experiment it is necessary to make statistically valid statements about the outcomes of the experiment. Suppose, for example, a queuing model is simulated in order to estimate a response variable Q (e.g., the long-run average time spent queuing for service). In addition to obtaining a point estimate $\hat{Q}$ of Q, it is desirable to estimate a confidence interval for Q. An estimated *$100 \cdot \alpha\%$ confidence interval* for Q is an interval $(\hat{Q}_1, \hat{Q}_2)$ whose endpoints $\hat{Q}_1$ and $\hat{Q}_2$ are estimated via simulation and have the property that $\Pr\{\hat{Q}_1 < Q < \hat{Q}_2\} = \alpha$. (Note that $\hat{Q}$, $\hat{Q}_1$ and $\hat{Q}_2$ are random variables while Q is a number.) Thus, an estimated confidence interval carries with it a statement that the response variable is contained in the interval with a given probability.

This paper focuses on the application of recently developed techniques for estimating confidence intervals when simulating a class of stochastic systems called regenerative systems. These techniques, called regenerative simulation techniques, are applied to the simulation of a queuing model of an automated tape library which serves as the mass storage portion of a computer installation. *Multiple comparison procedures* are used in conjunction with the regenerative techniques to make statistically valid statements about the effect on tape library performance of changing the values of certain tape library parameters. For example, let Q_1, Q_2 and Q_3 denote the values of a response variable corresponding to three sets of parameter values and let $D_1 = Q_2 - Q_1$, $D_2 = Q_3 - Q_1$ and $D_3 = Q_3 - Q_2$ denote the three pairwise differences. A simple multiple comparison procedure yields the $100 \cdot \alpha\%$ confidence statement $\Pr\{\hat{D}_{11} < D_1 < \hat{D}_{12}, \hat{D}_{21} < D_2 < \hat{D}_{22}, \hat{D}_{31} < D_3 < \hat{D}_{32}\} = \alpha$ where $\hat{D}_{11}$, $\hat{D}_{12}$, $\hat{D}_{21}$, $\hat{D}_{22}$, $\hat{D}_{31}$ and $\hat{D}_{32}$ are estimated via the regenerative techniques. Thus, D_1, D_2 and D_3 are each contained in their respective estimated intervals with a given joint probability α.

In the next section of the paper the automated tape library is described and a model of the library consisting of an open network of interconnected queues is presented. For the purpose of this paper, i.e., to present an example of the application of regenerative simulation to a complex queuing model, the model incorporates several simplifying assumptions, but even so the model is not analytically tractable. The section on simulation

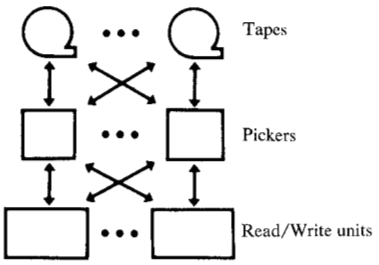

Figure 1 Automated tape library components.

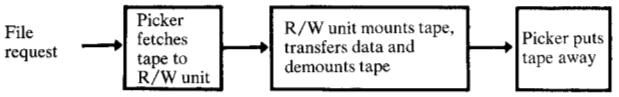

Figure 2 Services rendered for file request.

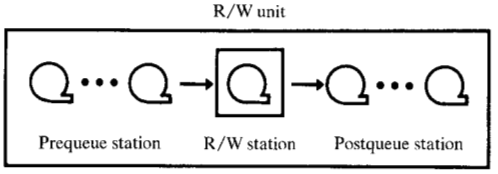

Figure 3 Structure of R/W unit.

methodology contains an exposition of the methods applied in simulating the tape library model. The method for estimating confidence intervals is not new, but the use of multiple comparison procedures in conjunction with the regenerative techniques is new. In addition, the theoretical and practical issues involved in applying the methodology to so complex a queuing model have not been explored elsewhere. In the experimental results section, simulation experiments designed to study the effect on performance of changing the values of certain parameters of the tape library model are described and the results of these experiments are presented and discussed. The last section contains concluding remarks.

The mathematical foundation for the simulation of regenerative stochastic systems was provided by Crane and Iglehart [1, 2]. Examples of the application of regenerative simulation techniques to queuing systems can be found in the literature, e.g., [2–8].

Library model

• *Introduction*

The system studied is the mass storage portion of a computer installation, here called the library. Libraries have traditionally consisted of reels of magnetic tape or boxes of cards that serve to hold seldom used or backup copies of data. Data in the library were maintained at relatively low cost and access to the data required human intervention and consequently time waits of several seconds or minutes. The evolution of large on-line data bases has resulted in a need for libraries that have reduced access times as well as low cost.

One proposal is an *automated tape library* (e.g., Ampex Terabit [9], IBM 3850 [10]), in which magnetic tapes or tape cassettes are the storage media, but the tapes are retrieved by mechanical means. Figure 1 illustrates the components of such a library. Data in the form of files are stored on a large number of magnetic tapes. The tapes are contained in fixed storage locations in the library. To access a file (read or write) requires removing the tape containing the file from its storage location, mounting the tape on a drive mechanism, here called an *R/W unit* (read/write unit), and positioning the head over the desired file. Due to cost considerations the number of R/W units is usually far smaller than the number of tapes, so mechanical devices called *pickers* are included to transport tapes to and from the R/W units.

An open queuing network model of an automated tape library is developed in this section to study the effect on performance of varying parameter values of the library. The performance measures studied are the *average response time* to satisfy an access request and the maximum rate at which requests can be satisfied (*maximum throughput*). The model developed is analytically intractable and will be numerically studied via simulation. Since, however, the main emphasis of this paper is to illustrate the careful application of simulation techniques, several simplifying assumptions are incorporated into the model. A detailed description of the operation of the automated tape library to be modeled is given next.

• *Operation of library*

When a request for file access arrives at the library the tape containing the file must first be located. If the tape is in its storage location, then the sequence of services illustrated in Fig. 2 must be rendered. (If the desired tape is already at an R/W unit, the initial picker service is not performed.) First, a picker must move the appropriate tape from its storage location to an R/W unit (*fetch service*). The R/W unit mounts the tape, positions the head to the desired file, transfers the data, and demounts the tape (*R/W service*). Lastly, the picker returns the tape to its storage location (*putaway service*).

Each R/W unit has a buffer area that can hold unmounted tapes. Figure 3 illustrates the structure of an R/W unit where the buffer area is represented by *prequeue* and *postqueue stations*. Fetched tapes are placed at the prequeue station by the pickers and tapes are re-

moved from the postqueue station by the pickers for putaway services. The *read/write station*, which represents the tape drive, selects a tape from the prequeue station, mounts the tape, and transfers the data. Following the data transfer, the tape is demounted and placed in the postqueue station and another tape is selected from the prequeue.

In practice the stations of an R/W unit have finite capacities. It is assumed that the prequeue station can hold at most C_1 tapes and the postqueue station can hold at most C_2 tapes. Additionally, it is assumed that the total number of tapes in the R/W unit (at both queue stations and at the R/W station) can never exceed C_3 tapes, where $C_3 \leq C_1 + C_2 + 1$. The condition $C_3 < C_1 + C_2 + 1$ allows representation of an R/W unit in which parts of both queue stations are realized in a common area. For example, the IBM 3850 [10] uses a carousel in the R/W unit to hold tape cartridges (Fig. 4). The carousel has three holes, each capable of holding one tape. The picker places a tape in the carousel and the carousel rotates first to bring the tape under the R/W station and then again to allow a putaway. Conceptually, the two holes not under the R/W station can serve as either prequeue or postqueue stations.

The automated tape library considered here uses a carousel in the R/W unit, which operates as follows. When a fetched tape is placed in the carousel and the R/W station is free, the tape immediately is rotated to the R/W station and an R/W service begins. However, when an R/W service is completed, the carousel is not rotated if the remaining holes all contain tapes that have received R/W services and are waiting for putaway services. The effect of this operation is that for a carousel with $C > 1$ holes, the prequeue station capacity is $C - 1$ and the postqueue station capacity is C (i.e., $C_1 = C - 1$, $C_2 = C_3 = C$). Thus, the carousel in Fig. 4 corresponds to $C_1 = 2$ and $C_2 = C_3 = 3$. If $C = 1$ then $C_1 = C_2 = C_3 = 1$.

The finite capacities introduce a blocking effect in that the starts of services are occasionally delayed even though the appropriate device is free. Thus, a fetch service is blocked if all prequeue stations are filled to capacity and an R/W service is blocked if the postqueue station of the same R/W unit is full. Since blocking tends to inhibit service, a degradation in performance is expected. The model to be developed will be studied to determine the effect on average response time and maximum throughput as the number of carousel holes C is varied. Determining that performance could be significantly improved by increasing C is of interest since this change involves relatively little additional cost.

- *Description of library model*

The automated tape library is modeled by the open network of queues shown in Fig. 5. The time sequence

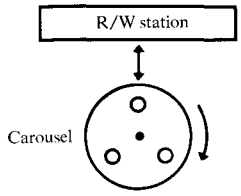

Figure 4 IBM 3850 R/W unit.

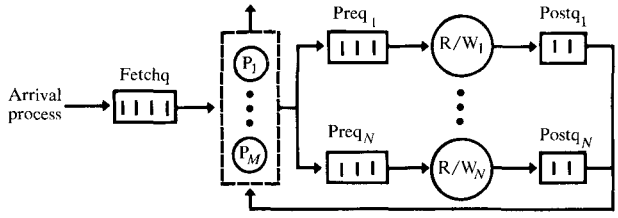

Figure 5 Structure of network queuing model for tape library.

of requests for file accesses is represented by an arrival process of customers at the network. Customers are of a single type, i.e., not identified by the file requested. The pickers and R/W stations are represented by servers in the model denoted by $P_1, \cdots, P_M$ and $R/W_1, \cdots, R/W_N$, respectively. Associated with each read/write server, R/W_i, are two queues, $preq_i$ and $postq_i$, to represent the prequeue and postqueue stations. Arrivals at the network join a conceptual queue called the *fetchq* (fetch queue). The fetchq has unlimited capacity while the preq's and postq's of each R/W unit are constrained by C_1, C_2 and C_3 as described above.

A customer arriving at the network is routed as suggested by the arrows in Fig. 5. A customer in the fetchq eventually moves to a picker server to receive a fetch service. (It is assumed that every file request requires a fetch service.) He is then placed in a preq and eventually moves to the associated R/W server for an R/W service. Next he is placed in the postq from which he eventually moves to a picker server for a putaway service. At the termination of the putaway service the customer departs from the network. Figure 6 summarizes the lifetime of a customer in the network. Also indicated is the *response time* for a request which is defined as the time interval between a customer's arrival and the completion of his R/W service. (Note that a customer receiving a service is considered to be at the server and not in a queue.)

The queuing discipline for each queue in the network is first-in, first-out. Algorithms are used to schedule fetch and putaway services for the pickers and to determine the routing to preq's of customers receiving fetch ser-

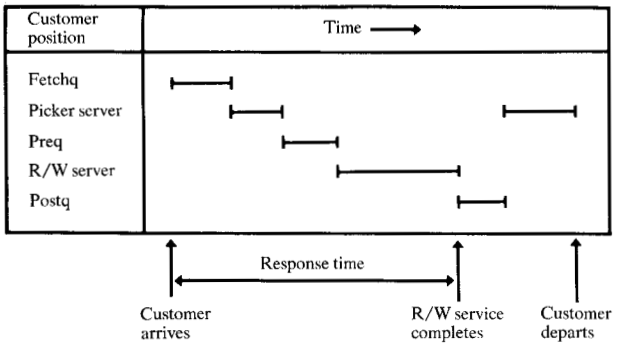

Figure 6 Lifetime of a customer in the network.

vices. It is convenient to call this set of algorithms the *picker scheduler*. An important characteristic of the picker scheduler is that the preq, say preq_i, that a customer receiving a fetch service will join is determined when the fetch service begins. While the fetch service is in progress it is said that the customer is *destined* for preq_i.

Let

$n(\text{preq}_i, t)$ = number of customers in preq_i at time t.
$n(\text{dest}_i, t)$ = number of customers destined for preq_i at time t.
$n(\text{postq}_i, t)$ = number of customers in postq_i at time t.
$n(\text{R/W}_i, t)$ = number of customers receiving an R/W service from server R/W_i at time t.

The picker scheduler is invoked whenever 1) a customer arrives at the fetchq, 2) an R/W service is completed, or 3) a picker service (fetch or putaway) is completed. Scheduling decisions are made in zero time and scheduled services start immediately. Picker services are never interrupted. Upon being invoked, the picker scheduler schedules all picker services that can be started using the following rules:

1. A fetch service is scheduled at time t if and only if there is an available picker, the fetchq is nonempty and there exists at least one eligible preq. Preq_i is eligible (to receive a customer) at time t if $n(\text{preq}_i, t) + n(\text{dest}_i, t) < C_1$ and $n(\text{preq}_i, t) + n(\text{dest}_i, t) + n(\text{R/W}_i, t) + n(\text{postq}_i, t) < C_3$. If A_1 is the set of eligible preq's at time t then the destination preq, say preq_d, is chosen as follows: if $|A_1| = 1$ ($|A_1|$ is the cardinality of A_1) then $\text{preq}_d = A_1$; otherwise form a sequence of sets $A_2 \subseteq A_1, A_3 \subseteq A_2, A_4 \subseteq A_3, A_5 \subseteq A_4$ in order as necessary until an A_i is reached such that $|A_i| = 1$, in which case $\text{preq}_d = A_i$ ($|A_5| = 1$ is guaranteed). A_2 is formed from A_1 by retaining each preq that minimizes $n(\text{preq}_i, t) + n(\text{dest}_i, t)$, A_3 is formed from A_2 by retaining each preq that minimizes $n(\text{R/W}_i, t)$, A_4 is formed from A_3 by retaining each preq that minimizes $n(\text{postq}_i, t)$, and A_5 is formed from A_4 by randomly choosing one preq with equal probabilities assigned to the members of A_4. An equivalent selection rule is to choose preq_d with equal probability from the set of preq's in A_1 that achieve

$$\min_i \{J^2[n(\text{preq}_i, t) + n(\text{dest}_i, t)] + Jn(\text{R/W}_i, t) + n(\text{postq}_i, t)\}$$

where $J > C_3$ is an arbitrary constant.
2. A putaway service is scheduled at time t if and only if there is an available picker, no fetch service can be scheduled, and there is a nonempty postq. The postq to be serviced is chosen with equal probability from the nonempty postq's that achieve

$$\max_i \{J^2 n(\text{postq}_i, t) + Jn(\text{R/W}_i, t) + n(\text{preq}_i, t) + n(\text{dest}_i, t)\}.$$

Observe that the picker scheduler first attempts to schedule a service using rule 1 and, if unsuccessful, an attempt is then made using rule 2. Thus, fetch services are given priority over putaway services. It is possible, however, for the picker scheduler to first successfully schedule a putaway service using rule 2 and then immediately successfully schedule a fetch service using rule 1. For example, consider a model with $C_1 = 2$ and $C_2 = C_3 = 3$. Suppose at time t there is at least one customer in the fetchq and at least one picker is available, but there are two customers in each preq and one customer receiving R/W service from each R/W server. Thus, there are no eligible preq's and all the postq's are empty. Suppose further that at time t a customer completes an R/W service and joins an empty postq, say postq_i. The picker scheduler, unable to schedule a fetch service, first schedules a putaway from postq_i. Instantaneously, the putaway service starts, an R/W service starts at R/W_i and preq_i becomes eligible. A fetch service destined for preq_i is then scheduled immediately (assuming there is another available picker). Even though the picker scheduler can make more than one decision at the same time, logically the decisions are made sequentially.

Each read/write server, R/W_i, takes customers from preq_i and places them in postq_i independent of the actions of other R/W servers (a separate data channel out of the library is assumed for each R/W unit). Thus an R/W service begins at R/W_i at time t if and only if $n(\text{preq}_i, t) > 0$, $n(\text{R/W}_i, t) = 0$ and $n(\text{postq}_i, t) < C_2$.

The following *probabilistic assumptions* are made for the model:

1. The arrival process is Poisson with rate λ arrivals/s.

2. Fetch and putaway service times for all pickers are mutually independent and identically distributed random variables. The common distribution is uniform on the interval a to b seconds $(0 \leq a < b)$.
3. R/W service times for all R/W servers are mutually independent and identically distributed random variables. The common distribution is uniform on the interval c to d seconds $(0 \leq c < d)$.
4. All interarrival times, picker service times, and R/W service times are mutually independent.

• *Discussion*

Table 1 summarizes the model parameters whose values are varied. The parameters a, b, c, and d of the service time distributions are fixed at 1, 3, 4, and 12 seconds, respectively. The chosen service time distributions are intended as a simple approximation to reality. Recall that a request for a file is modeled as an arrival but the name of the requested file is not included explicitly. The distribution of fetch and putaway service times is intended to reflect the probabilities over all tapes of the times to perform fetch and putaway operations. The times are known to have a nonzero minimum and a finite maximum; the uniform distribution was chosen for simplicity. The uniform distribution for R/W service time represents the case where the times to mount and demount a tape are constants, the time to transfer a file is constant (i.e., equal sized files) and a file is equally likely to start at any position on the tape. Clearly other distributions for the service times could be used to reflect the location of often used tapes in the storage area, unequal sized files and nonuniform file position on the tapes. The simulation techniques used in this paper are still applicable.

The chosen picker scheduler is intended to represent the intuitively appealing scheme of assigning a customer about to begin a fetch service to the R/W unit with the lightest load and removing a customer about to begin a putaway service from the R/W unit with the heaviest load. As stated, the effect on performance as the number of carousel holes C is varied will be studied. As C is increased the blocking effect presumably is reduced and the performance is expected to improve. (It is interesting to note that this increased performance may not occur for some picker scheduling algorithms. For example, a picker scheduler was considered that is identical to the one described above except that the destination preq was chosen with equal probability from the set of all eligible preq's. There was evidence from simulations that if $N > 1$ and λ is small, increasing C from 1 to 2 caused the average response time to increase! A possible explanation is that for $C = 1$, if two customers arrive at an empty system they necessarily will be assigned to different preq's. If $C = 2$ there is a chance that both will be destined for the same preq, the second customer incurring a larger delay than if he entered an empty R/W unit.)

Table 1 Library model parameters.

Parameter	*Range*	*Meaning*
λ	>0	arrival rate
M	≥ 1	number of picker servers
N	≥ 1	number of R/W servers
C	≥ 1	carousel capacity

• *Simulation programs*

Because of the complex structure of the library model (e.g., scheduling algorithms, finite capacity queues) and the non-exponential service time distributions, the library model is analytically intractable. Even if exponential service times were assumed, it would not be computationally feasible to analyze the model. Consequently, the model is studied via simulation. SIMPL/I [11] is chosen as the simulation language because of its suitability for queuing models, the ease of incorporating various stopping rules for the simulation and the ease of collecting and statistically analyzing data from the simulation runs. SIMPL/I is a PL/I based simulation language which is able to support asynchronous communicating processes. SIMPL/I processes are used to represent each picker server, each R/W server, the picker scheduler and the arrival process. Separate random number streams are used to generate the service times for each server and to generate interarrival times for the arrival process. For each simulation run the starting seeds are chosen to insure independent runs.

As described in the next section, different simulation techniques were used to estimate maximum throughput and average response time; two experimental simulation programs were written: Program 1, which estimates maximum throughput, consists of 1619 PL/I statements and Program 2, which estimates average response time, consists of 1842 PL/I statements. (SIMPL/I contains a PL/I preprocessor. The SIMPL/I source program for Programs 1 and 2 contained 265 and 399 statements, respectively.) The logical correctness of the simulation programs was verified primarily by running test cases. (The library model degenerates to an analytically solvable M/G/1 queue if $M = N = C_1 = C_2 = C_3 = 1$; see the last part of the experimental results section.) The PL/I programs were compiled with the optimizing compiler, but little additional effort was expended to further reduce the simulation time.

Simulation methodology

• *Estimation of confidence intervals*

Recently, several methods have been proposed for estimating confidence intervals for certain response variables when simulating stochastic systems having a regenerative structure [1, 2, 6]. Informally, a stochastic system is said to be *regenerative* if with probability one there exists an infinite sequence of increasing random times, called *regeneration points*, at which the system "stochastically restarts." The evolution in time of the system between successive regeneration points is called a *tour*, or cycle, and the stochastic behavior of the system during different tours is independent and identical. This underlying regenerative structure guarantees that for many response variables, estimates for the response variables based on a single run of the simulation are approximately normal if the run is sufficiently long and if certain random variables associated with a tour (e.g., the time duration of a tour) have finite first two moments. Furthermore, the variance of the estimates can be estimated either from independent replications of the simulation [6] or by observing a fixed number of tours during a single run of the simulation [1, 2]. Thus, provided a simulation run is sufficiently long, a theoretical basis exists for estimating confidence intervals for many response variables in regenerative stochastic systems.

The regenerative methods are applicable to simulating the tape library model if conjecture C1, which is presented later in this section, holds.

For the tape library model, let q_k denote the time spent in the fetch queue plus the time spent in the prequeue for the kth customer to arrive, and denote by r_k this customer's response time, i.e., the time from when he arrives until the completion of his R/W service. Assuming the limit exists with probability one, we wish to obtain point and confidence interval estimates for

$$Q = \lim_{n\to\infty} \sum_{k=1}^{n} q_k/n.$$

The average response time

$$R = \lim_{n\to\infty} \sum_{k=1}^{n} r_k/n$$

is related to Q by

$$R = Q + \mathrm{E}[T_{\mathrm{F}}] + \mathrm{E}[T_{\mathrm{R/W}}] \qquad (1)$$

where T_{F} and $T_{\mathrm{R/W}}$ are the fetch and R/W service times. Point and confidence interval estimates for R are obtained directly from the corresponding estimates for Q using (1).

Denote by $\mathscr{E}$ the event of a customer arriving at the empty system and assume event $\mathscr{E}$ occurs at time $t_0 = 0$. It is clear, due to the probabilistic assumptions made for this model in the previous section, that whenever event $\mathscr{E}$ recurs, the system stochastically restarts. In order to apply the regenerative simulation method to estimating Q, it is necessary that two additional conditions be satisfied:

1. Event $\mathscr{E}$ occurs infinitely often with probability one as the system evolves in time.
2. The random variables ν and σ, respectively the number of customers served during a tour and the sum of the times spent in the fetch queue and prequeues for all customers served during a tour, have finite first and second moments.

For certain open queuing systems (e.g., the M/G/1 queue [8]), these conditions hold if and only if the input rate is not too high (i.e., the traffic intensity is less than one) and the service times have finite fourth moments. A traffic intensity for the tape library model is defined next.

Suppose that instead of an arrival process of customers at the system, the tape library model has an infinite number of customers in the fetch queue at time zero (the fetch queue never empties). Call this the *saturated system*; let $D^*(t)$ denote the number of departures from the saturated system in the time interval $[0, t)$ and let

$$\lambda^* = \lim_{t\to\infty} D^*(t)/t,$$

(assuming this limit exists with probability one). The quantity λ^* is called the *saturated throughput* of the system and is itself an interesting performance measure. Define the traffic intensity ρ for the system to be $\rho = \lambda/\lambda^*$. The following *conjecture* is made for the tape library model:

C1. If $\rho < 1$ then Q is finite, the event $\mathscr{E}$ occurs infinitely often with probability one and $\mathrm{E}[\nu]$, $\mathrm{E}[\nu^2]$, $\mathrm{E}[\sigma]$ and $\mathrm{E}[\sigma^2]$ are finite. If $\rho > 1$ then Q is infinite and there is a positive probability the event $\mathscr{E}$ will not recur.

Let $\{t_k\colon k = 1, 2, \cdots\}$ denote the increasing sequence of random times at which event $\mathscr{E}$ recurs. The evolution of the system between t_{k-1} and t_k is the kth tour. Denote by ν_k the number of customers served during the kth tour and by σ_k the sum of the queuing times for all customers served during the kth tour ($\{\nu_k\colon k = 1, 2, \cdots\}$ and $\{\sigma_k\colon k = 1, 2, \cdots\}$ are sequences of i.i.d. random variables where for each k, ν_k is distributed as ν and σ_k is distributed as σ.) Let

$$\sigma(n) = \sum_{k=1}^{n} \sigma_k/n,$$

$$\nu(n) = \sum_{k=1}^{n} \nu_k/n,$$

$$Q(n) = \sigma(n)/\nu(n), \qquad (2)$$

and

$$V_1(n) = \sum_{k=1}^{n} [\sigma_k - \sigma(n)]^2/(n-1),$$

$$V_2(n) = \sum_{k=1}^{n} [\nu_k - \nu(n)]^2/(n-1),$$

$$V_{12}(n) = \sum_{k=1}^{n} [\sigma_k - \sigma(n)][\nu_k - \nu(n)]/(n-1),$$

$$V(n) = V_1(n) - 2Q(n)V_{12}(n) + [Q(n)]^2 V_2(n).$$

It can be shown (e.g., in the same manner as in [8]) as a direct consequence of conjecture C1 that if $\rho < 1$, then $\lim_{n\to\infty} Q(n) = Q$ with probability one (in which case the point estimate $Q(n)$ is said to be a strongly consistent estimate of Q) and that, for n sufficiently large,

$$I_Q(n, \alpha) = [Q(n) - \delta(n, \alpha), Q(n) + \delta(n, \alpha)] \tag{3}$$

is approximately a $100 \cdot \alpha\%$ confidence interval for Q where

$$\delta(n, \alpha) = \phi^{-1}[(1+\alpha)/2)][V(n)/n]^{\frac{1}{2}}/\nu(n),$$

$\phi(t) = (1/2\pi)^{\frac{1}{2}} \int_{-\infty}^{t} e^{-x^2/2}\, dx$ is the probability distribution function of a normal random variable having mean zero and variance one, and $\phi^{-1}(\cdot)$ is the inverse of the function $\phi(\cdot)$. (If $\alpha = 0.95$, corresponding to a 95% confidence interval, then $\phi^{-1}[(1+\alpha)/2] = 1.960$.) Thus, a point estimate for Q and an approximate confidence interval for Q can be computed based on observing values of σ and ν over n simulated tours. Point and confidence interval estimates for R, denoted respectively by $R(n)$ and $I_R(n, \alpha)$, are obtained by adding the sum of the mean fetch service time and mean R/W service time to $Q(n)$ in (2) and (3) respectively; i.e.,

$$R(n) = Q(n) + \mathrm{E}[T_{\mathrm{F}}] + \mathrm{E}[T_{\mathrm{R/W}}] \tag{4}$$

and

$$I_R(n, \alpha) = [R(n) - \delta(n, \alpha), R(n) + \delta(n, \alpha)]. \tag{5}$$

While conjecture C1 is strongly believed to hold for this model, the authors have been unable to prove its validity. Also, it is known that this conjecture is not true for some queuing models. For example, Whitt [12] shows that for the GI/G/s queue (s server queues with i.i.d. interarrival times and i.i.d. service times) if $s > 1$ the event $\mathscr{E}$ need not occur infinitely often with probability one if $\rho < 1$ unless there is a positive probability that an interarrival time exceeds a service time. (If a service time exceeds an interarrival time with probability one, then the system never empties with probability one.) Of course, the probability that an interarrival time exceeds a service time is positive if the arrival process is Poisson (M/G/s queue). The authors have assumed the validity of the conjecture when simulating the tape library model with Poisson arrivals since the probability that the system empties is then positive, although it may not be equal to one. If the arrival process is not Poisson, but the interarrival times are i.i.d. with distribution function having a density on the whole positive real line, then the probability that the system empties is again positive and the conjecture is believed to hold. Thus, regenerative simulation techniques are not necessarily restricted to systems with Poisson arrivals. The problem of determining the class of queuing models for which conjecture C1 holds is currently being investigated.

The saturated throughput λ^* is not known and must itself be estimated via simulation. The saturated system is not regenerative (except in special cases) so that the regenerative method is not applicable here. The saturated throughput is estimated by performing U independent replications of a simulation of the saturated system. Each replication is stopped when K customers have departed from the system. Let $\tau(K)$ denote the time at which the Kth departure from the system occurs and let $\lambda^*(K) = K/\tau(K)$. Let $\lambda_u^*(K)$ denote the value of $\lambda^*(K)$ observed on the uth replication, $u = 1, \cdots, U$. Then a point estimate of λ^* is given by

$$\lambda^*(K, U) = \sum_{u=1}^{U} \lambda_u^*(K)/U \tag{6}$$

and the variance of $\lambda^*(K)$ can be estimated by

$$V^*(K, U) = \sum_{u=1}^{U} [\lambda_u^*(K) - \lambda^*(K, U)]^2/(U-1).$$

By relying on the robustness of the t-statistic when the observations $\lambda_u^*(K)$ are non-normal (see Chapter 10 of [13]), an approximate $100 \cdot \alpha\%$ confidence interval for λ^* is given by

$$I^*(K, U, \alpha) = [\lambda^*(K, U) - \delta^*(K, U, \alpha), \lambda^*(K, U) + \delta^*(K, U, \alpha)], \tag{7}$$

where

$$\delta^*(K, U, \alpha) = \theta_{U-1}^{-1}[(1+\alpha)/2][V^*(K, U)/U]^{\frac{1}{2}},$$

$\theta_{U-1}(\cdot)$ is the probability distribution function of the t-statistic with $U-1$ degrees of freedom and $\theta_{U-1}^{-1}(\cdot)$ is the inverse of $\theta_{U-1}(\cdot)$. If $\alpha = 0.95$, corresponding to a 95% confidence interval, and $U = 10$ then

$$\theta_{U-1}^{-1}[(1+\alpha)/2] = 2.262.$$

According to conjecture C1, λ^* is the maximum (actually the supremum) input rate for which the average response time is finite. Let $D(t)$ denote the number of departures from the system with Poisson arrivals in the time interval $[0, t)$. It can be shown, as a consequence of C1, that if $\lambda < \lambda^*$ then $\lim_{t\to\infty} [D(t)/t] = \lambda$ with probability one, i.e., the output rate or throughput equals

the input rate. Thus, the saturated throughput λ^* is also the *maximum throughput* for which the average response time is finite.

In summary, to investigate a given configuration of the tape library model, the model is first simulated with a saturated fetch queue in order to estimate λ^*. Then the model is simulated with Poisson arrivals for any input rate $\lambda < \lambda^*$ and the average response time R is estimated using techniques based on the regenerative structure of the model.

• *Comparing system variants (multiple comparison procedures)*

A prime reason for simulating a system model, such as for the tape library, is to investigate the effect of different system designs on system performance. For the tape library model, it might be desired to compare the average response times for two or more system variants. Statistical methods for making such comparisons are called multiple comparison procedures [14]. One such procedure is to obtain point and confidence interval estimates for all pairwise differences between the average response times for the system variants.

Suppose there are L system variants to be compared. Let the superscript l refer to the lth system variant. Assume that each system variant has traffic intensity less than one. Each system variant is simulated independently for the same number n of tours. It can be shown that for each l, $l = 1, \cdots, L$, $n^{\frac{1}{2}}\nu^{(l)}(n)(Q^{(l)}(n) - Q^{(l)})$ is asymptotically distributed as a normal random variable having mean zero and variance $V^{(l)}$. Denote this asymptotic normality by

$$n^{\frac{1}{2}}\nu^{(l)}(n)\,[Q^{(l)}(n) - Q^{(l)}] \sim N(0, V^{(l)}).$$

By applying theorem 4.4.8 of Chung [15], it is permissible to replace $\nu^{(l)}(n)$ by $\mathrm{E}[\nu^{(l)}]$ with the result that

$$n^{\frac{1}{2}}\left[Q^{(l)}(n) - Q^{(l)}\right] \sim N\left[0, V^{(l)}/(\mathrm{E}[\nu^{(l)}])^2\right].$$

Since $Q^{(l)}(n)$ and $Q^{(j)}(n)$ are independent for $j \neq l$, it follows that

$$n^{\frac{1}{2}}\left[Q^{(l)}(n) - Q^{(j)}(n) - (Q^{(l)} - Q^{(j)})\right] \sim N\left[0, V^{(l)}/\mathrm{E}[\nu^{(l)}])^2 + V^{(j)}/\mathrm{E}[\nu^{(j)}])^2\right].$$

Now replacing $V^{(l)}$, $\mathrm{E}[\nu^{(l)}]$, $V^{(j)}$ and $\mathrm{E}[\nu^{(j)}]$ by $V^{(l)}(n)$, $\nu^{(l)}(n)$, $V^{(j)}(n)$ and $\nu^{(j)}(n)$, respectively, and again applying theorem 4.4.8 of Chung [15] yields

$$n^{\frac{1}{2}}\left[Q^{(l)}(n) - Q^{(j)}(n) - (Q^{(l)} - Q^{(j)})\right] \div \left\{V^{(l)}(n)/[\nu^{(l)}(n)]^2 + V^{(j)}(n)/[\nu^{(j)}(n)]^2\right\}^{\frac{1}{2}} \sim N(0, 1).$$

Hence, for n sufficiently large,

$$I^{(l,j)}(n, \alpha) = \left[Q^{(l)}(n) - Q^{(j)}(n) - \delta^{(l,j)}(n, \alpha),\ Q^{(l)}(n) - Q^{(j)}(n) + \delta^{(l,j)}(n, \alpha)\right] \tag{8}$$

is approximately a $100 \cdot \alpha\%$ confidence interval for $Q^{(l)} - Q^{(j)}$ where

$$\delta^{(l,j)}(n, \alpha) = \phi^{-1}[(1+\alpha)/2] \times \left\{\{V^{(l)}(n)/[\nu^{(l)}(n)]^2 + V^{(j)}(n)/[\nu^{(j)}(n)]^2\}/n\right\}^{\frac{1}{2}}.$$

Clearly

$$Q^{(l)}(n) - Q^{(j)}(n) \tag{9}$$

is a point estimate for $Q^{(l)} - Q^{(j)}$. If the mean fetch service times and mean R/W service times are the same for system variants l and j, then (8) and (9) are, respectively, confidence interval and point estimates for $R^{(l)} - R^{(j)}$. Otherwise, (8) and (9) are adjusted accordingly.

In order to make a simultaneous confidence statement about all pairwise differences $Q^{(l)} - Q^{(j)}$, $l = 1, \cdots, L-1$, $j = l+1, \cdots, L$, a conservative procedure is to use the Bonferroni inequality [14]. Thus,

$$\begin{aligned}\Pr\{Q^{(l)} - Q^{(j)} \in I^{(l,j)}(n, \alpha),\ l = 1, \cdots, L-1,\ j = l+1, \cdots, L\} &\geq 1 - \sum_{l=1}^{L-1}\sum_{j=l+1}^{L}(1 - \Pr\{Q^{(l)} - Q^{(j)} \in I^{(l,j)}(n, \alpha)\}) \\ &\approx 1 - (1-\alpha)L(L-1)/2. \end{aligned} \tag{10}$$

If L and $1-\alpha$ are small, say $L = 3$ and $1 - \alpha = 0.05$, this bound can provide a useful approximation to the exact joint probability [14].

It is also interesting to compare the saturated throughput for L system variants. This is accomplished by performing U independent replications for each system variant and computing for each l and $j \neq l$ a point estimate for $\lambda^{*(l)} - \lambda^{*(j)}$ given by

$$\lambda^{*(l)}(K, U) - \lambda^{*(j)}(K, U) \tag{11}$$

and an approximate $100 \cdot \alpha\%$ confidence interval for $\lambda^{*(l)} - \lambda^{*(j)}$ given by

$$I^{*(l,j)}(K, U, \alpha) = \left[\lambda^{*(l)}(K, U) - \lambda^{*(j)}(K, U) - \delta^{*(l,j)}(K, U, \alpha),\ \lambda^{*(l)}(K, U) - \lambda^{*(j)}(K, U) + \delta^{*(l,j)}(K, U, \alpha)\right], \tag{12}$$

Table 2 Configurations of the tape library model.

Basic configuration	*Configuration label M/N/C*	*M*	*N*	C_1	C_2	C_3
Balanced	1/2/1	1	2	1	1	1
	1/2/2	1	2	1	2	2
	1/2/3	1	2	2	3	3
Picker limited	1/4/1	1	4	1	1	1
	1/4/2	1	4	1	2	2
	1/4/3	1	4	2	3	3
R/W limited	2/2/1	2	2	1	1	1
	2/2/2	2	2	1	2	2
	2/2/3	2	2	2	3	3

Table 3 Estimates of saturated throughput λ^*.

Configuration	*Point estimate* $\lambda^*(1000, 10)$	*95% confidence interval* $I^*(1000, 10, 0.95)$
1/2/1	0.15871	(0.15800, 0.15942)
1/2/2	0.23417	(0.23299, 0.23535)
1/2/3	0.24448	(0.24372, 0.24523)
1/4/1	0.24815	(0.24724, 0.24906)
1/4/2	0.25115	(0.24999, 0.25231)
1/4/3	0.25139	(0.24994, 0.25284)
2/2/1	0.19506	(0.19432, 0.19580)
2/2/2	0.25069	(0.24890, 0.25248)
2/2/3	0.24957	(0.24813, 0.25101)

where

$$\delta^{*(l,j)}(K, U, \alpha) = \theta^{-1}_{2U-2}[(1+\alpha)/2] \times \{[V^{*(l)}(K, U) + V^{*(j)}(K, U)]/U\}^{\frac{1}{2}}.$$

As discussed in Scheffé [13], Chapter 10, even for small U the validity of the above confidence interval is insensitive to non-normality of the observations and to inequality of the variances of the two populations, i.e., to $\mathrm{Var}[\lambda_u^{*(l)}] \neq \mathrm{Var}[\lambda_u^{*(j)}]$, $j \neq l$. In order to make a simultaneous confidence statement about $\lambda^{*(l)} - \lambda^{*(j)}$, $l = 1, \cdots, L-1, j = l+1, \cdots, L$, the Bonferroni inequality can again be used.

Experimental results

In this section the maximum throughput and average response time for several configurations of the tape library model are studied via simulation experiments. The section begins with a description of the configurations considered. Next the simulation experiments and results are presented and, finally, the validity of the techniques used is discussed.

Recall that the goal of the experiments is to understand the effect of the carousel capacity C on performance. For C sufficiently large the blocking effect should be negligible and the performance of the tape library should not depend on C. In particular, consider the saturated system. It can be shown that for each R/W server and each picker server the utilizations $S_{R/W}$ and S_P, respectively, i.e., the fractions of time the servers are busy, are given by

$$S_{R/W} = \lambda^* \mathrm{E}[T_{R/W}]/N,$$

$$S_P = \lambda^* (\mathrm{E}[T_F] + \mathrm{E}[T_P])/M,$$

where T_P is the putaway service time.

Since utilizations are never greater than one it follows that

$$\lambda^* \leq \min \{N/\mathrm{E}[T_{R/W}], M/(\mathrm{E}[T_F] + \mathrm{E}[T_P])\}. \quad (13)$$

Inequality (13) holds for any value of the carousel capacity C. If $N/\mathrm{E}[T_{R/W}] < M/(\mathrm{E}[T_F] + \mathrm{E}[T_P])$ the system is said to be *R/W limited* since the R/W servers provide the bound on λ^*. The system is *picker limited* if the inequality is reversed and *balanced* if $N/\mathrm{E}[T_{R/W}] = M/(\mathrm{E}[T_F] + \mathrm{E}[T_P])$. Three basic configurations of the model are studied corresponding to picker limited ($M = 1, N = 4$), R/W limited ($M = 2, N = 2$) and balanced ($M = 1, N = 2$). For each basic configuration, carousel capacities $C = 1, 2$, and 3 are considered. Using labels of the form $M/N/C$, Table 2 summarizes the nine configurations. Observe that for all configurations the bound on λ^* in (13) is 0.25 customer/second. (Recall that T_F and T_P are uniformly distributed on [1, 3] and $T_{R/W}$ is uniformly distributed on [4, 12].)

The saturated system was simulated using Program 1. For each configuration ten independent replications were simulated. Each replication was started with the fetch queue nonempty (logically containing an infinite number of customers) and no customer elsewhere. Each replication was terminated when 1000 customers had departed from the network. Point and 95% confidence interval estimates for λ^* were calculated from (6) and (7), respectively, where $K = 1000$, $U = 10$, and $\alpha = 0.95$. The results are shown in Table 3.

Evidently, the effect on λ^* of blocking is most prominent for the balanced system. In all cases, however, the blocking effect appears minimal for $C \geq 2$ (the point estimates for λ^* are all within 7% of 0.25 if $C \geq 2$). For this reason, capacities larger than 3 were not considered.

Comparisons of the saturated throughputs at different capacities were made using the multiple comparison

Table 4 Saturated throughput comparisons.

Basic configuration M	N	Capacities l	j	$\lambda^{*(j)} - \lambda^{*(l)}$ Point estimate	95% confidence interval
1	2	1	2	0.07546	(0.07418, 0.07674)
		1	3	0.08577	(0.08481, 0.08673)
		2	3	0.01031	(0.00901, 0.01161)
1	4	1	2	0.00300	(0.00163, 0.00437)
		1	3	0.00324	(0.00165, 0.00483)
		2	3	0.00024	(−0.00148, 0.00196)
2	2	1	2	0.05563	(0.05383, 0.05743)
		1	3	0.05451	(0.05300, 0.05601)
		2	3	−0.00112	(−0.00326, 0.00101)

procedures. For given M and N let $\lambda^{*(l)} = \lambda^*$ for basic configuration $M/N/l$. Table 4 shows point and 95% confidence interval estimates for $\lambda^{*(j)} - \lambda^{*(l)}$ calculated from (11) and (12) for $(l, j) = (1, 2), (1, 3)$, and $(2, 3)$ for each basic configuration. Using inequality (10) allows one to make the statement for each basic configuration that all three differences are contained in their respective intervals with probability at least $1 - (0.05) \times 3 = 0.85$, where 0.85 is an approximation to the true bound.

The multiple comparison techniques allow stronger statements to be made about the effect of varying the capacity C than can be made solely from the estimates in Table 3. As an example, consider the statements that can be made about the effect of increasing C from 2 to 3 for the balanced system ($M = 1$, $N = 2$). From Table 3, if the saturated throughputs for configurations 1/2/2 and 1/2/3 are both contained in their respective intervals then $\lambda^{*(3)} - \lambda^{*(2)}$ is contained in the interval (0.00837, 0.01224), which has width 0.00387. This event occurs with probability $(0.95)^2 = 0.9025$. From the multiple comparison technique, $\lambda^{*(3)} - \lambda^{*(2)}$ has an estimated 95% confidence interval (0.00901, 0.01161) in Table 4, which has width 0.00260. In either case the mean of the difference is estimated to be 0.01031, but the latter statement is stronger (95% vs 90%) and is for a narrower interval (0.00260 vs 0.00387).

At first glance the point estimates of the saturated throughput in Table 3 suggest some possible difficulties. Namely, the estimated values of λ^* for configurations 1/4/2, 1/4/3 and 2/2/2 exceed the known bound of 0.25. Inspection of the 95% confidence intervals, however, reveals that rates below 0.25 are never excluded. Similarly, λ^* appears to decrease as C increases from 2 to 3 for $M = N = 2$. In Table 4 the estimated 95% confidence interval for this difference is (−0.00326, 0.00101), suggesting that no strong conclusion about this effect should be made.

Point and confidence interval estimates for the average response time R were obtained for three values of the input rate λ. Based on the estimates of λ^* the three values chosen were $\lambda = 0.05$, 0.10, and 0.15 customer/second (the lowest λ^* estimate was 0.15871 with a 95% confidence interval of (0.15800, 0.15942) for the 1/2/1 configuration). It is assumed subsequently that $\lambda^* > 0.15$ for all configurations and thus that the traffic intensities are less than one.

The simulation experiments were made using Program 2. For $\lambda = 0.05$ and 0.10, 500 tours were simulated for each configuration and for $\lambda = 0.15$, 1000 tours were simulated for each configuration. All simulations were independent. The number of tours for each value of λ was chosen by performing pilot simulation runs and observing the number of tours necessary to yield estimated 95% confidence intervals that had widths not greater than 10% of the point estimates. (A later investigation of the coverages of these intervals, described at the end of this section, raises the possibility that the number of tours chosen, especially for $\lambda = 0.15$, may be too small to strongly conclude that the estimated confidence intervals are valid.)

Point and 95% confidence interval estimates for the average response time R were obtained using (4) and (5) for each configuration and each value of λ. The results are shown in Table 5. The effect of blocking is seen to be most dramatic at $\lambda = 0.15$, especially for the balanced configuration: Increasing C from 1 to 2 causes the estimated average response time to decrease from 66.487 to 13.674 seconds. As with λ^*, increasing C from 2 to 3 has relatively little effect on the average response time.

Comparisons of the average response time for different configurations were made using the multiple comparison procedures. For given M, N, and λ let $R^{(l)}$ = average response time for configuration $M/N/l$ with input rate λ. Point estimates and 95% confidence interval estimates for $R^{(j)} - R^{(l)}$ over the nine configurations and over the three input rates were calculated from (9) and (8) and are shown in Table 6. The hypothesis that increasing the capacities always decreases the average response time corresponds to all $R^{(j)} - R^{(l)}$, $(l, j) = (1, 2), (1, 3), (2, 3)$ being negative. It is seen that for five cases the point estimate of $R^{(j)} - R^{(l)}$ is positive, but in all cases the estimated 95% confidence intervals include negative values and the hypothesis cannot be rejected.

The total simulation time using Program 1 to estimate λ^* was 360 seconds and the total simulation time using Program 2 to estimate R was 526 seconds. In the latter case 33, 67 and 426 seconds were expended for the cases $\lambda = 0.05$, 0.10, and 0.15, respectively.

The confidence interval $I_R(n, \alpha)$ for the average response time R given by (5) is an approximate $100 \cdot \alpha\%$ confidence interval; i.e., the probability that R is con-

Table 5 Point and 95% confidence interval estimates for average response time R.

Configuration	$\lambda = 0.05$ *(500 tours)*	$\lambda = 0.10$ *(500 tours)*	$\lambda = 0.15$ *(1000 tours)*
1/2/1	10.907(10.679, 11.135)	14.572(13.785, 15.360)	66.487(52.921, 80.053)
1/2/2	10.481(10.370, 10.592)	11.615(11.364, 11.865)	13.674(13.314, 14.035)
1/2/3	10.579(10.438, 10.721)	11.645(11.377, 11.912)	13.739(13.440, 14.037)
1/4/1	10.255(10.204, 10.305)	10.779(10.605, 10.954)	12.209(11.837, 12.581)
1/4/2	10.257(10.198, 10.316)	10.545(10.478, 10.611)	11.174(11.025, 11.322)
1/4/3	10.222(10.176, 10.269)	10.630(10.539, 10.722)	11.096(10.991, 11.201)
2/2/1	10.467(10.293, 10.640)	12.181(11.658, 12.704)	18.159(17.197, 19.121)
2/2/2	10.387(10.262, 10.513)	11.179(10.985, 11.372)	12.996(12.636, 13.357)
2/2/3	10.324(10.219, 10.429)	11.037(10.877, 11.197)	12.723(12.475, 12.971)

tained in $I_R(n, \alpha)$ is approximately equal to α, and the approximation becomes better as n is increased. The probability that R is contained in $I_R(n, \alpha)$ is called the true coverage. In addition, the point estimate $R(n)$ given by (4) is, in general, biased for finite n, i.e., $\mathrm{E}[R(n)] \neq R$, but for n sufficiently large the bias is small. A question arises as to how large n must be for the confidence interval approximation to be satisfactory and for the bias to be small. If R were know, the bias and true coverage could be estimated via simulation. If, in the model, there is one picker, one R/W unit and if $C_1 = C_2 = C_3 = 1$, then at most one customer can be in the R/W unit or at the picker at any one time. The model thus degenerates to an M/G/1 queue with service time equal to the sum of the fetch, R/W and putaway service times. The average response time for this degenerate model is known. In [8] the bias of $R(n)$ and true coverage of $I_R(n, \alpha)$ for $\alpha = 0.95$ are estimated by performing 100 independent replications of a simulation of the M/G/1 queue; each replication is terminated after n tours have been completed. Traffic intensities of 0.2, 0.5 and 0.8 and several values of n are considered. It is found that the bias of $R(n)$ is not significant even for values of n as small as 50. The estimated coverage, however, is quite low for small n. For $\rho = 0.2$ and $\rho = 0.5$ the estimated coverage is approximately 0.80 for $n = 50$ but increases to above 0.90 for $n = 500$. For $\rho = 0.8$ the estimated coverage is only 0.68 at $n = 50$ and 0.87 at $n = 1000$. (In [8] the average waiting time $W = R + \mathrm{E}[T_P]$ is considered but the bias and coverage results are identical for W and R.)

It is of interest to consider the validity of the confidence intervals for the library model for other values of M, N and C, but since the average response time is not known, a method difference from that in [8] must be used. The balanced system, i.e., the one with $M = 1$, $N = 2$, and $C = 1$ was selected for study at an input rate $\lambda = 0.10$. A single long simulation of 30 000 tours was performed and point and 95% confidence interval esti-

Table 6 Average response time comparisons.

Basic configuration M	N	*Capacities* l	j	$\lambda = 0.05$ $R^{(j)} - R^{(l)}$
1	2	1	2	−0.42673(−0.68033, −0.17313)
		1	3	−0.32817(−0.59663, −0.05972)
		2	3	0.09856(−0.08135, 0.27847)
	4	1	2	0.00207(−0.07545, 0.07959)
		1	3	−0.03258(−0.10121, 0.03605)
		2	3	−0.03465(−0.10957, 0.04028)
2	2	1	2	−0.07936(−0.29349, 0.13477)
		1	3	−0.14295(−0.34570, 0.05981)
		2	3	−0.06359(−0.22731, 0.10014)
				$\lambda = 0.10$
M	N	l	j	$R^{(j)} - R^{(l)}$
1	2	1	2	−2.9578 (−3.7843, −2.1313)
		1	3	−2.9279 (−3.7597, −2.0960)
		2	3	0.02991(−0.33651, 0.39633)
1	4	1	2	−0.23484(−0.42164, −0.04804)
		1	3	−0.14896(−0.34607, 0.04815)
		2	3	0.08588(−0.02721, 0.19897)
2	2	1	2	−1.0023 (−1.5601, −0.4444)
		1	3	−1.1442 (−1.6913, −0.5970)
		2	3	−0.14189(−0.39277, 0.10900)
				$\lambda = 0.15$
M	N	l	j	$R^{(j)} - R^{(l)}$
1	2	1	2	−52.813 (−66.384, −39.242)
		1	3	−52.749 (−66.318, −39.179)
		2	3	0.06447(−0.40363, 0.53258)
1	4	1	2	−1.0355 (−1.4355, −0.6355)
		1	3	−1.1133 (−1.4994, −0.7272)
		2	3	−0.07779(−0.25926, 0.10368)
2	2	1	2	−5.1627 (−6.1896, −4.1358)
		1	3	−5.4360 (−6.4292, −4.4428)
		2	3	−0.27334(−0.71063, 0.16395)

Table 7 Estimates for 1/2/1 configuration and $\lambda = 0.10$.

Average response time		
Point estimate	*95% confidence interval*	*Experiment*
14.755	(14.606, 14.905)	30 000 tours
14.599	(13.888, 15.310)	Averages over 100 runs of 1000 tours each

mates of 14 755 and (14 606, 14 905), respectively, were obtained for R. Next, 100 independent simulation runs of 1000 tours each were performed. The resulting estimates, averaged over the 100 runs, are shown in Table 7 together with the results for 30 000 tours. It was assumed that the confidence interval based on 30 000 tours was valid, i.e., that this interval, call it $I(30\,000)$, contains R with probability 0.95. Thus, if $I(30\,000)$ is contained totally within an estimated confidence interval based on 1000 tours, then R is contained in the latter interval with probability at least 0.95. $I(30\,000)$ was observed to fall totally within 79, partially within 16, and totally outside of 5 of the 100 estimated confidence intervals based on 1000 tours. This suggests that the true coverage for 1000 tours is between 0.79 and 0.95.

From this experiment one cannot strongly claim or disclaim the validity of the 95% confidence intervals estimated from simulation runs of 1000 tours. However, it is reasonable to conclude that the true coverages for the confidence intervals in Tables 5 and 6, calculated from runs of 500 or 1000 tours, are not so small that the conclusions based on these tables are meaningless. Further investigation is suggested here, but was not carried out due to the computer time required. (It took substantially more computer time to perform the experiment in Table 7 than it took to perform all the experiments in Tables 5 and 6.)

Conclusions

The simulation study presented in this paper illustrates several issues that arise concerning the application of regenerative simulation techniques to complex queuing models. The model considered consists of an open network of interconnected queues and incorporates scheduling algorithms, finite capacity queues and non-exponential service times. Since all service times and interarrival times for this model are i.i.d. it is clear that the system stochastically restarts whenever a customer arrives at the empty system. However, in order that the regenerative techniques be applicable, stronger conditions must be satisfied. It is necessary that the system stochastically restart with probability one and that certain random variables associated with a tour have finite first two moments. Subject to certain restrictions on the interarrival and service times, e.g., the interarrival time distribution has a density on the whole positive real line and all service times have finite fourth moments, it was conjectured that the above conditions hold if the traffic intensity (input rate/saturated throughput) is less than one. Proving the validity of this conjecture is an open problem. Nonetheless, the regenerative techniques were applied when simulating the model for traffic intensities less than one. The system was observed to empty out often during the simulation runs, although this does not guarantee the validity of the conjecture.

When the regenerative techniques are applied, the simulation duration should be large enough so that the estimated confidence intervals are approximately valid and so that the widths of the intervals are small enough to provide useful information about the response variables being estimated. (These issues arise when any technique is applied to estimate confidence intervals.) The validity issue is the more difficult to address. The test of validity proposed at the end of the previous section may be too costly to apply in practice. Testing the point estimates for normality, using say the Kolmogorov-Smirnov test of fit, is also costly. Results for the M/G/1 queue [8] and results from the test in the previous section provide some evidence that simulation durations of 500 or 1000 tours are sufficient to provide reasonably valid results for the tape library model, with the longer duration required at higher traffic intensities. Since, the confidence interval width is (approximately) inversely proportional to the square root of the number of tours, a pilot run should provide a rough indication of how large the simulation duration should be for a desired confidence interval width.

In summary, regenerative simulation was found to be a viable tool for numerically studying a complex queuing model which is not analytically tractable. Moderate simulation durations (durations of 500 and 1000 tours were used where the average computer time to simulate a tour was 0.03 second using a large computer) were sufficient to obtain fairly accurate confidence interval estimates. The model was first simulated under saturated conditions with independent replications used to estimate a confidence interval for λ^*, the maximum input rate for which regenerative simulation is applicable.

Acknowledgment

The technical assistance of Dion L. Johnson in creating and maintaining a convenient interface with SIMPL/I is appreciated.

References

1. M. A. Crane and D. L. Iglehart, "Simulating Stable Stochastic Systems: III. Regenerative Processes and Discrete-Event Simulations," *Oper. Res.* **23,** 33 (1975).
2. D. L. Iglehart, "Simulating Stable Stochastic Systems, V: Comparison of Ratio Estimators," *Naval Res. Logist. Quart.*, to be published.
3. M. A. Crane and D. L. Iglehart, "Simulating Stable Stochastic Systems, I: General Multiserver Queues," *J. ACM* **21,** 103 (1974).
4. G. S. Fishman, "Estimation in Multiserver Queuing Simulations," *Oper. Res.* **22,** 72 (1974).
5. A. M. Law, "Efficient Estimators for Simulated Queuing Systems," *ORC 74-7*, Operations Research Center, University of California, Berkeley, California 1974.
6. S. S. Lavenberg and G. S. Shedler, "Derivation of Confidence Intervals for Work-Rate Estimators in a Closed Queuing Network," *SIAM J. Comput.* **4,** 108 (1975).
7. S. S. Lavenberg, "Efficient Estimation via Simulation of Work-Rates in Closed Queueing Networks," *Proceedings in Computational Statistics*, Physica Verlag, Vienna, Austria, 1974, pp. 353-362.
8. S. S. Lavenberg and D. R. Slutz, "Introduction to Regenerative Simulation," *IBM J. Res. Develop.* **19,** 458 (1975, this issue).
9. S. Damron, J. Lucas, J. Miller, E. Salbu and M. Wildmann, "A Random Access Terabit Magnetic Memory," *Proc. Fall Joint Computer Conference* (AFIPS) **33,** Part 2, 1381 (1968).
10. *Introduction to IBM 3850 Mass Storage System (MSS)*, Form No. GA32-0028, IBM Corporation, White Plains, New York, 1975.
11. *SIMPL/I (Simulation Language Based on PL/I)*, Form No. GH19-5035, IBM Corporation, White Plains, New York, 1972.
12. W. Whitt, "Embedded Renewal Processes in the GI/G/s Queue," *J. Appl. Prob.* **9,** 650 (1972).
13. H. Scheffé, *The Analysis of Variance*, John Wiley & Sons, Inc., New York, 1959.
14. J. P. C. Kleijnen, *Statistical Techniques in Simulation*, Part II, Marcel Dekker, Inc., New York, 1975.
15. K. L. Chung, *A Course in Probability Theory*, Harcourt, Brace & World, Inc., New York, 1968.

Received February 4, 1975; revised April 15, 1975

The authors are located at the IBM Research Division Laboratory, Monterey and Cottle Roads, San Jose, California 95193.