MANAGEMENT SCIENCE
Vol. 38, No. 3, March 1992
Printed in U.S.A.

EXPERIMENTS WITH INITIAL TRANSIENT DELETION FOR PARALLEL, REPLICATED STEADY-STATE SIMULATIONS*

PETER W. GLYNN AND PHILIP HEIDELBERGER
Department of Operations Research, Stanford University, Stanford, California 94305
IBM Thomas J. Watson Research Center, Hawthorne, P.O. Box 704,
Yorktown Heights, New York 10598

A simple and effective way to exploit parallel processors in discrete event simulations is to run multiple independent replications, in parallel, on multiple processors and to average the results at the end of the runs. We call this the method of parallel replications. This paper is concerned with using the method of parallel replications for estimating steady-state performance measures. We report on the results of queueing network simulation experiments that compare the statistical properties of several possible estimators that can be formed using this method. The theoretical asymptotic properties of these estimators were determined in Glynn and Heidelberger (1989a, b). Both the theory and the experimental results reported here strongly indicate that a nonstandard (in the context of steady-state simulation), yet easy to apply, estimation procedure is required on highly parallel machines. This nonstandard estimator is a ratio estimator. The experiments also show that use of the ratio estimator is advantageous even on machines with only a moderate degree of parallelism.
(SIMULATION; REPLICATIONS; PARALLEL PROCESSING; STEADY-STATE; ESTIMATION)

1. Introduction

A simple and effective way to exploit parallel processors for computationally intensive discrete event simulations is to run multiple independent replications, in parallel, on multiple processors and to average the results at the end of the runs. We call this the method of parallel replications. This paper is concerned with using the method of parallel replications for estimating steady-state performance measures. In particular, we report on the results of queueing network simulation experiments that compare the statistical properties of several possible estimators that can be formed using this method. The theoretical asymptotic properties of these estimators were determined in Glynn and Heidelberger (1989a, b). Both the theory and the experimental results reported here strongly indicate that a nonstandard (in the context of steady-state simulation), yet easy to apply, estimation procedure is required on highly parallel machines. This nonstandard estimator takes the form of a ratio estimator. The experiments also show that use of the ratio estimator is advantageous even on machines with only a moderate degree of parallelism.

We remark that an alternative approach to parallel processing of simulations is distributed simulation, in which multiple processors cooperate together to generate a single realization of the stochastic process being simulated. For an excellent introduction to distributed simulation and a thorough bibliography on this topic, see Fujimoto (1990). A theoretical comparison of the statistical efficiencies of parallel replications and distributed simulation for estimating steady-state parameters can be found in Heidelberger (1986).

Intuitively, when using the method of parallel replications on a large number of processors, one expects to get highly accurate estimates after only a relatively short amount of time. However, there are some potentially serious statistical problems inherent in this approach, and careful estimation procedures must be applied in order to obtain estimates with the proper (or desired) statistical properties. These problems basically arise because

* Accepted by James R. Wilson; received May 17, 1990. This paper has been with the authors 1 month for 1 revision.

0025-1909/92/3803/0400$01.25

any bias effects are magnified on highly parallel machines, i.e., because of the bias, one obtains highly accurate estimates of the wrong quantity.

In the context of estimating transient performance measures (or steady-state performance measures in regenerative simulations), these problems have been identified and addressed in Heidelberger (1988) and Glynn and Heidelberger (1991). These papers show that nonstandard estimators are required on highly parallel machines. Other issues related to parallel replications for estimating transient quantities are described in Bhavsar and Isaac (1987).

For estimating steady-state performance measures, the traditional approaches (on a single processor) are to use either the method of batch means, or independent replications with initial transient deletion (see, e.g., Law 1977, Law and Carson 1979, Law and Kelton 1982, or Bratley, Fox and Schrage 1987). When using replications, it is generally advised to use only "a few long" replications (say 10 to 20) with deletion to reduce susceptibility to the effects of initialization bias.

With the prospect of parallelism as motivation, Glynn and Heidelberger (1989a, b) have addressed, from a theoretical point of view, how one should control the number of replications (processors), the length of each replication, and the length of the initial transient deletion interval in order to obtain valid central limit theorems for steady-state parameters. Such central limit theorems can then be used as the basis for confidence interval formation. These papers, which extend the single-processor results of Glynn (1987 and 1990), show that valid confidence intervals can be obtained even for a very large number of processors P (relative to the replication length) provided the deletion interval grows appropriately *and* the proper (nonstandard) ratio estimator is used.

On the other hand, if each processor is run for a prespecified amount of computer time, c, then it was shown that initial transient deletion does not, in fact, remove the dominant term in the bias expansion (i.e., the term of order $1/c$) of the traditional (standard) independent replications estimator, $\alpha_T(P, c)$. In this case, the amount of simulated time generated by each processor is a rv (random variable) and thus the traditional estimator becomes a ratio estimator. The bias expansion of this estimator reveals two sources of bias of order $1/c$:

1. "Initialization" bias, i.e., bias essentially due to the simulation not being started in steady-state conditions. (This also includes a contribution because the amount of simulation time is random.)
2. "Ratio" bias, i.e., bias due to the fact that the denominator of the ratio estimator is a rv.

When done appropriately, initial transient deletion effectively removes the initialization bias. However, initial transient deletion does not remove the ratio bias. The nonstandard estimator, $\alpha_R(P, c)$, corresponds to the classical ratio estimator which is typically used in sample surveys (see, e.g., Cochran 1963) and regenerative simulations (see, e.g., Crane and Iglehart 1975 or Iglehart 1975). The initialization bias (of order $1/c$) in $\alpha_R(P, c)$ is the same as the initialization bias in $\alpha_T(P, c)$, but the ratio bias in $\alpha_R(P, c)$ is P times smaller than the ratio bias in $\alpha_T(P, c)$. Thus initial transient deletion effectively removes all bias of order $1/c$ from $\alpha_R(P, c)$.

The net effect of this analysis is that, when using the ratio estimator $\alpha_R(P, c)$, valid confidence intervals for steady-state parameters can be formed when very highly parallel machines (large P) are run for a relatively short amount of time (qualitatively, small c/P). In this situation, valid confidence intervals are not obtained when estimating the steady-state parameter by the traditional estimator $\alpha_T(P, c)$. Using $\alpha_T(P, c)$, valid confidence intervals are only obtained when (qualitatively) c/P is very large, i.e., when the length of each replication is large with respect to the number of processors. We emphasize that $\alpha_T(P, c)$ and $\alpha_R(P, c)$ both make use of exactly the same underlying data; they merely average these data differently.

The purpose of this paper is to demonstrate, experimentally, that this dramatic difference between the theoretical asymptotic behaviors of these two estimators is exhibited in sample sizes that are not unreasonable in practice. In simulations of simple queueing systems, we show that noticeable effects (increased bias and decreased confidence interval coverage) are present on as few as 32 to 64 processors. Severe effects are observed on 128 or more processors. Thus, from both a theoretical and practical viewpoint, the traditional estimator, $\alpha_T(P, c)$ should be avoided on even moderately sized parallel processors. There are other causes of confidence interval coverage degradation besides point estimator bias, e.g., underestimation of the variance, or sample sizes that are not large enough to produce point estimators having an approximate normal distribution. However, the focus of this paper is the effect of ratio bias on both point estimators and confidence interval coverage.

The rest of the paper is organized as follows. In §2, we summarize the relevant theoretical results from Glynn and Heidelberger (1989a, b) and Glynn (1987 and 1990). In §3, we describe the queueing models that we used for experimentation. In §4, we describe the design of the experiments, including point and interval estimation procedures. The results of the simulation experiments are presented in §5. Finally, §6 contains a summary of our findings, a discussion of their relevance to traditional steady-state estimation on single-processor systems, and an indication of related future research topics.

2. Summary of Theoretical Results

The results that we quote from Glynn and Heidelberger (1989a, b) and Glynn (1987 and 1990) were derived under reasonable, yet fairly technical assumptions. These basically involve assumptions concerning the existence of central limit theorems and their associated uniform integrability (i.e., moment convergence in the central limit theorem), an exponential convergence rate to the steady-state distribution, and certain boundedness conditions. Since a precise statement of these conditions would be rather tedious (and not particularly illuminating for the present purposes), we will make the simplifying assumption that the process being simulated is an irreducible, finite state space, continuous time Markov Chain (CTMC) with state space denoted by $\mathbf{E}$. Such processes automatically satisfy all of the necessary assumptions.

We let $\{X(s), s \geq 0\}$ denote the CTMC. The parameter s denotes simulated time so that $X(s)$ is the state of the process at simulated time s. There then exists a rv X such that $X(s) \Rightarrow X$ as $s \to \infty$ where $\Rightarrow$ denotes convergence in distribution. The rv X has the steady-state distribution of the CTMC; this steady-state distribution is independent of the distribution of $X(0)$ (the initial conditions of the CTMC). We shall be interested in estimating quantities of the form $\alpha = \mathrm{E}[f(X)]$ for some function f. (Since we have assumed that the state state space is finite, $\mathrm{E}[|f(X)|] < \infty$ so that α exists.)

There are P processors. Simultaneously an independent simulation of the CTMC is started on each processor. We let $X_i(s)$ denote the state of the process at simulated time s on processor i, $i = 1, \ldots, P$. Let $T_i(c)$ denote the simulation time on processor i after c units of computer time. (The discussion in this paper also holds if c is measured in units of "wall clock" time, or for that matter, any other way of measuring time.) Let $C_i(s)$ denote the (random) amount of computer time required on processor i to obtain s units of simulated time.

There are a variety of ways to set the run length. We will consider two reasonable and practical approaches. In the first approach, a fixed amount of simulated time, say t_P, is generated on each processor. In this case, the completion time of the simulation experiment is $C(t_P) = \max\{C_1(t_P), \ldots, C_P(t_P)\}$, which is a rv. Since we can view $\{C_i(s), s \geq 0\}$ as a cumulative process, it is reasonable to assume that each $C_i(t_P)$ obeys a central limit theorem, i.e., there exist finite positive constants λ and σ_1 such that

$$\frac{C_i(t_P) - t_P/\lambda}{\sqrt{t_P}} \Rightarrow \sigma_1 N(0, 1) \tag{2.1}$$

as $t_P \to \infty$ where $N(0, 1)$ denotes a normally distributed rv with mean zero and variance one. The parameter λ^{-1} is the long-run rate at which computer time is expended per unit of simulated time. Alternatively, λ is the long-run rate at which simulated time is generated per unit of computer time. Since the completion time is the maximum of iid (independent and identically distributed) rvs that are approximately normally distributed, the expected completion time is approximately equal to $(t_P/\lambda) + \sigma_1\sqrt{2t_P \ln(P)}$ provided t_P and $P \to \infty$ in such a way that $t_P/P^2 \to \infty$ (see Glynn and Heidelberger 1989b). In this expression, (t_P/λ) is the expected completion time of an individual processor and $\sigma_1\sqrt{2t_P \ln(P)}$ is the additional time until the last processor finishes. The factor $\sqrt{2 \ln(P)}$ arises as the maximum of P iid $N(0, 1)$ rvs, which then gets multiplied by the standard deviation of an individual completion time, $\sigma_1\sqrt{t_P}$ (see, e.g., p. 14 of Leadbetter, Lindgren and Rootzén 1983). Notice that if $\sigma_1 = 0$, then $C_i(t_P) \equiv t_P/\lambda$, i.e., computer time is deterministically proportional to simulated time and there is no completion time penalty. Since the holding time in a state (in simulated time units) is a rv and since the amount of work (computer time) to generate a transition may depend on the state of the system (e.g., the time to put an event on the future event list typically grows with the length of the list), we view such proportionality as the exception rather than the rule. Thus, in general, the completion time penalty grows as $\sqrt{2t_P \ln(P)}$ which is clearly undesirable. This stopping rule, in conjunction with an initial transient deletion rule that deletes a fixed amount of simulated time, was analyzed in Glynn and Heidelberger (1989a).

In the second approach, we stop each simulation at exactly the same computer time, c. In this approach, the completion time of the experiment is deterministic, but the amount of simulated time generated on each processor, $T_i(c)$, is now a rv. Note that $T_i(c) = \sup\{t \geq 0 : C_i(t) \leq c\}$, so that $\{T_i(c), c \geq 0\}$ is the inverse process of $\{C_i(t), t \geq 0\}$. For simplicity, we will assume that the cumulative process $C_i(t)$ can be represented as an integral, i.e.,

$$C_i(t) = \int_0^t \chi(X_i(s))ds \qquad \text{where} \qquad 0 < \chi(j) < \infty \text{ for all } j \in \mathbf{E}. \tag{2.2}$$

It is well known that such CTMCs satisfy a bivariate central limit theorem:

$$\sqrt{c}\left(\frac{\int_0^{T_i(c)} f(X_i(s))ds}{T_i(c)} - \alpha, \frac{T_i(c)}{c} - \lambda\right) \Rightarrow \mathbf{N}(\mathbf{0}, \mathbf{A}) \tag{2.3}$$

as $c \to \infty$ where $\mathbf{N}(\mathbf{0}, \mathbf{A})$ denotes a bivariate normally distributed random vector with means zero and covariance matrix $\mathbf{A}$. This bivariate central limit theorem can be shown by applying the Cramér-Wold device (see p. 49 of Billingsley 1968) to a univariate central limit theorem for regenerative processes (see, e.g., Crane and Iglehart 1975). The explicit form of the entries of $\mathbf{A}$ will not be needed in this paper, but its entries can be defined in terms of the first two moments (and cross moments) of integrals over regenerative cycles as described in Crane and Iglehart (1975). In fact, a slightly stronger version of this central limit theorem is valid (and required), namely a functional central limit theorem version of equation (2.3). See Billingsley (1968) for a discussion of functional central limit theorems. In practice, this is not a restriction.

We next assume that all the data generated in the first $\kappa_P(c)$ units of computer time are deleted for the purpose of reducing initialization bias. We assume that $\kappa_P(c)$ is a deterministic quantity. Since this deletion rule is not standard, it requires some justification. First, it represents one reasonable way to implement a fixed deletion rule, i.e., a

deletion rule that discards a fixed portion of the simulation output. The above deletion rule has the intuitive appeal that initial transient deletion is done on the same time scale as run length control. It also has the advantage of strong theoretical support, in terms of both the analytically known bias expansions and the asymptotic central limit theorems to be listed below. Note that many other deletion rules are possible and reasonable, e.g., deleting a fixed amount of simulated time, or data-driven deletion procedures that discard a random portion of the simulation depending on the outcome of statistical tests for stationarity. The situation in which run length control is based on CPU time and initial transient deletion is based on simulation time has yet to be analyzed. Although we expect the traditional estimator to suffer from ratio bias problems (since its denominator is still a rv), it is unclear how such a deletion procedure affects the bias expansions. In addition, it may happen that some replications have not yet reached the transient deletion time within the required CPU time limit. While many heuristic data-driven deletion rules have been proposed, few, if any, have been theoretically analyzed. Finally, understanding the properties of fixed deletion rules should help in the design and analysis of data-driven deletion rules.

Define $\gamma_i(c) \equiv T_i(\kappa_P(c))$ to be the (random) simulated time at which processor i begins collecting data for steady-state estimation and let

$$Y_i(c) \equiv \int_{\gamma_i(c)}^{T_i(c)} f(X_i(s))ds. \tag{2.4}$$

The length of the interval in which it is assumed that the process is "in steady-state" is then $\tau_i(c) \equiv T_i(c) - \gamma_i(c)$. The traditional steady-state simulation estimator of α (assuming we were simulating on a single processor and associating the index i with an independent replication) is

$$\alpha_T(P, c) \equiv \frac{1}{P} \sum_{i=1}^{P} \frac{Y_i(c)}{\tau_i(c)}. \tag{2.5}$$

We call this estimator "traditional" since estimators of this form occur as the default in many simulation packages and languages when steady-state estimation is performed using the method of independent replications with initial transient deletion. However, the ratio form of $\alpha_T(P, c)$ immediately suggests an alternative estimator, which is more suitable for ratio estimation:

$$\alpha_R(P, c) \equiv \frac{\sum_{i=1}^{P} Y_i(c)}{\sum_{i=1}^{P} \tau_i(c)} = \frac{\bar{Y}(c)}{\bar{\tau}(c)} \tag{2.6}$$

where $\bar{Y}(c) = (1/P) \sum_{i=1}^{P} Y_i(c)$ and $\bar{\tau}(c) = (1/P) \sum_{i=1}^{P} \tau_i(c)$. Other simulation contexts in which such ratio estimators have been considered are regenerative simulation (identify $\tau_i(c)$ and $Y_i(c)$ as the length of the ith regenerative cycle and an integral over the ith cycle, respectively) and the method of batch means with random sized batches (identify $\tau_i(c)$ and $Y_i(c)$ as the length of the ith batch and an integral over the ith batch, respectively—see Fox and Glynn 1987).

We begin by stating bias expansions for $\alpha_T(P, c)$ and $\alpha_R(P, c)$ when both P and c are large. Formally, we let the computer time c be a function of P, i.e., $c = c_P$, and consider behavior when both $P \to \infty$ and $c_P \to \infty$. We first consider the case when no initial transient deletion is performed, i.e., $\kappa_P(c_P) = 0$ and $\gamma_i(c_P) = 0$. The conditions stated above are sufficient to guarantee the following asymptotic bias expansions:

$$\mathrm{E}\,[\alpha_T(P, c_P)] = \alpha + \frac{b_T}{c_P} + o(1/c_P),$$

$$\mathrm{E}\,[\alpha_R(P, c_P)] = \alpha + \frac{b_R}{c_P} + o(1/c_P), \qquad \text{where} \tag{2.7}$$

$$b_T = a - \frac{A_{12}}{\lambda}, \qquad b_R = a. \tag{2.8}$$

(In equation (2.7) we write $h(c_P) = g(c_P) + o(1/c_P)$ if $\lim_{c_P \to \infty} c_P |h(c_P) - g(c_P)| = 0$.) The expansion for $\alpha_T(P, c)$ was derived in Glynn (1990) while the expansion for $\alpha_R(P, c)$ was derived in Glynn and Heidelberger (1989b). The precise form of the constant a is defined in terms of expectations of integrals over regenerative cycles. In order to display a, define a fixed state, say 0, to serve as a regenerative state, let ψ_n be the simulation time of the nth entrance to state 0 and let $\tilde{\psi}_2 = \psi_2 - \psi_1$. (We briefly drop the unnecessary replication index i to define a.) Then

$$a = \frac{\mathrm{E}\left[\int_0^{\psi_1} (f(X(s)) - \alpha) ds\right]}{\lambda} - \frac{\mathrm{E}\left[\int_0^{\tilde{\psi}_2} \int_0^s \chi(X(u + \psi_1)) du [f(X(s + \psi_1)) - \alpha] ds\right]}{\mathrm{E}[\tilde{\psi}_2]}. \tag{2.9}$$

Although not obvious from equation (2.9), as will be seen below (in equation (2.12)), a arises in the bias expansions because

$$\mathrm{E}\left[\int_0^{T_i(c)} (f(X_i(s)) - \alpha) ds\right] \neq 0. \tag{2.10}$$

There are two reasons why $a \neq 0$. First, the initial distribution of the process $X_i(s)$ may not equal the steady-state distribution. Second, the upper limit, $T_i(c)$, is, in general, random. Note that if the process is started in the steady-state distribution and if $T_i(c)$ is deterministic, then the left-hand side of equation (2.10) is equal to zero, in which case $a = 0$. However, even if the process is started in the steady-state distribution, a may still be nonzero if $T_i(c)$ is random. Because appropriate deletion of an initial portion of the simulation output removes a from the bias expansion of $\alpha_R(P, c_P)$, we will call a the "initialization" bias term.

Note that the traditional estimator contains an extra bias term, $-A_{12}/\lambda$, which can be thought of as ratio bias, i.e., bias because the denominator of the ratio is a rv.

To see why the bias expansions of $\alpha_T(P, c)$ and $\alpha_R(P, c)$ differ, we give the following brief heuristic arguments (which are made rigorous in the above-mentioned papers). Notice that both $\mathrm{E}[\alpha_T(P, c_P)]$ and $\mathrm{E}[\alpha_R(P, c_P)]$ can be written as $\mathrm{E}[A(c_P)/B(c_P)]$. Now let $\epsilon(c_P) = (B(c_P) - \mathrm{E}[B(c_P)])/\mathrm{E}[B(c_P)]$ and write

$$\frac{A(c_P)}{B(c_P)} = \frac{A(c_P)}{\mathrm{E}[B(c_P)](1 + \epsilon(c_P))} \approx \frac{A(c_P)(1 - \epsilon(c_P) + \epsilon^2(c_P) - \cdots)}{\mathrm{E}[B(c_P)]}. \tag{2.11}$$

Taking expectations of equation (2.11) yields

$$\mathrm{E}\left[\frac{A(c_P)}{B(c_P)}\right] = \frac{\mathrm{E}[A(c_P)]}{\mathrm{E}[B(c_P)]} - \frac{\mathrm{Cov}[A(c_P), B(c_P)]}{(\mathrm{E}[B(c_P)])^2} + \cdots. \tag{2.12}$$

For both $\alpha_T(P, c_P)$ and $\alpha_R(P, c_P)$, the initialization bias term a arises from the fact that $\mathrm{E}[A(c_P)]/\mathrm{E}[B(c_P)] = \mathrm{E}[Y_i(c_P)]/\mathrm{E}[\tau_i(c_P)] \neq \alpha$. The ratio bias arises from the covariance term in equation (2.12). For $\alpha_T(P, c_P)$, $\mathrm{E}[B(c_P)] = \mathrm{E}[\tau_i(c_P)] \approx \lambda c_P$ and $\mathrm{Cov}[A(c_P), B(c_P)] = \mathrm{Cov}[Y_i(c_P), \tau_i(c_P)] \approx c_P \lambda A_{12}$ by the central limit theorem in equation (2.3) (and its uniform integrability). Thus the ratio bias for $\alpha_T(P, c_P)$ is $-A_{12}/(\lambda c_P)$ as stated. For $\alpha_R(P, c_P)$, the ratio bias is reduced by a factor of P since

$$\mathrm{Cov}[A(c_P), B(c_P)] = \mathrm{Cov}[\bar{Y}(c_P), \bar{\tau}(c_P)] = \frac{\mathrm{Cov}[Y_i(c_P), \tau_i(c_P)]}{P} \approx \frac{c_P \lambda A_{12}}{P}. \tag{2.13}$$

Combining equations (2.12), (2.13) and the expression for E $[B(c_P)]$ shows that the ratio bias of $\alpha_R(P, c_P)$ is $O(1/Pc_P)$ $(=o(1/c_P)$ as $P \to \infty)$.

The effect of the bias expansions of equation (2.7) is that, without deletion, $\alpha_T(P, c_P)$ and $\alpha_R(P, c_P)$ obey the following central limit theorems:

THEOREM 1. *For $i = 1, \ldots, P$, let $\{X_i(s), s \geq 0\}$ denote one of the iid sample paths of an irreducible, finite state space CTMC satisfying equations* (2.2) *and* (2.3). *Define $\sigma^2 = A_{11}$ and let $\kappa_P(c_P) = 0$ and $\gamma_i(c_P) = 0$. As $P \to \infty$,*

1. *If $P/c_P \to \infty$, $c_P \to \infty$ and $b_T \neq 0$, then $\sqrt{Pc_P}|\alpha_T(P, c_P) - \alpha| \Rightarrow \infty$.*
2. *If $P/c_P \to m$ $(0 < m < \infty)$ and $b_T \neq 0$, then $\sqrt{Pc_P}(\alpha_T(P, c_P) - \alpha) \Rightarrow N(0, \sigma^2) + b_T m^{1/2}$.*
3. *If $P/c_P \to 0$, then $\sqrt{Pc_P}(\alpha_T(P, c_P) - \alpha) \Rightarrow N(0, \sigma^2)$.*

THEOREM 2. *Theorem* 1 *is also valid for $\alpha_R(P, c_P)$ with b_R replacing b_T.*

Theorems 1 and 2 imply that, without deletion, one must let $P/c_P \to 0$ in order to obtain valid confidence intervals for α, i.e., the length of each replication must be large with respect to the number of replications (processors).

We next consider the case of asymptotically negligible deletion, i.e., $\kappa_P(c_P) \to \infty$ but $\kappa_P(c_P)/c_P \to 0$. In this case, it is shown in Glynn and Heidelberger (1989b) that

$$\mathrm{E}\,[\alpha_T(P, c_P)] = \alpha + \frac{d_T}{c_P} + o(1/c_P),$$

$$\mathrm{E}\,[\alpha_R(P, c_P)] = \alpha + o(1/c_P), \qquad \text{where} \tag{2.14}$$

$$d_T = -\frac{A_{12}}{\lambda}. \tag{2.15}$$

Equations (2.14) and (2.15) imply that, for $\alpha_T(P, c_P)$, initial transient deletion is effective in removing initialization bias, but does not remove ratio bias (unless $A_{12} = 0$ in which case simulated time and computer time are deterministically proportional). The effect of this bias expansion on the central limit theorem for $\alpha_T(P, c_P)$ is that valid confidence intervals will, again, only be obtained if $P/c_P \to 0$. On the other hand, initial transient deletion removes all sources of bias of order $1/c_P$ from the bias expansion of $\alpha_R(P, c_P)$. This will permit a valid central limit theorem for $\alpha_R(P, c_P)$ even if $P/c_P \to \infty$ provided the length of the deletion interval does not grow too slowly.

THEOREM 3. *For $i = 1, \ldots, P$, let $\{X_i(s), s \geq 0\}$ denote one of the iid sample paths of an irreducible, finite state space CTMC satisfying equations* (2.2) *and* (2.3). *Assume $\kappa_P(c_P) \to \infty$ and $\kappa_P(c_P)/c_P \to 0$. As $P \to \infty$,*

1. *If $P/c_P \to \infty$, $c_P \to \infty$ and $d_T \neq 0$, then $\sqrt{Pc_P}|\alpha_T(P, c_P) - \alpha| \Rightarrow \infty$.*
2. *If $P/c_P \to m$ $(0 < m < \infty)$ and $d_T \neq 0$, then $\sqrt{Pc_P}(\alpha_T(P, c_P) - \alpha) \Rightarrow N(0, \sigma^2) + d_T m^{1/2}$.*
3. *If $P/c_P \to 0$, then $\sqrt{Pc_P}(\alpha_T(P, c_P) - \alpha) \Rightarrow N(0, \sigma^2)$.*

THEOREM 4. *For $i = 1, \ldots, P$, let $\{X_i(s), s \geq 0\}$ denote one of the iid sample paths of an irreducible, finite state space CTMC satisfying equations* (2.2) *and* (2.3). *Assume $\kappa_P(c_P) \to \infty$ and $\kappa_P(c_P)/c_P \to 0$. As $P \to \infty$, if either*

1. *$P/c_P \to \infty$ and $\kappa_P(c_P)/\ln(P) \to \infty$, or*
2. *$P/c_P \to m$ $(0 < m < \infty)$ and $\kappa_P(c_P) \to \infty$, or*
3. *$P/c_P \to 0$*

then

$$\sqrt{Pc_P}(\alpha_R(P, c_P) - \alpha) \Rightarrow N(0, \sigma^2). \tag{2.16}$$

The ln (P) term of Theorem 4 arises because a finite state space CTMC converges exponentially fast to its steady-state distribution. As indicated earlier, the bias expansions and Theorems 1–4 are valid under more general conditions. Basically, one needs a functional version of the central limit theorem in equation (2.3), uniform integrability of second moments in this joint central limit theorem, exponential convergence to steady-state, and some sort of regularity conditions on $C_i(t)$ and $\{X_i(s), s \geq 0\}$. In Glynn and Heidelberger (1989b), it was assumed that the cumulative process $C_i(t)$ can be represented as an integral, i.e., $C_i(t) = \int_0^t \chi_i(s)ds$ where $\delta < \chi_i(s) < k$ for positive, finite constants δ and k. In addition it was assumed that $|X_i(s)| < k$ for a finite constant k and that $\{(X_i(s), \chi_i(s)), s \geq 0\}$ is a regenerative process. The regenerative assumption is not really as restrictive as it might seem since the estimation procedures do not make use of the regenerative structure. It is mainly used as a proof device, and, in addition, many stochastic processes possess a (hidden) regenerative structure (see, e.g., Glynn 1989). We further believe the result to be true for more general cumulative processes $\{C_i(t), t \geq 0\}$ where, e.g., $C_i(t)$ is discontinuous.

3. Queueing Models Used for Experimentation

In this section, we describe four queueing models that we used for determining, experimentally, the behavior of $\alpha_T(P, c)$ and $\alpha_R(P, c)$. These represent simplified versions of models (with analytically tractable solutions) that often arise in simulations of computer or communications systems. We ran experiments on the waiting time process in an $M/M/1$ queue and on three CTMCs: the queue length processes in an $M/M/1$ queue with feedback, an open Jackson network and a closed product form network (see, e.g., Kleinrock 1975).

For the $M/M/1$ waiting time simulations, we let ϕ be the arrival rate, μ be the service rate, and $\rho = \phi/\mu$ be the traffic intensity. Let W_n be the waiting time of the nth customer. If $\rho < 1$, there exists a (steady-state) rv W such that $W_n \Rightarrow W$ as $n \to \infty$. The performance measure of interest is $\alpha = \mathrm{E}[W] = \rho/[\mu(1 - \rho)]$, the mean steady-state waiting time.

For the $M/M/1$ queue with feedback, we let ϕ denote the arrival rate, μ the service rate and q the feedback probability. The expected number of visits a customer makes to the queue is $1/(1 - q)$ and the traffic intensity is $\rho = \phi/[\mu(1 - q)]$. We let $Q(s)$ denote the queue length at (simulated) time s, including the customer in service. Then if $\rho < 1$, there exists a (steady-state) rv Q such that $Q(s) \Rightarrow Q$ as $s \to \infty$. The output performance measure of interest is $\alpha = \mathrm{E}[Q] = \rho/(1 - \rho)$, the mean steady-state queue length. We set $\phi = 1$, $\mu = 20$, and $q = 0.9$, so that $\rho = 0.5$ and $\alpha = 1.0$. We ran experiments with two sets of initial conditions: $Q(0) = 0$ and $Q(0) = 5$.

A diagram of the open Jackson network is shown in Figure 1. This network is sometimes called an open central server model (see Buzen 1973) with server 0 representing a CPU (central processing unit), and servers 1 to 4 representing I/O (input/output) devices. There is a single type of job. Jobs arrive at the network (at the CPU) according to a Poisson process with rate ϕ. All servers operate using the FCFS (first-come-first-served) service discipline and the service times of jobs at server i are iid exponentially distributed rvs with mean $1/\mu_i$. When a job leaves the CPU, it goes to I/O device i with probability q_i $(1 \leq i \leq 4)$, and when a job leaves an I/O device, it goes to the CPU with probability q_0 or exits the system with probability $(1 - q_0)$. Let $Q_i(s)$ denote the queue length at server i at time s (including the customer in service) and let ρ_i denote the traffic intensity at server i. The total arrival rate at each server can be found by solving a system of flow balance equations (see p. 149 of Kleinrock 1975) and the ρ_i's can then be directly computed. In this particular case, $\rho_0 = \phi/[\mu_0(1 - q_0)]$ and $\rho_i = q_i\phi/[\mu_i(1 - q_0)]$ for $i \geq 1$. Provided $\rho_i < 1$ for all i, then there exists a (steady-state) random vector $(Q_0, \ldots, Q_4)$ such that $(Q_0(s), \ldots, Q_4(s)) \Rightarrow (Q_0, \ldots, Q_4)$ as $s \to \infty$. Under the above assumptions,

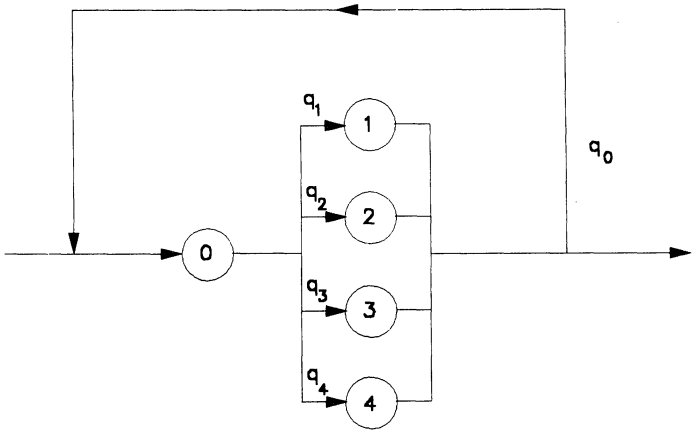

FIGURE 1. Open Central Server Model.

the distribution of $(Q_0, \ldots, Q_4)$ has a product form, and in particular $E[Q_i] = \rho_i/(1 - \rho_i)$. The output performance measure of interest is $\alpha = E[Q_0]$, the mean steady-state queue length at the CPU. We set $\phi = 1.0$, $q_0 = 0.75$, $q_i = 0.25$ for $i \geq 1$, $1/\mu_0 = 0.1875$, and $1/\mu_i = 0.50$ for $i \geq 1$. With these parameters, $\rho_0 = 0.75$, $\rho_i = 0.50$ for $i \geq 1$, $\alpha = E[Q_0] = 3.00$, and $E[Q_i] = 1.0$ for $i \geq 1$. The model was simulated with initial conditions $Q_i(0) = 0$ for all i.

The closed, product form queueing network model is shown in Figure 2. This model is sometimes called a closed central server model. Again server 0 represents a CPU and servers 1 to 4 represent I/O devices. There are a fixed number of jobs N circulating in the network. As in the open model, the service discipline is FCFS at all servers and we assume iid exponentially distributed service times with mean $1/\mu_i$ at server i. When a job leaves the CPU, it goes to I/O device i with probability q_i and when a job leaves an I/O device it goes back to the CPU. Let $Q_i(s)$ denote the queue length at server i at time s (including the customer in service) and let ρ_i denote the steady-state utilization of server i. Then there exists a random vector $(Q_0, \ldots, Q_4)$ such that $(Q_0(s), \ldots, Q_4(s)) \Rightarrow (Q_0, \ldots, Q_4)$ as $s \to \infty$. The performance measure of interest is again $\alpha = E[Q_0]$. A variety of numerical algorithms are available for computing such steady-state performance measures (see Buzen 1973 or Chapter 3 of Lavenberg 1983). We set $N = 10$, $q_i = 0.3$ for $1 \leq i \leq 3$, $q_4 = 0.1$, $1/\mu_0 = 1.0$, $1/\mu_i = 2.0$ for $1 \leq i \leq 3$, and $1/\mu_4$

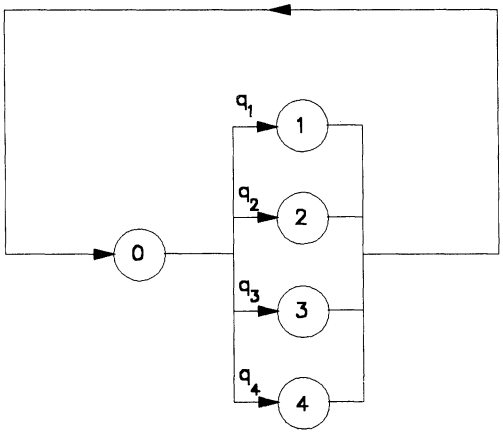

FIGURE 2. Closed Central Server Model.

$= 11.0$. With these parameters, $\rho_0 = 0.82$, $\rho_i = 0.49$ for $1 \le i \le 3$, $\rho_4 = 0.90$, $\alpha = \mathrm{E}[Q_0] = 3.06$, $\mathrm{E}[Q_i] = 0.92$ for $1 \le i \le 3$, and $\mathrm{E}[Q_4] = 4.17$. This model was simulated with initial conditions $Q_0(0) = 6$ and $Q_i(0) = 1$ for $i \ge 1$.

Because we did not have convenient access to a very highly parallel machine, all experiments were run on a single processor (an IBM 3090 mainframe). The effect of running parallel replications on multiple processors with a computer time stopping constraint was simulated as follows. For the queue length processes, we assumed that each event (external arrival or service completion) took one unit of computer time to process. Thus $C_i(t)$ is the number of events completed at simulated time t (on replication i) and $T_i(c)$ is the amount of simulated time generated after processing c events. Now the constant event time assumption is clearly an approximation. Whether or not this assumption holds is highly dependent on the model being simulated, the simulation software and the computer hardware. However, this assumption does capture and model many of the features assumed in our underlying model relating output performance measures, simulated time and CPU time. First, CPU time should be closely related to the number of events simulated. Second, with c representing the number of events, then the bivariate central limit theorem in equation (2.3) is indeed valid. While the integral representation $C_i(t) = \int_0^t \chi(X_i(s))ds$ is not strictly valid, as mentioned earlier, we believe Theorems 1–4 are valid under more general assumptions.

To validate the constant event time assumption in a particular case, we conducted the following experiment. We ran simulations of the open central server model on an IBM PS/2 Model 80. In these timing runs, we used the IBM Research Queueing (RESQ) package (see Gordon, MacNair, Gordon and Kurose 1990). (While our customized Jackson network simulator as described in §4 is much faster than RESQ on the 3090 thereby allowing larger sample sizes, it was much simpler to use RESQ on the PS/2 than to port our simulator from the 3090 to the PS/2 and then instrument it with the necessary timing routines.) We ran 20 iid replications, each for 120 seconds of CPU time. Since RESQ only checks the CPU time after a fixed number (1001) of events (an event is an arrival or service completion), each replication ran for slightly longer than 120 seconds. However every replication stopped after exactly the same number of events (38038). We then estimated the coefficient of variation (standard deviation divided by mean) of the final simulated times and of the actual CPU times required to generate this fixed number of events. The estimated coefficient of variation of the simulated times was 0.025 while the estimated coefficient of variation of the CPU times was 0.0007. This provides remarkably good validation of the constant event time model in this example.

The goal of the $M/M/1$ waiting time simulations is to compare the asymptotic loss in coverage predicted by Theorem 3 to finite sample simulation results for a model in which the ratio bias term d_T can be calculated analytically. To this end, we need to appropriately define what is meant by "simulation" time and "computer" time. Specifically, let $C_i(t)$ be the arrival time of customer number t and $T_i(c)$ be the number of customers that arrive in the interval $(0, c)$. Thus the "simulation" time at "computer" time c is the number of arrivals in $(0, c)$. In this example, the integrals in equations (2.3) and (2.4) get replaced by sums. For example, with no deletion ($\kappa_P(c) = 0$ and $\gamma_i(c) = 0$), then

$$\alpha_T(P, c) = \frac{1}{P} \sum_{i=1}^{P} \frac{1}{T_i(c) + 1} \sum_{t=0}^{T_i(c)} W_t(i) \tag{3.1}$$

where $W_t(i)$ is the waiting time of customer t on replication i. Notice that $T_i(c)$ is a rv so that $\alpha_T(P, c)$ involves ratios with random denominators. Thus ratio bias will be a factor in this setup.

Using these definitions for $C_i(t)$ and $T_i(c)$, the required limit theorems hold and the ratio bias term d_T can be calculated analytically, as follows. Let $\bar{W}_n = \sum_{k=0}^{n-1} W_k/n$ be the

average waiting time of the first n customers and let $\bar{A}_n = \sum_{k=1}^{n} A_k/n$ be the average interarrival time of the first n customers. (In the derivation that follows we drop the processor/replication subscript i.) Note that $\bar{A}_n = C_i(n)/n$. Since $T_i(n)/n \to \phi$ (the arrival rate), in this example we have $\phi = \lambda$. By regenerative process theory, we also have the bivariate central limit theorem

$$\sqrt{n}(\bar{W}_n - \alpha, \bar{A}_n - \lambda^{-1}) \Rightarrow \mathbf{N}(\mathbf{0}, \mathbf{B}) \tag{3.2}$$

for some covariance matrix $\mathbf{B}$. By Theorem 5 of Glynn and Heidelberger (1989b), $\mathbf{B}$ and $\mathbf{A}$ (the covariance matrix in equation (2.3)) are related by $A_{11} = B_{11}/\lambda$, $A_{12} = -\lambda B_{12}$, and $A_{22} = \lambda^3 B_{22}$. Thus by equation (2.15), $d_T = B_{12}$ $(= -A_{12}/\lambda)$. All the terms of $\mathbf{B}$ can be explicitly calculated (we set $\mu = 1$): $B_{11} = \rho[2 + 5\rho - 4\rho^2 + \rho^3]/(1-\rho)^4$ (see, e.g., Blomqvist 1967 or Whitt 1989), $B_{22} = \text{Var}[A_k] = 1/\lambda^2$, and, by using regenerative process theory (see, e.g., Crane and Iglehart 1975)

$$B_{12} = \frac{\text{Cov}[(Y_i - \alpha N_i), (t_i - \lambda^{-1} N_i)]}{\text{E}[N_i]} \tag{3.3}$$

where Y_i is the sum of the waiting times in the ith regenerative cycle, t_i is the sum of the interarrival times in the ith cycle and N_i is the number of arrivals in the ith cycle. Now $\text{E}[N_i] = 1/(1-\rho)$, while the covariance term in equation (3.3) can be calculated from results contained in Lavenberg, Moeller and Sauer (1977) (the research report version of Lavenberg, Moeller and Sauer 1979). (This covariance term arises in the context of control variables for simulation variance reduction.) Specifically, for $M/G/1$ queues $\text{Cov}[(Y_i - \alpha N_i), (t_i - N_i/\lambda)] = -b_2/[2(1-\rho)^3]$ where b_2 is the second moment of the service times. Since, for $M/M/1$ (with $\mu = 1$) $b_2 = 2$, equation (3.3) reduces to $d_T = B_{12} = -1/(1-\rho)^2$.

The effect of ratio bias on confidence interval coverage can now be calculated analytically. For given values of P and c_P, deciding on which part of Theorem 3 to apply depends on whether one interprets the value P/c_P as being "nearly infinite", fixed, or nearly zero. We will assume that part (2) applies, and use the value $m = P/c_P$. Notice that if $m \approx 0$ (and $d_T\sqrt{m} \approx 0$) then parts (2) and (3) of Theorem 3 are, practically speaking, identical. Similarly, if m and $d_T\sqrt{m}$ are very large, then parts (1) and (2) of Theorem 3 are, practically speaking, identical. With this in mind, let $\Phi(x) \equiv \text{P}\{N(0,1) \le x\}$ and define $z_{\delta/2}$ by $\Phi(z_{\delta/2}) = 1 - \delta/2$. From part (2) of Theorem 3, if $P/c_P = m$, and $\beta = d_T\sqrt{m/A_{11}}$, then

$$\begin{aligned}\text{P}\{\sqrt{Pc_p}\,|\,\alpha_T(P, c_P) - \alpha|/\sqrt{A_{11}} \le z_{\delta/2}\} &\approx \text{P}\{|N(0,1) + \beta| \le z_{\delta/2}\} \\ &= \Phi(z_{\delta/2} - \beta) - \Phi(-z_{\delta/2} - \beta). \end{aligned} \tag{3.4}$$

Thus for any given ρ, P, and c_P, the coverage of presumed $100 \times (1-\delta)\%$ confidence intervals can be predicted. In §5, we will compare the predicted coverage with the sample coverage observed in simulation experiments. Note that by using the heavy traffic approximation $B_{11} \approx 4\rho/(1-\rho)^4$ (see Whitt 1989) we obtain $\beta \approx -0.5\sqrt{m}$. This approximation also works well for moderate values of ρ. Thus for given P and c_P, we expect the loss in coverage due to ratio bias to be approximately independent of the traffic intensity (provided ρ is not too small and P and c_P are large enough that the central limit theorem is valid). This behavior will be observed in §5.

4. Design of the Simulation Experiments

In this section we describe how the simulation experiments were performed. As mentioned earlier, the effect of running parallel replications on multiple processors with a computer time stopping constraint was simulated on a single processor. For the various models, and different values of P, c and $\kappa_P(c)$, we were interested in estimating the mean, variance and confidence interval coverage of $\alpha_T(P, c)$ and $\alpha_R(P, c)$. We built a simple queueing network simulator suitable for these purposes. (We used the combined generator

described in L'Ecuyer 1988 as a source of random numbers.) To estimate these quantities for given values of P, c, and $\kappa_P(c)$, M "super replications" were performed where each super replication consisted of P replications, each of length c and having truncation interval $\kappa_P(c)$. Thus for super replication j ($1 \le j \le M$) samples $\alpha_T(P, c, j)$ and $\alpha_R(P, c, j)$ of $\alpha_T(P, c)$ and $\alpha_R(P, c)$, respectively, were obtained (according to equations (2.5) and (2.6). E $[\alpha_T(P, c)]$ and E $[\alpha_R(P, c)]$ were estimated by

$$\bar{\alpha}_T(P, c) \equiv \sum_{j=1}^{M} \alpha_T(P, c, j)/M \quad \text{and} \quad \bar{\alpha}_R(P, c) \equiv \sum_{j=1}^{M} \alpha_R(P, c, j)/M,$$

respectively. The sample standard deviations, $S_T(P, c)$ and $S_R(P, c)$ of $\bar{\alpha}_T(P, c)$ and $\bar{\alpha}_R(P, c)$, respectively, were computed in the usual way, e.g.,

$$S_T^2(P, c) = \sum_{j=1}^{M} (\alpha_T(P, c, j) - \bar{\alpha}_T(P, c))^2/[M(M-1)].$$

On each super replication we also obtained asymptotic standard deviation estimates $\hat{\sigma}_T(P, c, j)$ and $\hat{\sigma}_R(P, c, j)$ for $\alpha_T(P, c)$ and $\alpha_R(P, c)$, respectively. These were estimated as follows. Let $Y_i(c, j)$ and $\tau_i(c, j)$ be the samples of $Y_i(c)$ and $\tau_i(c)$ obtained in the jth super replication. Then

$$\hat{\sigma}_T^2(P, c, j) \equiv \frac{\sum_{i=1}^{P} (Y_i(c, j)/\tau_i(c, j) - \alpha_T(P, c, j))^2}{P-1}. \tag{4.1}$$

Computation of $\hat{\sigma}_R^2(P, c, j)$ is analogous to variance estimation in regenerative simulation:

$$\hat{\sigma}_R^2(P, c, j) \equiv \frac{(1/P) \sum_{i=1}^{P} (Y_i(c, j) - \alpha_R(P, c, j)\tau_i(c, j))^2}{((1/P) \sum_{i=1}^{P} \tau_i(c, j))^2}. \tag{4.2}$$

From these point and variance estimates, presumed $100 \times (1-\delta)\%$ confidence intervals for α can be formed as follows. Using the traditional estimator the confidence interval for the jth super replication is $\alpha_T(P, c, j) \pm t_{\delta/2}(P-1)\hat{\sigma}_T(P, c, j)/\sqrt{P}$ where $t_{\delta/2}(P-1)$ is defined by $1 - \delta/2 = \mathrm{P}\{Z \le t_{\delta/2}(P-1)\}$ and Z has a Student's t distribution with $(P-1)$ degrees of freedom. Using the classical ratio estimator the confidence interval for the jth super replication is $\alpha_R(P, c, j) \pm z_{\delta/2}\hat{\sigma}_R(P, c, j)/\sqrt{P}$. (This is analogous to forming confidence intervals in regenerative simulation.) For a given estimator, we define its coverage to be the fraction of these confidence intervals that actually contain α. If valid confidence intervals are being formed for α, then, by definition, the coverage should converge to $(1-\delta)$ as $M \to \infty$. In all cases we set $\delta = 0.1$ corresponding to 90% confidence intervals.

The simulator was organized in such a way that statistics could be collected for multiple values of P, c and $\kappa_P(c)$ from the same set of runs. Thus the data generated for a particular model are correlated. We took values of P to be powers of two, ranging from $P = 8$ to $P = 512$ for the CTMCs and $P = 128$ to $P = 1024$ for the $M/M/1$ waiting time simulations. (We begin reporting the $M/M/1$ waiting time experiments at $P = 128$, since the coverage loss is negligible for smaller values of P.) Each super replication for P processors also comprised 2 super replications for $P/2$ processors, 4 super replications for $P/4$ processors, etc. We used 200 super replications for the largest values of P in each case. Thus, e.g., 12,800 super replications of the CTMCs were obtained for $P = 8$. These sample sizes were generally large enough so that very accurate point estimates were obtained.

5. Experimental Results

The first set of experiments are for the $M/M/1$ waiting times. The purpose of these experiments is to compare the analytic results of §3 to actual simulation results. To

concentrate on the effect of the ratio bias, these simulations were started in the steady-state distribution. We simulated until (simulated) time $c = 1{,}000$. By deleting customers arriving before times $\kappa_P(c) = 100$, 250 and 500, we obtained runs of effective lengths $c - \kappa_P(c) = 900$, 750 and 500, respectively. Notice that even though the simulation is started in the steady-state distribution, as discussed earlier, the initialization bias term a may still be nonzero because the number of customers arriving in $(0, c)$ is a rv. Therefore, according to Theorems 2 and 4, we need to delete some initial data in order to remove a from the bias expansion of $\alpha_R(P, c)$. Therefore, we do not include the no deletion case, $\kappa_P(c) = 0$, in Table 1. We simulated at $\rho = 0.50$ and $\rho = 0.75$.

The results of these experiments are listed in Table 1. Table 1 lists the predicted coverages for $\alpha_T(P, c)$ as calculated by equation (3.4) (using the effective run length for c in that equation), as well as the sample coverages for $\alpha_T(P, c)$ and $\alpha_R(P, c)$ observed in the simulations. Table 1 indicates generally excellent agreement between the predictions and the experiments. Notice that, for a given P and $\kappa_P(c)$, the predicted and sample coverage for $\alpha_T(P, c)$ is quite insensitive to the value of ρ, as explained in §3. In addition, for fixed $\kappa_P(c)$, as P increases the coverage for $\alpha_T(P, c)$ decreases. This is in agreement with part (2) of Theorem 3 and is explained by the fact that as P increases, increasingly accurate estimates of (the biased) $\mathrm{E}\,[\alpha_T(P, c)]$ are obtained. This loss in coverage is greatest for the largest value of $\kappa_P(c)$ since that corresponds to the smallest effective run length. On the other hand, the coverage for $\alpha_R(P, c)$ stays close to its nominal value of 0.90.

Figures 3 to 5 plot results from simulations of the $M/M/1$ queue with feedback. Figure 3 plots $\bar{\alpha}_T(P, c)$ and $\bar{\alpha}_R(P, c)$ as functions of $\kappa_P(c)$ for $c = 1000$ events, $P = 512$, and two different initial queue lengths. Actually, when $Q(0) = 0$, $\bar{\alpha}_T(P, c)$ appears almost unbiased without truncation ($\kappa_P(c) = 0$), but $\bar{\alpha}_T(P, c)$ increases above the steady-state value of $\alpha = 1.0$ as $\kappa_P(c)$ increases. In this case, the initialization bias and ratio bias are of opposite signs and, in effect, approximately cancel each other out when $\kappa_P(c) = 0$. When $Q(0) = 5$, $\bar{\alpha}_T(P, c)$ decreases as $\kappa_P(c)$ increases, but, again, does not come close to α. For $\kappa_P(c) = 500$ the values of $\bar{\alpha}_T(P, c)$ are very nearly the same for both $Q(0) = 0$ and $Q(0) = 5$, but are about 8% above the steady-state value. On the other hand, $\bar{\alpha}_R(P, c)$ approaches α as $\kappa_P(c)$ increases for both $Q(0) = 0$ and $Q(0) = 5$. These point

TABLE 1

Predicted and Sample 90% Confidence Interval Coverages in M/M/1 Queue Waiting Time Simulations with $c = 1000$

		$\rho = 0.5$			$\rho = 0.75$		
$c - \kappa_P(c)$	P	Predicted $\alpha_T(P, c)$	Sample $\alpha_T(P, c)$	Sample $\alpha_R(P, c)$	Predicted $\alpha_T(P, c)$	Sample $\alpha_T(P, c)$	Sample $\alpha_R(P, c)$
500	128	0.888	0.874	0.894	0.889	0.871	0.898
	256	0.876	0.856	0.892	0.878	0.862	0.908
	512	0.852	0.848	0.909	0.856	0.875	0.912
	1024	0.806	0.795	0.930	0.813	0.850	0.920
750	128	0.892	0.888	0.903	0.893	0.885	0.898
	256	0.884	0.876	0.894	0.885	0.882	0.902
	512	0.868	0.870	0.900	0.871	0.870	0.912
	1024	0.837	0.845	0.915	0.842	0.865	0.950
900	128	0.893	0.886	0.900	0.894	0.882	0.898
	256	0.887	0.881	0.880	0.888	0.876	0.896
	512	0.874	0.880	0.900	0.876	0.865	0.912
	1024	0.847	0.855	0.930	0.851	0.870	0.920

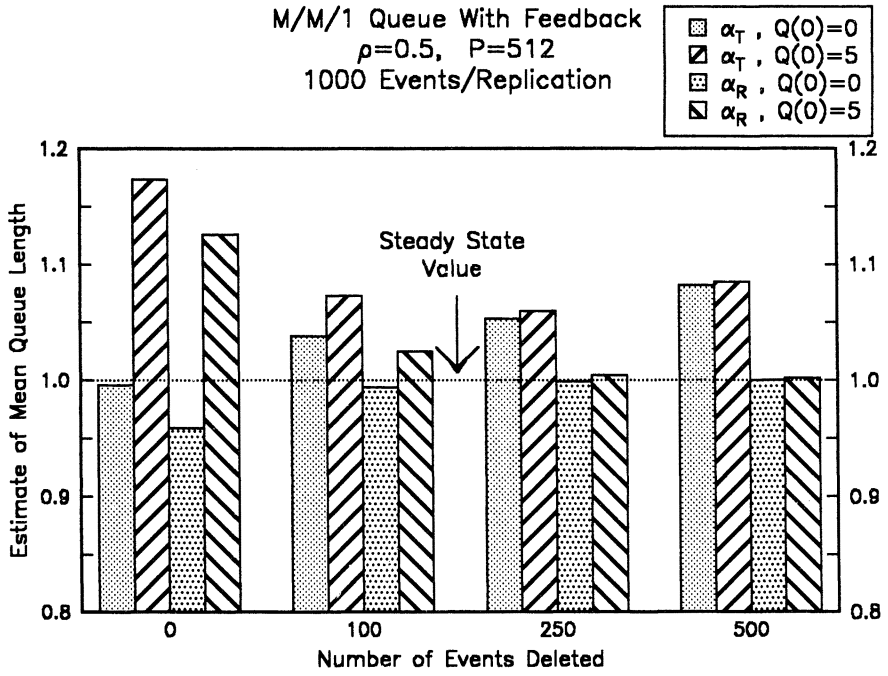

FIGURE 3. Estimated Mean Queue Length in $M/M/1$ Queue with Feedback.

estimates are very accurate. For example, when $Q(0) = 0$ and $\kappa_P(c) = 500$, $\bar{\alpha}_T(512, c) = 1.084$, $S_T(512, c) = 0.002$, $\bar{\alpha}_R(512, c) = 1.002$ and $S_R(512, c) = 0.002$.

Figure 4 plots the coverages for these estimators (without deletion) as a function of P. Because, by coincidence, $\mathrm{E}[\alpha_T(P, c)] \approx \alpha$ when $Q(0) = 0$ and $\kappa_P(c) = 0$, the coverage

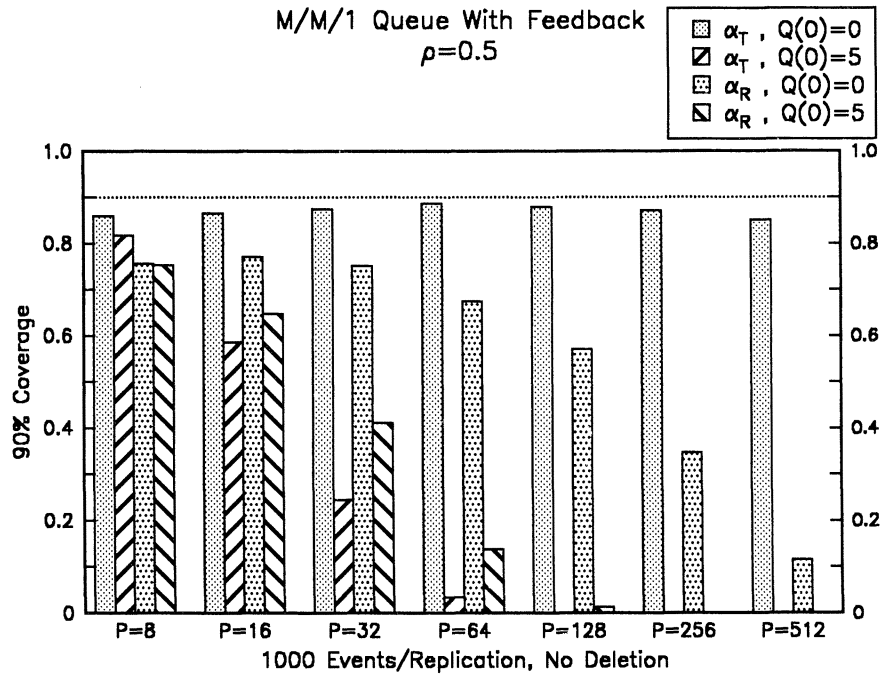

FIGURE 4. Estimated 90% Coverage in $M/M/1$ Queue with Feedback (without Deletion).

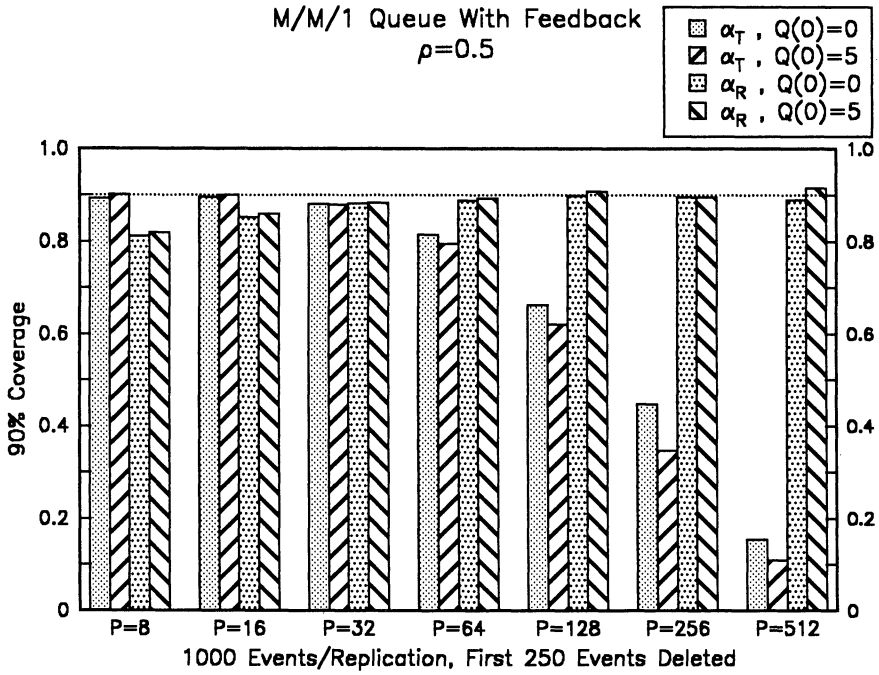

FIGURE 5. Estimated 90% Coverage in $M/M/1$ Queue with Feedback (with Deletion).

for $\alpha_T(P, c)$ remains at or near the nominal value of 0.90. However, the coverage for $\alpha_T(P, c)$ decreases (to zero) as P increases when $Q(0) = 5$ because of the stronger initialization bias. Similarly, because of initialization bias, the coverage for $\alpha_R(P, c)$ is seriously degraded for both $Q(0) = 0$ and $Q(0) = 5$.

Figure 5 shows the coverages when $\kappa_P(c) = 250$. With this value of $\kappa_P(c)$, the initialization bias is essentially eliminated, although ratio bias is still present: for example, when $Q(0) = 5$, $\bar{\alpha}_R(512, c) = 1.004$ compared to $\alpha = 1.0$ while $\bar{\alpha}_T(512, c) = 1.060$. Because of the ratio bias, the coverage for $\alpha_T(P, c)$ decreases from around 0.90 to less than 0.20 as P increases from 8 to 512 for both initial conditions. Significant coverage loss begins to be observed in the range from $P = 32$ to $P = 64$. On the other hand, the coverage for $\alpha_R(P, c)$ starts out slightly below 0.90 for $P = 8$ and then rapidly approaches 0.90 as P increases. The low coverage when $P = 8$ is due both to a less robust variance estimate as well as to the use of a normal multiplier, rather than a (larger) t-multiplier, in the confidence interval. For example, when $Q(0) = 0$ and a t-multiplier with 7 degrees of freedom is used instead of the normal multiplier, the coverage for $\alpha_R(8, P)$ increases from 0.820 to 0.864.

Figures 6 and 7 display results of simulating the open central server model. The network was simulated for $c = 2500$ events. Figure 6 plots $\bar{\alpha}_T(P, c)$ and $\bar{\alpha}_R(P, c)$ as functions of $\kappa_P(c)$ for $P = 8$, 64 and 512. (Because of the organization of the simulator's data collection facilities, for fixed c, the value of $\bar{\alpha}_T(P, c)$ is the same for all values of P.) Initialization bias is essentially eliminated by $\kappa_P(c) = 1000$, but significant ratio bias is still evident in $\bar{\alpha}_T(P, c)$. Note also that there are only slight differences between $\bar{\alpha}_R(8, c)$, $\bar{\alpha}_R(64, c)$ and $\bar{\alpha}_R(512, c)$. Because of the initialization bias, without deletion, the coverage for both $\alpha_T(P, c)$ and $\alpha_R(P, c)$ are well below the 0.90 level: the coverage for $\alpha_T(64, c)$ is 0.627 while the coverage for $\alpha_R(64, c)$ is 0.457. Note that when $\kappa_P(c) = 250$, $\alpha_T(P, c)$ is, by chance, almost unbiased. Thus, with this amount of deletion, the coverage for $\alpha_T(P, c)$ will be (approximately) correct, but for the wrong reason. For example, $\alpha_T(512, c)$ has coverage 0.91 while $\alpha_R(512, c)$ has coverage of only 0.60.

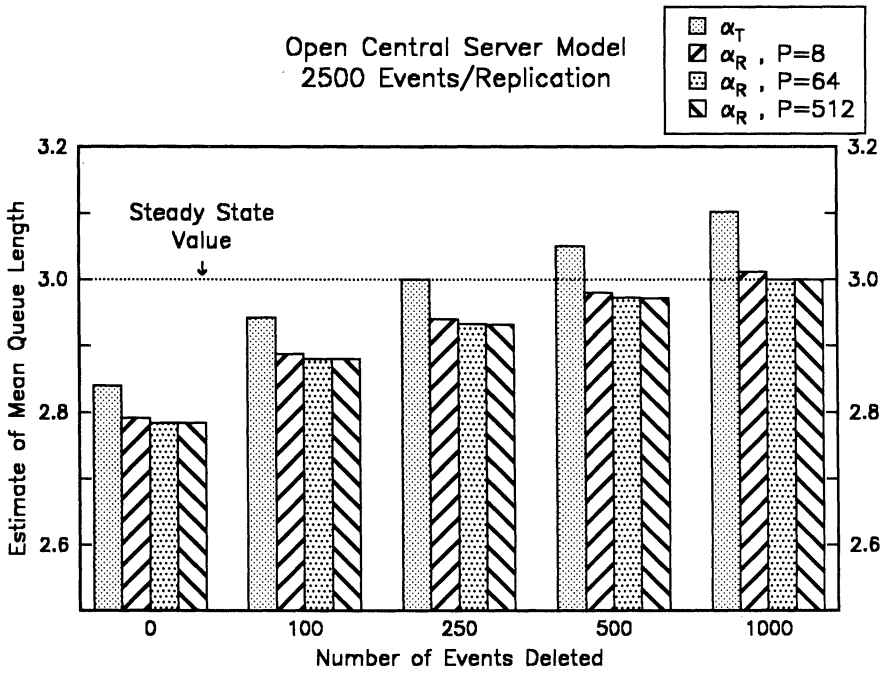

FIGURE 6. Estimated Mean Queue Length in Open Central Server Model.

Figure 7 plots the coverages for $\alpha_T(P, c)$ and $\alpha_R(P, c)$ as a function of P when initialization bias is essentially removed ($\kappa_P(c) = 1000$). Again, the coverage for $\alpha_T(P, c)$ decreases as P increases, while the coverage for $\alpha_R(P, c)$ increases to and then remains at or near the nominal 0.90 level.

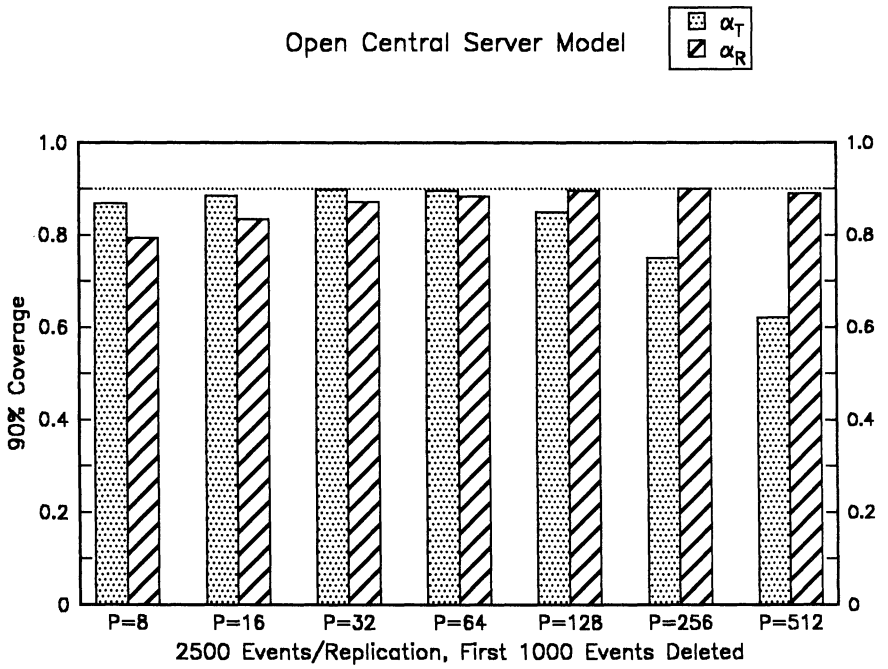

FIGURE 7. Estimated 90% Coverage in Open Central Server Model.

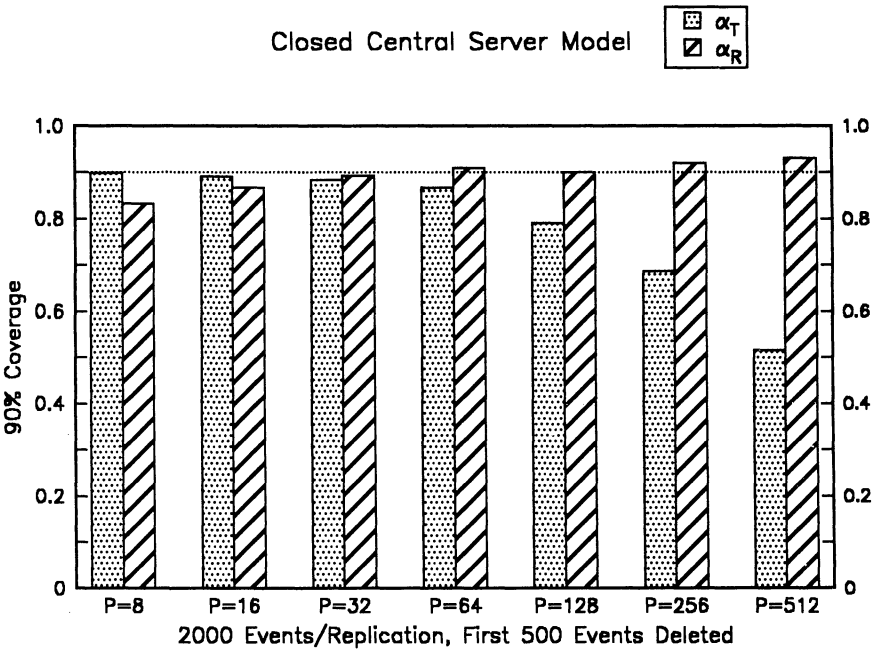

FIGURE 8. Estimated 90% Coverage in Closed Central Server Model.

Figure 8 displays a similar pattern for simulations of the closed central server model. This figure plots coverage results when $c = 2000$ and $\kappa_P(c) = 500$. With these parameters, initialization bias is removed but ratio bias is still present. The steady-state value being estimated is $\alpha = 3.057$, and $\bar{\alpha}_R(512, c) = 3.056$ ($S_R(512, c) = 0.002$) while $\bar{\alpha}_T(512, c) = 3.111$ ($S_T(512, c) = 0.002$).

As has been indicated several times above, for given values of P, c and $\kappa_P(c)$, the values of $S_T(P, c)$ and $S_R(P, c)$ have been very nearly the same. This has been observed throughout our experiments. This is explained by the fact that, even with ratio bias still present, $\alpha_T(P, c)$ and $\alpha_R(P, c)$ both obey central limit theorems with the same asymptotic standard deviation (see Theorems 3 and 4).

6. Summary and Conclusions

This paper has considered the problem of estimating steady-state parameters on multiple processors via the method of parallel replications. While the method is conceptually straightforward to apply, statistical considerations point to the need for using an alternative steady-state estimation procedure. This need arises because the traditional estimator, $\alpha_T(P, c)$, contains two sources of bias having the same order of magnitude: initialization bias and ratio bias. While appropriate deletion of an initial portion of each simulation effectively removes initialization bias, it does not affect the ratio bias. When using a large number of processors, this residual ratio bias results in a biased estimate and corresponding loss in confidence interval coverage.

The alternative estimator, $\alpha_R(P, c)$, corresponds to the classical ratio estimator that is commonly used in regenerative simulation. Its ratio bias is order P times smaller than its initialization bias. Thus appropriate deletion is effective in removing the major source of bias. The net effect is that using $\alpha_R(P, c)$ rather than $\alpha_T(P, c)$ allows one to either:

1. use many more processors for a given amount of computing time per processor, or
2. make shorter runs for a given number of processors.

This paper examined these issues empirically via simulations of a variety of queueing systems. Our experiments confirm the theoretical results, and indicate that the ratio bias can become a problem even on moderately sized parallel processors with, say, 32 to 64 processors.

The results of this paper also have some applicability to the standard single-processor method of independent replications. In this method, the replication length is often determined by either the total number of events, a simulated time limit, a computer time limit, or the number of events of a particular type such as the number of departures from a queue. (Sometimes a combination of these limits is used.) When estimating many parameters in a queueing network, there will always be some parameters that are estimated on a different time scale than that used to determine the replication length. Thus the denominator of some parameter estimates will be random, resulting in ratio bias. For example, if simulated time is used to control the replication length, then response time estimates will have a random denominator (the number of customers departing from the queue). On the other hand, if an event count is used to control the replication length, then queue length estimates will have a random denominator (the simulated time). Thus ratio bias could be a concern, even on a single processor. However, there is usually little motivation to run a very large number of short replications on a single processor, since either batch means or running a few long replications will be less sensitive to initialization bias. Nevertheless, the issue of ratio bias should be kept in mind. In fact, for a small number of replications, the ratio form of $\alpha_R(P, c)$ suggests the use of jackknifing (see Miller 1974) for both (ratio) bias reduction and for robust variance estimation. However, the properties and validity of jackknifing in this situation have not yet been established, and remain as open problems for research.

In addition, if the replication length is determined within a sequential procedure (see, e.g., Law and Kelton 1982), then the denominator of the estimates will typically be random resulting in possible ratio bias. This will also be true if the length of the truncation interval is determined by statistical tests of the simulation output (see, e.g., Schruben 1982). The effect of ratio bias in these situations also has yet to be analyzed.[1]

[1] The work of Peter Glynn was supported by the IBM Corporation under SUR-SST Contract 12480042 and by the U.S. Army under contract number DAAL-03-88-K-0063. We wish to thank Bob Gordon for his assistance with the RESQ timing runs. We are also grateful to the Area Editor, Associate Editor and an anonymous referee for their many comments which helped to improve the paper.

References

BHAVSAR, B. C. AND J. R. ISAAC, "Design and Analysis of Parallel Monte Carlo Algorithms," *SIAM J. Scientific and Statist. Computing*, 8 (1987), s73–s95.

BILLINGSLEY, P., *Convergence of Probability Measures*, Wiley, New York, 1968.

BLOMQVIST, N., "The Covariance Function of the $M/G/1$ Queueing System," *Skand. AktuarTidskr.*, 8 (1967), 157–174.

BRATLEY, P., B. L. FOX AND L. E. SCHRAGE, *A Guide to Simulation*, (Second Ed.), Springer-Verlag, New York, 1987.

BUZEN, J. P., "Computational Algorithms for Closed Queueing Networks with Exponential Servers," *Comm. ACM*, 16 (1973), 527–531.

COCHRAN, W. G., *Sampling Techniques*, (Second Ed.), Wiley, New York, 1963.

CRANE, M. A. AND D. L. IGLEHART, "Simulating Stable Stochastic Systems. III. Regenerative Processes and Discrete Event Simulations," *Oper. Res.*, 23 (1975), 33–45.

FOX, B. L. AND P. W. GLYNN, "Estimating Time Averages via Randomly Spaced Observations," *SIAM J. Appl. Math.*, 47 (1987), 186–200.

FUJIMOTO, R. M., "Parallel Discrete Event Simulation," *Comm. ACM*, 33, 10 (1990), 31–53.

GLYNN, P. W., "Limit Theorems for the Method of Replication," *Stochastic Models*, 4 (1987), 344–350.

———, "A GSMP Formalism for Discrete Event Systems," *Proc. IEEE*, 77 (1989), 14–23.

———, "A Low Bias Steady-State Estimator for Equilibrium Processes," Technical Report, Department of Industrial Engineering, University of Wisconsin, Madison, WI, 1990.

GLYNN, P. W. AND P. HEIDELBERGER, "Analysis of Initial Transient Deletion for Replicated Steady-State Simulations," IBM Research Report RC 15259, Yorktown Heights, NY, 1989a. *Oper. Res. Lett.*, (to appear).

——— AND ———, "Analysis of Initial Transient Deletion for Parallel Steady State Simulations," IBM Research Report RC 15260, Yorktown Heights, NY, 1989b. *SIAM J. Scientific and Statist. Computing*, (to appear).

——— AND ———, "Analysis of Parallel Replicated Simulations Under a Completion Time Constraint," *ACM Trans. Computer Simulation*, 1, 1 (1991), 3–23.

GORDON, R. F., E. A. MACNAIR, K. J. GORDON AND J. F. KUROSE, "Hierarchical Modeling in a Graphical Simulation System," *Proc.* 1990 *Winter Simulation Conf.*, O. Balci, R. P. Sadowski and R. E. Nance (Eds.). IEEE Press, 1990, 499–503.

HEIDELBERGER, P., "Statistical Analysis of Parallel Simulations," 1986 *Winter Simulation Conf. Proc.*, J. Wilson and J. Henriksen (Eds.), IEEE Press, 1986, 290–295.

———, "Discrete Event Simulations and Parallel Processing: Statistical Properties," *SIAM J. Scientific and Statist. Computing*, 9 (1988), 1114–1132.

IGLEHART, D. L., "Simulating Stable Stochastic Systems. V. Comparison of Ratio Estimators," *Naval Res. Logist. Quart.*, 22 (1975), 553–565.

KLEINROCK, L., *Queueing Systems. Vol.* 1. *Theory*, Wiley, New York, 1975.

LAVENBERG, S. S. (ED.), *Computer Performance Modeling Handbook*, Academic Press, New York, 1983.

———, T. L. MOELLER AND C. H. SAUER, "Concomitant Control Variables Applied to the Regenerative Simulation of Queueing Systems," IBM Research Report RC 6413. Yorktown Heights, NY, 1977.

———, ——— AND ———, "Concomitant Control Variables Applied to the Regenerative Simulation of Queuing Systems," *Oper. Res.*, 27 (1979), 134–160.

LAW, A. M., "Confidence Intervals in Discrete Event Simulation: A Comparison of Replication and Batch Means," *Naval Res. Logist. Quart.*, 24 (1977), 667–678.

——— AND J. S. CARSON, "A Sequential Procedure for Determining the Length of Steady-State Simulations," *Oper. Res.*, 27 (1979), 1011–1025.

——— AND W. D. KELTON, "Confidence Intervals for Steady-State Simulations. II. A Survey of Sequential Procedures," *Management Sci.*, 28 (1982), 550–562.

LEADBETTER, M. R., G. LINDGREN AND H. ROOTZÉN, *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, New York, 1983.

L'ECUYER, P., "Efficient and Portable Combined Random Number Generators," *Comm. ACM*, 31 (1988), 742–749 and 774.

MILLER, R. G., "The Jackknife—A Review," *Biometrika*, 61 (1974), 1–15.

SCHRUBEN, L. W., "Detecting Initialization Bias in Simulation Output," *Oper. Res.*, 30 (1982), 569–590.

WHITT, W., "Planning Queueing Simulations," *Management Sci.*, 35 (1989), 1341–1366.