SIAM J. SCI. STAT. COMPUT.
Vol. 13, No. 4, pp. 904–922, July 1992

005

ANALYSIS OF INITIAL TRANSIENT DELETION FOR PARALLEL STEADY-STATE SIMULATIONS*

PETER W. GLYNN† AND PHILIP HEIDELBERGER‡

Abstract. This paper investigates theoretical properties of a simple method for using parallel processors in discrete event simulations: running independent replications, in parallel, on multiple processors and averaging the results at the end of the runs. Specifically, the problem of estimating steady-state parameters from such an experiment is considered. Sampling plans are considered in which the replication lengths are given by limits on either simulated or computer time, and in which the beginning portion of each run may be deleted for the purpose of controlling initialization bias. The critical relative growth rates for the number of processors, the length of each replication, and the length of the deletion period that are required in order to produce valid confidence intervals for steady-state parameters are determined. When the replication length is determined by computer time, the straightforward estimator with deletion may not work for a large number of processors. In this case, the deletion is essentially useless due to an additional bias term that arises because the simulated time at the end of a replication is random. In this case, a new estimator can be used to remove this source of bias.

Key words. simulation, parallel processing, steady-state, initial transient

AMS(MOS) subject classifications. 68U20, 65C05, 60F05, 60K05, 60K20, 60K30

1. Introduction.

A simple way to exploit the power of parallel processing in computationally intensive discrete event simulations is to run multiple independent replications, in parallel, on the processors and to appropriately average the results at the end of the runs. At first glance, this method produces a p-fold speedup, i.e., reduction in completion time, over a sequential (one processor) simulation having the same variance where p is the number of processors. We call this the parallel replications approach.

An alternative approach is distributed simulation, in which all p processors cooperate on a single realization of the simulation. While significant speedups have been achieved using distributed simulation in specific problem domains (see, e.g., Fujimoto (1989, 1990); Goli, Heidelberger, Towsley, and Yu (1990); Lubachevsky (1989); Nicol (1988); Unger and Fujimoto (1989); Unger and Jefferson (1988); and Yu, Towsley, and Heidelberger (1989)), in our opinion, distributed simulation has not yet been demonstrated to be a robust and effective general purpose technique for dealing with the types of complex models arising in manufacturing, computer, and communications systems.

In contrast, parallel replications are conceptually and practically simple to apply and are almost universally applicable. The widespread applicability stems from the fact that a major reason why many models must run for a long time is the slow rate at which a simulation estimate's variance decreases. Essentially, parallel replications are inappropriate only when either

*Received by the editors December 20, 1989; accepted for publication (in revised form) March 11, 1991. This research was supported by the IBM Corporation under Shared University Research contract 12480042, by U.S. Army Research Office contract DAAL-03-88-K-0063, and by a grant from the Natural Sciences and Engineering Research Council of Canada.

†Department of Operations Research, Stanford University, Stanford, California 94305.

‡IBM Thomas J. Watson Research Center, P.O. Box 704, Yorktown Heights, New York 10598.

1. The model is such that a single replication cannot be completed on a single processor within a reasonable amount of time. This may be due to either an exceptionally large model, which, e.g., may not fit into the memory of a single processor, or, in steady-state simulations, to a model with a very slowly dissipating initial transient.
2. The variance from each replication is very small, in which case the output is nearly deterministic and having a large number of replications is merely a waste of computing resources.

A simple analytic model for comparing the statistical efficiencies of distributed simulation and parallel replications that justifies the above conclusion was developed in Heidelberger (1986).

However, there are some potentially serious statistical problems associated with the parallel replications approach, especially for a large number of processors. In the case of estimating expected values of so-called transient quantities, these problems have been studied in Heidelberger (1988) and Glynn and Heidelberger (1991a). The source of the problem can be illustrated as follows. Suppose that replications are run on each of p processors and that one sets a completion time constraint of t units of (computer) time per processor. The number of replications completed on each processor by time t is a random variable. While there are a number of different ways the output can be averaged, there is some sampling bias due to the fixed completion time t and the associated random number of replications. If t remains fixed, then as $p \to \infty$, the estimates can converge to the wrong quantity. A variety of estimators that alleviate this situation can be devised, but they have the property that some or all of the replications in progress at time t must be completed before the estimator can be formed. Thus one must pay a completion time penalty in order to obtain the correct convergence behavior. Even in the case of a single processor, some care needs to be taken in order to best handle the bias associated with stopping the simulation at time t; see Meketon and Heidelberger (1982), Glynn (1989b), and Glynn and Heidelberger (1990). In our discussion, the above results are described primarily in terms of estimating transient quantities, but they also apply to steady-state estimation in regenerative simulations (see Smith (1955) or Crane and Iglehart (1975)) since, in this case, steady-state performance measures can be expressed as a ratio of expected values of "transient" quantities. Bhavsar and Isaac (1987) discuss other properties associated with parallel replication schemes for transient quantities.

In this paper, we consider the parallel replications approach to the steady-state estimation problem in more generality. Specifically, we consider sampling schemes that delete some initial part of each run in order to reduce "initialization bias," i.e., bias due to the fact that the model cannot typically be started in its steady-state distribution (otherwise, there would be no need to simulate). We determine the critical relative growth rates for

1. the number of processors (replications),
2. the length of each replication, and
3. the length of the deletion period

that are required in order for the method of parallel replications with initial transient deletion to obey a usable central limit theorem. By a usable central limit theorem, we mean one that is centered about the unknown steady-state parameter and upon which confidence intervals for the steady-state parameter can be based.

Determining the length of each replication is a basic issue in such an experiment. We consider two approaches: the first, based on simulated time, and the second, based on computer time. The approach chosen makes a difference, in terms of both completion time and estimator bias. Deletion helps to reduce bias in the first approach;

however, the completion time is a random variable. On the other hand, when the replication length is based on computer time, the completion time is deterministic. However, an additional source of estimator bias is introduced in this case. Furthermore, this additional source of bias is not eliminated by deletion, and thus initial transient deletion is essentially useless in this case. This bias, which is of order one over the replication length, is due to the fact that the estimate from each processor takes the form of a ratio of two random variables. The denominator in this ratio is the (random) length of the replication in simulated time (minus the initialization period). When the run is based on simulated time, the denominator is deterministic and this additional bias is not introduced. A new estimator (also based on computer time) that gets around this problem is proposed and analyzed. This estimator has a deterministic stopping time and benefits from the initial transient deletion.

Some comments about the analysis techniques used in this paper are appropriate. First, we state limit theorems in a triangular array setting, i.e., we simultaneously let $p \to \infty$ and $t(p) \to \infty$ where p is the number of processors and $t(p)$ is the length of each replication. Since in practice only a fixed, finite number of processors are available, these results should be interpreted as determining (qualitatively) appropriate values for $t(p)$ in order to obtain proper convergence behavior for large values of p. Second, many of the results are established under regenerative assumptions that would appear to limit the applicability of the results. However, this is done mainly as a mathematical convenience and does not impose significant restrictions since

1. Many systems have (hidden) regenerative structure. For example, Glynn (1989a) shows that many finite state space generalized semi-Markov processes (GSMPs) are regenerative. Essentially all discrete event simulation models can be described as GSMPs.

2. The estimators involved do not make any specific use of the regenerative structure, i.e., one need not identify regeneration points and group the data by regenerative cycles. To further amplify this point, Glynn (1989b) derives bias expansions for certain integrals of regenerative processes and then proposes a bias-reducing technique based on the form of these expansions. However, to employ that method in practice requires identifying the regeneration times and estimating expectations of random variables defined over the regenerative cycles. No such requirement is made here.

Section 4 contains further discussion of these points.

In order to present a complete and self-contained description of the relevant results, certain theorems are restated from other papers, specifically, Glynn (1987, 1989b) and Glynn and Heidelberger (1991b). The rest of the paper is organized as follows. In §2 we describe the formal mathematical framework, and in §3 we consider estimators based on a fixed amount of simulated time, both with and without deletion. Section 4 treats estimators based on computer time but without deletion, while §5 treats estimators based on computer time with deletion. The proofs are contained in §6. Finally, the results are summarized in §7.

The focus of this paper is theoretical. However, in Glynn and Heidelberger (1992) we describe experiments with these estimators on simple queueing network models of computer systems. The experimental results reported in that paper confirm the theoretical results developed in this paper.

2. The framework. Suppose that our goal is to estimate the steady-state mean of $X = (X(t) : t \geq 0)$, where X is a real-valued stochastic process. We let $C = (C(t) : t \geq 0)$ be the associated *cumulative computer time* process, so that $C(t)$ represents the

random amount of computer time required to generate X over the interval $[0, t]$. Let $\Rightarrow$ denote weak convergence, or convergence in distribution (see Billingsley (1968)). We assume that (X, C) satisfies the following set of hypotheses:

(2.1)

(i) C is a nondecreasing process.

(ii) There exist (deterministic) constants α, λ^{-1} $(0 < \lambda^{-1} < \infty)$, and a 2×2 matrix G such that

$$Z_\varepsilon(\cdot) \Rightarrow GB(\cdot)$$

as $\varepsilon \downarrow 0$ in $D[0, \infty)$ (the Skorohod space of functions $x : [0, \infty) \to \mathbb{R}^2$ that are right continuous and have left limits), where B is a two-dimensional standard Brownian motion and

$$Z_\varepsilon(t) = \varepsilon^{-1}\left(\varepsilon^2 \int_0^{t/\varepsilon^2} X(s)ds - \alpha t,\ \varepsilon^2 C(t/\varepsilon^2) - \lambda^{-1} t\right).$$

Assumption (2.1) (i) is reasonable in view of the interpretation of $C(t)$. As for (2.1) (ii), we note that one consequence of the assumption is that X and C satisfy a joint central limit theorem (CLT):

$$T^{1/2}\left(\frac{1}{T}\int_0^T X(s)ds - \alpha,\ \frac{C(T)}{T} - \lambda^{-1}\right) \Rightarrow N(0, H) \tag{2.2}$$

as $T \to \infty$, where $H = GG^t$. (In this case note that $GN(0, I)$ has the same distribution as $N(0, H)$ where I is the identity matrix. In one dimension, $\sigma N(0, 1)$ has the same distribution as $N(0, \sigma^2)$.) Thus, assumption (2.1) (ii) is best viewed as a strengthened version of the ordinary CLT. Because of the fact that (2.1) (ii) deals with weak convergence of stochastic processes (as opposed to ordinary random variables), it is typically termed a functional central limit theorem (FCLT) hypothesis. See Billingsley (1968) for further discussion of FCLTs.

It turns out that a great variety of stochastic processes exhibit FCLT behavior. For example, suppose that $C(\cdot)$ has a positive derivative so that it can be represented as

$$C(t) = \int_0^t \chi(s)ds, \quad \text{where} \quad \chi(t) > 0 \quad \text{a.s. for} \quad t \geq 0. \tag{2.3}$$

Under (2.3), $\chi(s)$ is the rate at which computer time is consumed at simulated time s. If the process $((X(t), \chi(t)) : t \geq 0)$ is regenerative and satisfies certain moment conditions, then (2.1) (ii) is known to be valid (see Glynn and Whitt (1987)). Regenerative structure is present in many of the stochastic models that are commonly simulated; see, for example, Glynn (1982). In addition, (2.1) (ii) holds if $((X(t), \chi(t)) : t \geq 0)$ is a martingale process or mixing process satisfying certain regularity hypotheses (see Ethier and Kurtz (1986)) or if it is an associated sequence (see Newman and Wright (1981)). Also, (2.1) (ii) is known to be valid for a large class of Markov processes (see Maigret (1978) and Nummelin (1984)). Because of the broad validity of (2.1) (ii), we view this condition as a mild regularity hypothesis that is satisfied by virtually all "real world" simulations.

One consequence of the joint CLT (2.2) is that

$$\frac{1}{T}\int_0^T X(s)ds \Rightarrow \alpha, \tag{2.4}$$

$$\frac{1}{T}C(T) \Rightarrow \lambda^{-1}, \tag{2.5}$$

as $T \to \infty$. The law of large numbers (2.4) states that the process X "settles down," on average, to the constant α; the parameter α is known as the *steady-state mean* of X. The goal of the steady-state simulation algorithms to be described in this paper is the efficient estimation of α. Also, (2.5) states that λ^{-1} may be interpreted as the long-run rate at which computer time is expended per unit of simulated time. Equivalently, λ is the long-run rate at which simulated time is generated per unit of computer time.

Two different (reasonable) strategies for estimating α can be employed by the simulation analyst. The first possibility is to generate a fixed amount of simulated time on each of the p processors, and to obtain an estimator for α by averaging the resulting observations over each of the p processors. This class of estimators, together with related initial transient deletion strategies, is discussed in §3. The second approach is to fix the amount of computer time available on each of the p processors, and to obtain an estimator for α by averaging over the random amount of simulated time generated on each of the p processors within the computational budget constraint. For $c \geq 0$, let $T_i(c) = \sup\{t \geq 0 : C_i(t) \leq c\}$ be the inverse process to $C_i(\cdot)$. The random variable (r.v.) $T_i(c)$ can be interpreted as the amount of simulated time that is generated on processor i in the first c units of computer time; we refer to $T_i(c)$ as the *cumulative simulated time* process associated with processor i. Hence, the second estimator involves averaging the process X_i over the random interval $[0, T_i(c)]$ and all p processors. Section 4 is devoted to studying this class of estimators when no initial transient deletion is applied, whereas §5 considers these estimators when initial transient deletion is applied.

3. Steady-state estimation using simulated time. Suppose that the process X is simulated up to (deterministic) time t on each of the p available processors. If the first $\beta(t)$ simulated time units are deleted from the initial segment of each copy X_i, we obtain the estimator

$$\alpha_s(p,t) = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{t-\beta(t)} \int_{\beta(t)}^{t} X_i(s)ds$$

(the subscript s stands for simulated time).

In our limit theorems, we shall be interested in determining how large the time horizon t needs to be for the parallel estimation algorithm to work efficiently. We shall therefore view t as a deterministic function $t = t(p)$ of the number of processors. Then $\beta(t) = \beta(t(p))$ is also a deterministic function of p. The limit theorems will include growth conditions on $t(p)$ and $\beta(t(p))$. To simplify the notation, we shall write $\alpha_s(p)$ and $\beta_s(p)$ as shorthand for $\alpha_s(p, t(p))$ and $\beta(t(p))$, respectively.

A consequence of (2.2) (and hence (2.1)) is that

$$\begin{aligned} T\left(\frac{1}{T}\int_0^T X(s)ds - \alpha\right)^2 &\Rightarrow \sigma_1^2 N(0,1)^2, \\ T\left(\frac{C(T)}{T} - \lambda^{-1}\right)^2 &\Rightarrow \sigma_2^2 N(0,1)^2, \end{aligned} \tag{3.1}$$

as $T \to \infty$, where $\sigma_i^2 = H_{ii}$ $(i = 1, 2)$. In order to carry out certain arguments, we will need to assume that the expectation operator can be passed through (3.1):

$$\begin{aligned} TE\left(\frac{1}{T}\int_0^T X(s)ds - \alpha\right)^2 &\to \sigma_1^2, \\ TE\left(\frac{C(T)}{T} - \lambda^{-1}\right)^2 &\to \sigma_2^2, \end{aligned} \tag{3.2}$$

as $T \to \infty$. A variety of stochastic processes obey (3.2), including regenerative processes (Smith (1955)), mixing and martingale process sequences (Ethier and Kurtz (1986)), and associated sequences (Newman and Wright (1981)) under appropriate moment conditions. Condition (3.2) is equivalent to asserting that $\{T(T^{-1}\int_0^T X(s)ds - \alpha)^2 : T > t_0\}$ and $\{T(C(T)/T - \lambda^{-1})^2 : T > t_0\}$ are uniformly integrable for some (finite) t_0 (see Chung (1974, p. 97)).

We will also need an assumption that controls the extent to which the initial transient biases the observations of the simulation. To be precise, let $b(t) = EX(t) - \alpha$. We assume that

$$\int_0^\infty |b(s)|ds < \infty. \tag{3.3}$$

Assumption (3.2) holds whenever $EX(t) \to \alpha$ exponentially fast (i.e., there exist constants $A, \lambda > 0$ such that $|b(t)| \leq Ae^{-\lambda t}$). This exponential rate of convergence is typical of most "real world" simulations. For example, finite-state irreducible aperiodic discrete-time Markov chains and finite-state irreducible continuous-time Markov chains both exhibit exponential convergence to their steady-state values (see Karlin and Taylor (1975)). Also, Nummelin and Tuominen (1982) prove exponential convergence rate results in a general state space Markov chain setting.

Our first theorem considers the case in which no initial bias deletion is performed, so that $\beta(t) \equiv 0$. Let $b = \int_0^\infty b(s)ds$.

THEOREM 1. *Assume* (2.1), (3.2), *and* (3.3) *are in force and that* $\beta(t) \equiv 0$. *Then,*

(i) *if* $p/t(p) \to \infty$, $t(p) \to \infty$, *and* $b \neq 0$, *then* $\sqrt{pt(p)}\ |\alpha_s(p) - \alpha| \Rightarrow \infty$ *as* $p \to \infty$,

(ii) *if* $p/t(p) \to m$ $(0 < m < \infty)$, *then* $\sqrt{pt(p)}\ (\alpha_s(p) - \alpha) \Rightarrow \sigma_1 N(0,1) + b\sqrt{m}$ *as* $p \to \infty$,

(iii) *if* $p/t(p) \to 0$, *then* $\sqrt{pt(p)}\ (\alpha_s(p) - \alpha) \Rightarrow \sigma_1 N(0,1)$ *as* $p \to \infty$.

The proof of this result appears in Glynn (1987). Note that if no truncation is used, $\alpha_s(1,t) = t^{-1}\int_0^t X_1(s)ds$. Then, (2.2) asserts that

$$\alpha_s(1,t) \overset{\mathcal{D}}{\approx} \frac{\sigma_1}{\sqrt{t}} N(0,1) \tag{3.4}$$

for large t ($\overset{\mathcal{D}}{\approx}$ denotes "approximately equal in distribution"). On the other hand, Theorem 1 (iii) states that

$$\alpha_s(p, t(p)) \overset{\mathcal{D}}{\approx} \frac{\sigma_1}{\sqrt{pt(p)}} N(0,1) \tag{3.5}$$

under the conditions stated there. Comparing (3.5) to (3.4), we see that (3.5) implies a p-fold speedup in the algorithm over that achieved with a single processor, i.e.,

$$\alpha_s(1, pt(p)) \overset{\mathcal{D}}{\approx} \alpha_s(p, t(p)).$$

Of course, a p-fold speedup is the best possible rate increase that we can expect in a parallel processing environment. Hence, Theorem 1 can be interpreted as stating that the time horizon $t = t(p)$ to be simulated on each of the p processors should satisfy $t >> p$, in order that the parallel algorithm achieve optimal efficiency.

Our next theorem considers the situation in which initial transient deletion is implemented. We note that $\beta_s(p)/t(p)$ is the fraction of the total simulated time that is deleted. The strategies that are considered here delete an asymptotically negligible fraction of the total observation set that is simulated.

THEOREM 2. *Assume* (2.1) *and* (3.2) *are in force and that* $b(t) \to 0$ *exponentially fast. Suppose* $\beta_s(p)/t(p) \to 0$ *as* $p \to \infty$. *If*

(i) $p/t(p) \to \infty$ *and* $\beta_s(p)/\log p \to \infty$, *or*

(ii) $p/t(p) \to m$ $(0 < m < \infty)$ *and* $\beta_s(p) \to \infty$, *or*

(iii) $p/t(p) \to 0$

as $p \to \infty$, *then*

$$\sqrt{pt(p)}(\alpha_s(p) - \alpha) \Rightarrow \sigma_1 N(0,1)$$

as $p \to \infty$.

For a proof, see Glynn and Heidelberger (1991b). Theorem 2 shows that with a modest amount of initial transient deletion, one can reduce the length of the time horizon significantly without affecting the p-fold speedup factor. In particular, if $t(p) = p^r$ $(r > 0)$, the p-fold speedup is retained so long as $\beta_s(p) = p^\varepsilon$ $(0 < \varepsilon < r)$. Thus, when initial transient deletion is suitably implemented, p-fold increases in efficiency ensue from time horizons that grow essentially arbitrarily slowly in p. A result similar to Theorem 2 can also be derived when the bias function $b(\cdot)$ decays polynomially fast (i.e., there exists $A, r > 0$, such that $|b(t)| \leq At^{-r}$ for $t \geq 0$); see Glynn and Heidelberger (1991b) for further details.

In our next theorem, we consider the case in which the fraction of the total simulation time that is deleted is fixed, so that the fraction deleted is no longer negligible. As might be expected, this rule, although reasonable from an implementation viewpoint, has a cost in terms of (greater) asymptotic variability.

THEOREM 3. *Assume* (2.1) *and* (3.2) *are in force and that* $b(t) \to 0$ *exponentially fast. If* $\beta_s(p) = \beta t(p)$ $(0 < \beta < 1)$ *and* $t(p) = p^r$ $(r > 0)$, *then*

$$\sqrt{pt(p)}(\alpha_s(p) - \alpha) \Rightarrow (1-\beta)^{-(1/2)}\sigma_1 N(0,1)$$

as $p \to \infty$.

For a proof, see Glynn and Heidelberger (1991b). Theorem 3, like Theorem 2, asserts that the time horizon over which we achieve a p-fold increase in efficiency is broadened considerably by using initial bias deletion. However, because of the fact that one deletes a fixed fraction of the total observation set, one pays a cost in the sense that the asymptotic variance is inflated by a factor of $(1-\beta)^{-1}$. On the other hand, for $\beta = 0.1$, the increase in the variance is only about 11 percent. Thus, the statistical cost incurred in using such a procedure is quite modest.

We conclude this section with a discussion of the computational cost associated with using estimators based on simulated time. In particular, assuming that the machine is (temporarily) dedicated to the estimation of α, the question of the algorithm's completion time is of significant importance. Recalling that $C_i(t)$ is the time at which

the ith processor completes the simulation of X over $[0,t]$, the completion time for a simulated time horizon of t is given by

$$C(p,t) = \max_{1 \leq i \leq p} C_i(t).$$

An additional relevant performance characteristic is the total idle time cumulated over all p processors, assuming that each processor must remain idle until all p processors have completed their assigned tasks. The idle time quantity is defined as

$$I(p,t) = \sum_{i=1}^{p} [C(p,t) - C_i(t)].$$

Set $C_s(p) = C(p, t(p))$ and $I_s(p) = I(p, t(p))$. Our final theorem of this section examines the behavior of $C_s(p)$ and $I_s(p)$ when the number of processors p is large. We will need one further assumption. Note that one consequence of (2.2) is that if $\sigma_2^2 > 0$, then

$$F_t(x) \to \Phi(x) \tag{3.6}$$

as $t \to \infty$, where $F_t(x) = P\{t^{-(1/2)}(C(t) - \lambda^{-1}t)/\sigma_2 \leq x\}$ and $\Phi(x) = P\{N(0,1) \leq x\}$. We wish to strengthen (3.6) to

$$\sup_{-\infty < x < \infty} |F_t(x) - \Phi(x)| = 0(t^{-(1/2)}) \tag{3.7}$$

as $t \to \infty$. This condition is known, in the probability literature, as a Berry–Esséen condition. The convergence result (3.7) is usually valid for stochastic processes satisfying (2.1). For example, regenerative processes (Bolthausen (1980)) and general state space Markov chains (Bolthausen (1982)) are known to satisfy (3.7) under suitable regularity hypotheses.

THEOREM 4. *Assume that* (2.1) *is in force with* $\sigma_2^2 > 0$ *and that* (3.2) *and* (3.7) *hold. If* $t(p)/p^2 \to \infty$, *then*

(i)
$$\frac{C_s(p) - \lambda^{-1}t(p)}{\sigma_2\sqrt{2t(p)\log p}} \Rightarrow 1,$$

(ii)
$$\frac{I_s(p)}{\sigma_2 p\sqrt{2t(p)\log p}} \Rightarrow 1$$

as $p \to \infty$.

According to Theorem 4, if $t(p) >> p$, then we may approximate $C_s(p)$ and $I_s(p)$ as

$$\begin{aligned} C_s(p) &\overset{\mathcal{D}}{\approx} \lambda^{-1}t(p) + \sigma_2\sqrt{2t(p)\log p}, \\ I_s(p) &\overset{\mathcal{D}}{\approx} \sigma_2 p\sqrt{2t(p)\log p}. \end{aligned} \tag{3.8}$$

Since λ^{-1} and σ_2^2 are typically unknown, the magnitude of the completion time $C_s(p)$ is, to some extent, a priori unpredictable. This is clearly undesirable. Furthermore, (3.8) shows that the total idle time can potentially be quite large. As a consequence, the machine may be significantly underutilized with estimators of the

type considered in this section (assuming that processors are not freed until time $C_s(p)$). For these reasons, the remainder of this paper explores estimators in which the completion time is fixed.

4. Steady-state estimation using computer time: No initial transient deletion. In this section, we suppose that each processor simulates X for a fixed (deterministic) amount of computer time c. At the completion time c, processor i will have generated X_i up to time $T_i(c)$. Given that no initial transient deletion is employed, two different estimators immediately suggest themselves:

$$\alpha_1(p,c) = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{T_i(c)} \int_0^{T_i(c)} X_i(s)ds,$$

$$\alpha_2(p,c) = \frac{\sum_{i=1}^{p} \int_0^{T_i(c)} X_i(s)ds}{\sum_{i=1}^{p} T_i(c)}.$$

Note that $\alpha_2(p,c)$ bears a strong resemblance to the usual ratio estimator for steady-state parameters in regenerative simulations (identify $T_i(c)$ as the length of the ith regenerative cycle).

As in §3, we shall be interested in limit theorems that describe how large the (computer) time horizon c should be, relative to the number of processors, in order that a p-fold speedup ensues. We shall therefore take c as a deterministic function $c = c(p)$ of the number of processors p. The limit theorems will then provide appropriate growth conditions on $c(p)$. As a shorthand notation, we will write $\alpha_i(p) = \alpha_i(p, c(p))(i = 1, 2)$.

Our first task is to understand the behavior of the sample means obtained from the individual processors.

THEOREM 5. *Assume that* (2.1) *and* (2.3) *hold. Then,*

(i) $\tilde{Z}_\varepsilon(\cdot) \Rightarrow \tilde{G}B(\cdot)$ *as* $\varepsilon \downarrow 0$ *in* $D[0,\infty)$, *where*

$$\tilde{Z}_\varepsilon(c) = \varepsilon^{-1}\left(\varepsilon^2 \int_0^{T(c/\varepsilon^2)} X(s)ds - \alpha\varepsilon^2 T(c/\varepsilon^2),\ \varepsilon^2 T(c/\varepsilon^2) - \lambda c\right),$$

$$\tilde{G}\tilde{G}^t = \tilde{H} \quad \text{and} \quad \tilde{H}_{11} = \lambda H_{11}, \quad \tilde{H}_{12} = \tilde{H}_{21} = -\lambda^2 H_{12}, \quad \tilde{H}_{22} = \lambda^3 H_{22}.$$

(ii)

$$c^{1/2}\left(\frac{\int_0^{T(c)} X(s)ds}{T(c)} - \alpha,\ \frac{T(c)}{c} - \lambda\right) \Rightarrow N(0, \underset{\sim}{H})$$

as $c \to \infty$, *where* $\underset{\sim}{H}_{11} = \lambda^{-1}H_{11}$, $\underset{\sim}{H}_{21} = -\lambda H_{12}$, *and* $\underset{\sim}{H}_{22} = \lambda^3 H_{22}$.

As in §3, we shall require that the cumulative processes defined on the time scale of computer time be appropriately uniformly integrable. Specifically, we shall need to assume that

$$\begin{aligned} cE\left(\frac{1}{T(c)}\int_0^{T(c)} X(s)ds - \alpha\right)^2 &\to \lambda^{-1}\sigma_1^2, \\ c^{-1}E\left(\int_0^{T(c)} [X(s) - \alpha]ds\right)^2 &\to \lambda\sigma_1^2, \\ cE\left(\frac{T(c)}{c} - \lambda\right)^2 &\to \lambda^3\sigma_2^2, \end{aligned} \tag{4.1}$$

as $c \to \infty$ (recall that $H_{ii} = \sigma_i^2$).

Theorem 5 asserts that the central limit behavior of X is preserved, in a qualitative sense, when the time scale is changed from that of simulated time to computer time. Since initial transient bias plays a critical role in the study of the estimators considered in this paper, it is incumbent upon us to also consider the extent to which the bias characteristics of X are altered by a change in the time scale.

We start by noting that if $E\int_0^{T(c)} |X(s)|ds < \infty$ and if (2.3) holds where $\chi(t) > 0$ almost surely for $t \geq 0$, a change-of-variables formula can be applied to obtain

$$\int_0^{T(c)} X(s)ds = \int_0^c Y(s)ds \quad \text{a.s.},$$

where $Y(s) = X(T(s))/\chi(T(s))$. The process $Y = (Y(t) : t \geq 0)$ is, in some sense, the original process X with its time scale transformed from simulated time to computer time. It turns out that the transformation that sends X into Y preserves any regenerative structure that may be present on the time scale of simulated time.

PROPOSITION 1. *Assume that* (2.3) *holds. If* $((X(t), \chi(t)) : t \geq 0)$ *is regenerative with respect to the random times* $(\tau_n : n \geq 1)$, *then* $(Y(t) : t \geq 0)$ *is regenerative with respect to the sequence* $(\eta_n : n \geq 1)$, *where* $\eta_n = C(\tau_n)$.

As indicated earlier in this paper, regenerative structure is to be found in many of the stochastic processes X that are simulated in practice. Given that X is regenerative, it is reasonable to further assume that the pair (X, χ) is also regenerative. Thus, we view regenerative hypotheses on the pair (X, χ) as a fairly mild restriction on the class of processes to be analyzed here. It is worth noting that while the proof techniques that appear in the remainder of this paper demand, to some extent, regenerative structure, our estimation strategies will be completely independent of the nature of the regenerations. As a consequence, our estimators will not require explicit identification of the regeneration points during the course of the simulation. The regenerative hypotheses will appear solely as mathematical regularity conditions and not as a vital component of the methodology itself.

The following hypotheses help to simplify certain proofs; they are not necessary to the development, however, and can be relaxed significantly, at the cost of greater mathematical complexity. Because we do not require that the simulation be started in the regeneration state, we need to make a distinction between the first (atypical) regeneration time τ_1 and the subsequent regeneration times. Let $\tilde{\eta} = \eta_2 - \eta_1$ and $\tilde{\tau} = \tau_2 - \tau_1$.

(4.2) There exists a constant $0 < x_1 < \infty$ such that $\chi(s) \geq x_1$ a.s.,

$$\begin{aligned} &E[\eta_2] < \infty, \quad E[\tau_2] < \infty, \quad E[\eta_2\tau_2] < \infty, \\ &E\left[\int_0^{\tau_2} |\chi(s)|ds\right] < \infty \quad \text{and} \quad E\left[\eta_2 \int_0^{\tau_2} |X(s)|ds\right] < \infty, \\ &\text{Both } \eta_1 \text{ and } \tilde{\eta} \text{ have probability density functions.} \end{aligned} \tag{4.3}$$

Assumption (4.2) merely states that there is a minimum rate at which computer time is consumed.

We are now ready to describe the behavior of the bias of the estimators $\alpha_1(p)$ and $\alpha_2(p)$.

THEOREM 6. *Assume that* (2.1), (2.3), (4.1), (4.2), *and* (4.3) *hold. If, in addition,* $c(p) \to \infty$ *as* $p \to \infty$, *then*

$$E\alpha_i(p) = \alpha + \frac{b_i}{c(p)} + o\left(\frac{1}{c(p)}\right)$$

as $p \to \infty$ $(i = 1, 2)$, *where* $b_1 = a + H_{12}$, $b_2 = a$, *and*

$$a = E\left[\int_0^{\tau_1} (X(s) - \alpha)ds\right]/\lambda - E\left[\int_0^{\tilde{\tau}} \int_0^s \chi(\tau_1 + u)du(X(\tau_1 + s) - \alpha)ds\right]/E\tilde{\tau}.$$

Roughly speaking, the bias in $\alpha_1(p)$ is due to two factors. First, the estimator $\alpha_1(p)$ is a ratio estimator, i.e., it is expressible as the ratio of two r.v.'s. The nonlinearity inherent in ratio estimators gives rise to the term $H_{12}/c(p)$. In addition, the expectation of the centered numerator r.v. $\int_0^{T(c)} [X(s) - \alpha]ds$ is nonzero, and this gives rise to the additional bias term $a/c(p)$. As for the bias of $\alpha_2(p)$, the ratio estimator bias is reduced by a factor of p because of the p-fold increase in the sample size of the numerator and denominator r.v.'s that appear in $\alpha_2(p)$.

Suppose that $\chi(s) = \lambda^{-1}$ almost surely. In this case, computer time is proportional to simulated time, i.e., $C(T) = T/\lambda$. Thus, $C(T)/T$ is deterministic, in which case $\sigma_2^2 = 0$ and $H_{12} = 0$. Therefore, $b_1 = b_2$ and the additional bias due to the randomness in $C(T)$ disappears.

The following theorem provides the analogue of Theorem 1 in the current setting, in which estimation occurs on the time scale of computer time rather than simulated time.

THEOREM 7. *Under the same hypotheses as in Theorem* 6, *the following results hold:*

(i) *If* $p/c(p) \to \infty$, $c(p) \to \infty$, *and* $b_i \neq 0$, *then* $\sqrt{pc(p)}|\alpha_i(p) - \alpha| \Rightarrow \infty$ *as* $p \to \infty$ $(i = 1, 2)$.

(ii) *If* $p/c(p) \to m$ $(0 < m < \infty)$, *then* $\sqrt{pc(p)}(\alpha_i(p) - \alpha) \Rightarrow \lambda^{-(1/2)}\sigma N(0, 1) + b_i m^{(1/2)}$ *as* $p \to \infty$ $(i = 1, 2)$.

(iii) *If* $p/c(p) \to 0$ *as* $p \to \infty$, *then* $\sqrt{pc(p)}(\alpha_i(p) - \alpha) \Rightarrow \lambda^{-(1/2)}\sigma N(0, 1)$ *as* $p \to \infty$ $(i = 1, 2)$.

According to Theorem 7, we typically need to choose $c(p) >> p$ when no initial bias deletion is used in order to achieve the desired p-fold increase in efficiency.

5. Steady-state estimation using computer time: Initial transient deletion. This section is devoted to analyzing modified versions of the estimators introduced in §4. Specifically, we shall modify the two estimators so that an initial segment is deleted from the observations generated by each processor. To precisely define the modified estimators, we let $\kappa(c) \leq c$ be a deterministic deletion point specified on the time scale of computer time. In other words, all observations generated in the first $\kappa(c)$ units of computer time are discarded. Of course, this just amounts to throwing away the initial segment $(X_i(t) : 0 \leq t \leq \gamma_i(c))$, where $\gamma_i(c) = T_i(\kappa(c))$. The modified versions of the two estimators studied in §4 are then defined as

$$\alpha_3(p, c) = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{T_i(c) - \gamma_i(c)} \int_{\gamma_i(c)}^{T_i(c)} X_i(s)ds,$$

$$\alpha_4(p, c) = \frac{\sum_{i=1}^{p} \int_{\gamma_i(c)}^{T_i(c)} X_i(s)ds}{\sum_{i=1}^{p} [T_i(c) - \gamma_i(c)]}.$$

Once again, to simplify notation, we set $\alpha_i(p) = \alpha_i(p, c(p))$ $(i = 3, 4)$ and $\kappa_c(p) = \kappa(c(p))$. We note that the assumption $\kappa_c(p)/c(p) \to 0$ as $p \to \infty$ is just a statement that the fraction of computer time devoted to observations that will eventually be deleted tends to zero in the limit.

Just as Theorem 2 required exponentially decreasing bias (on the simulated time scale), we will need some additional hypotheses to generate exponentially decreasing bias on the computer time scale.

$$E \exp t\eta_2 < \infty \quad \text{for some } t > 0, \tag{5.1}$$

$$X \text{ is a bounded process, i.e., } \sup\{|X(t, w)| : t \geq 0,\ w\epsilon\Omega\} \triangleq \|X\| < \infty. \tag{5.2}$$

Note that (5.1) is true if $E \exp t_2\tau_2 < \infty$ for some $t_2 > 0$ and $\chi(s) \leq M < \infty$ almost surely for all $s \geq 0$ (since $\eta_2 \leq M\tau_2$ in this case). Thus if X is a bounded process, exponential tail behavior of η_2 is inherited from that of τ_2.

THEOREM 8. *Suppose that* $\kappa_c(p)/c(p) \to 0$ *as* $p \to \infty$. *If, in addition to the hypotheses of Theorem* 6, (5.1) *and* (5.2) *hold, then*

(i) *if* $p/c(p) \to \infty$, $c(p) \to \infty$, *and* $H_{12} \neq 0$, *then* $\sqrt{pc(p)}|\alpha_3(p) - \alpha| \Rightarrow \infty$ *as* $p \to \infty$;

(ii) *if* $p/c(p) \to m$ $(0 < m < \infty)$, *then* $\sqrt{pc(p)}(\alpha_3(p) - \alpha) \Rightarrow \lambda^{-(1/2)}\sigma N(0, 1) + H_{12}m^{(1/2)}$ *as* $p \to \infty$;

(iii) *if* $p/c(p) \to 0$, *then* $\sqrt{pc(p)}(\alpha_3(p) - \alpha) \Rightarrow \lambda^{-(1/2)}\sigma N(0, 1)$ *as* $p \to \infty$.

Theorem 8 shows that even in the presence of initial bias deletion, the estimator $\alpha_3(p)$ does not effectively achieve a p-fold increase in efficiency unless the computer time $c(p)$ assigned to each processor is large (i.e., $c(p) >> p$). Thus, the estimator $\alpha_3(p)$ has essentially the same behavior as $\alpha_1(p)$ (see Theorem 7). In other words, $\alpha_3(p)$ does not benefit from the initial bias deletion present in the estimator. The basic problem is that deleting the initial segment from each processor's observations does not deal with the bias introduced by the nonlinearity of the ratio estimator obtained from each processor. This is reflected in Theorem 8 through the fact that the bias term that appears in part (ii) depends only on H_{12} and not also on the constant a that appears in Theorem 6.

Our next theorem describes the behavior of $\alpha_4(p)$.

THEOREM 9. *Suppose that* $\kappa_c(p)/c(p) \to 0$ *as* $p \to \infty$ *and that the same hypotheses as in Theorem* 8 *hold. If*

(i) $p/c(p) \to \infty$ *and* $\kappa_c(p)/\log p \to \infty$, *or*

(ii) $p/c(p) \to m$ $(0 < m < \infty)$ *and* $\kappa_c(p) \to \infty$, *or*

(iii) $p/c(p) \to 0$,

as $p \to \infty$, *then*

$$\sqrt{pc(p)}(\alpha_4(p) - \alpha) \Rightarrow \lambda^{-(1/2)}\sigma N(0, 1)$$

as $p \to \infty$.

According to Theorem 9, initial bias deletion has a significant positive impact on the estimator $\alpha_4(p)$. With a modest amount of initial bias deletion from the observations associated with each processor, a p-fold speedup in efficiency can be obtained with computer time horizons that are significantly shorter than those associated with no initial transient deletion. Since the estimator $\alpha_4(p)$ incurs none of the completion time and idle time costs associated with the "simulated time" estimators of §3,

this result suggests that the estimator $\alpha_4(p)$ is preferable to all the other estimators considered in this paper.

Our final theorem describes the behavior of $\alpha_4(p)$ when $\kappa(c) = \kappa c$ for $0 < \kappa < 1$, $c \geq 0$, i.e., when a proportion κ of all the observations are deleted before forming the estimator $\alpha_4(p)$.

THEOREM 10. *Assume the same hypotheses as in Theorem* 6. *If* $\kappa_c(p) = \kappa c(p)$ $(0 < \kappa < 1)$ *and* $c(p) = p^r$ $(r < 0)$, *then*

$$\sqrt{pc(p)}(\alpha_4(p) - \alpha) \Rightarrow (1-\kappa)^{-1/2}\lambda^{-1/2}\sigma N(0,1)$$

as $p \to \infty$.

The proof of Theorem 10 is similar to that of Theorem 9 and is therefore omitted. As in Theorem 3, deleting a positive fraction of all the observations leads to an asymptotic increase in the variability of the estimator of $(1-\kappa)^{-1}$. Of course, as pointed out in §3, this increase is quite modest if we choose κ small (say $\kappa = 0.1$).

6. Proofs.

Proof of Theorem 4. Let

$$\bar{\Phi}(x) = 1 - \Phi(x) \quad \text{and} \quad W_i(p) = (C_i(t(p)) - \lambda^{-1}t(p))/t(p)^{(1/2)}\sigma_2.$$

Note that the independence of the $W_i(p)$'s yields

$$\begin{aligned} &P\left\{C_s(p) \leq x\sigma_2\sqrt{2t(p)\log p} + \lambda^{-1}t(p)\right\} \\ &= P\left\{W_i(p) \leq x\sqrt{2\log p},\ 1 \leq i \leq p\right\} \\ &= F_{t(p)}(x\sqrt{2\log p})^p \\ &= \left(1 - \bar{\Phi}(x\sqrt{2\log p}) + O(1/\sqrt{t(p)})\right)^p. \end{aligned}$$

The quantity $\bar{\Phi}(x\sqrt{2\log p})$ may be estimated by using Lemma 2 of Feller (1968, p. 179). Noting that $O(1/\sqrt{t(p)}) = o(1/p)$, it is then straightforward to show that

$$P\{C_s(p) \leq x\sigma_2\sqrt{2t(p)\log p} + \lambda^{-1}t(p)\} \to \begin{cases} 0, & x < 1, \\ 1, & x > 1, \end{cases}$$

proving part (i). For part (ii), note that it is sufficient to prove that

$$\sum_{i=1}^{p} W_i(p)/p\sqrt{\log p} \Rightarrow 0$$

as $p \to \infty$. Fix $\varepsilon > 0$ and use Markov's inequality to obtain

$$P\left\{\left|\frac{1}{p}\sum_{i=1}^{p} W_i(p)\right| > \varepsilon\sqrt{\log p}\right\} \leq \frac{E|W_1(p)|}{\varepsilon\sqrt{\log p}}.$$

Assumption (3.2) states that $\{W(p)^2 : p > p_0\}$ is uniformly integrable for some (finite) p_0, from which one may conclude that $E|W_1(p)|$ is bounded in p. Hence, the right-hand side converges to zero as $p \to \infty$, proving (ii).

Proof of Theorem 5. We first note that (2.1) implies that

$$C_\varepsilon \Rightarrow \lambda^{-1}e$$

as $\varepsilon \downarrow 0$, where $C_\varepsilon(t) = \varepsilon^2 C(t/\varepsilon^2)$ and $e(t) = t$. Since χ is positive, it is evident that $C_\varepsilon^{-1}(\cdot)$ is continuous and satisfies $C_\varepsilon^{-1} \circ C_\varepsilon = e$. We may then apply Theorem 3.3 of Whitt (1980) to conclude that $C_\varepsilon^{-1} \Rightarrow \lambda e$ as $\varepsilon \downarrow 0$. But $C_\varepsilon^{-1} = T_\varepsilon$, where $T_\varepsilon(c) = \varepsilon^2 T(c/\varepsilon^2)$. Since composition is continuous as a mapping on the space of continuous functions (see Billingsley (1968, §17)), we obtain $Z_\varepsilon \circ T_\varepsilon \Rightarrow GB(\lambda e)$ as $\varepsilon \downarrow 0$. As a consequence, the continuous mapping principle implies that $h(Z_\varepsilon \circ T_\varepsilon) \Rightarrow h(GB(\lambda e))$, where $h(x,y) = (x, \lambda y)$; this proves part (i).

For part (ii), we let $\hat{X}(s) = X(s) - \alpha$. We note that part (i) implies that

$$c^{-(1/2)}\left(\int_0^{T(c)} \hat{X}(s)ds,\ T(c) - \lambda c\right) \Rightarrow N(0, \tilde{H})$$

as $c \to \infty$. Therefore, the continuous mapping principle shows that

$$\begin{aligned} &\lambda^{-1}c^{-(1/2)}\int_0^{T(c)} \hat{X}(s)ds - c^{(1/2)}\int_0^{T(c)} \hat{X}(s)ds/T(c) \\ &= \left(\lambda - \frac{c}{T(c)}\right)c^{-(1/2)}\int_0^{T(c)} \hat{X}(s)ds \Rightarrow 0 \end{aligned}$$

as $c \to \infty$. The proof of part (ii) is complete, if we note that

$$c^{-(1/2)}\left(\lambda^{-1}\int_0^{T(c)} \hat{X}(s)ds,\ T(c) - \lambda c\right) \Rightarrow N(0, \underset{\sim}{H})$$

as $c \to \infty$, and apply a converging-together argument.

Proof of Proposition 1. We first note that $\sigma(Y(t) : t \leq \eta_n) \subseteq \sigma((X(t), \chi(t)) : t \leq \tau_n)$, so it suffices to show that

$$P\{Y(\eta_n + \cdot)\varepsilon A | X(t),\ \chi(t) : t \leq \tau_n\} = P\{Y(\eta_1 + \cdot)\varepsilon A\} \tag{6.1}$$

for arbitrary (measurable) sets A. Now, $Y(\eta_n + t) = V(T(\eta_n + t))$, where $V(t) = X(t)/\chi(t)$, and $T(\eta_n + t) = T(\eta_n) + \Delta T_n(t) = \tau_n + \Delta T_n(t)$, where $\Delta T_n(t) = T(\eta_n + t) - T(\eta_n)$. We now observe that $\Delta T_n(\cdot)$ is the inverse to the process $\Delta C_n(t) = C(\tau_n + t) - C(\tau_n)$, in the sense that $\Delta C_n \circ \Delta T_n = e$. As a consequence, ΔT_n can be represented as a function $g(\tilde{V}_n)$, where $\tilde{V}_n(t) = (V(\tau_n + t),\ \chi(\tau_n + t))$. Hence, $Y(\eta_n + t) = k(\tilde{V}_n(t))$, where $k(\tilde{x}) = x_1((g \circ \tilde{x})(t))$ and $\tilde{x} = (x_1, x_2)$. Thus, $Y(\eta_n + \cdot)$ is a functional of the "shifted path" $\tilde{V}_n$ and (7.1) is trivially satisfied.

Proof of Theorem 6. The proof for $\alpha_1(p)$ can be found in Glynn (1989b) (the assumptions 4.3 are heavily used there). For part (ii), we let $S_i(c) = \int_0^{T_i(c)}[X_i(s) - \alpha]ds$. Then,

$$\alpha_2(p) - \alpha = \frac{\sum_{i=1}^{p} S_i(c(p))}{\sum_{i=1}^{p} T_i(c(p))}.$$

We note that for $\varepsilon > 0$,

$$P\left\{\left|\sum_{i=1}^{p}\frac{T_i(c(p))}{pc(p)} - \lambda\right| > \varepsilon\right\} \leq \frac{1}{pc(p)\varepsilon^2}\ \text{var}\left[\frac{T_i(c(p)) - \lambda c(p)}{\sqrt{c(p)}}\right] = O\left(\frac{1}{pc(p)}\right) \to 0 \tag{6.2}$$

as $p \to \infty$, where the uniform integrability condition (4.1) was used to guarantee that the variance term was bounded in p. Hence, $\sum_{i=1}^{p} T_i(c(p))/pc(p) \Rightarrow \lambda$ as $p \to \infty$ (this result will be used in Theorem 7). Expand $pc(p)/\sum_{i=1}^{p} T_i(c(p))$ in a first-order Taylor expansion about λ^{-1}, to obtain

$$\frac{pc(p)}{\sum_{i=1}^{p} T_i(c(p))} = \lambda^{-1} - \frac{1}{\xi(p)^2}\left(\sum_{i=1}^{p} \frac{T_i(c(p))}{pc(p)} - \lambda\right), \tag{6.3}$$

where $\xi(p)$ lies between λ and $\sum_{i=1}^{p} T_i(c(p))/pc(p)$.

Since $\chi(s)$ is bounded from below, there exists a finite constant M such that $|1/\xi(p)^2| \le M$. By using the Taylor expansion and taking expectations we have

$$\begin{aligned} E[(\alpha_2(p) - \alpha)] &= E\left[\sum_{i=1}^{p} \frac{S_i(c(p))}{\lambda pc(p)}\right] \\ &\quad - E\left[\sum_{i=1}^{p} \frac{S_i(c(p))}{pc(p)} \sum_{i=1}^{p} \frac{(T_i(c(p)) - \lambda c(p))}{\xi(p)^2 pc(p)}\right]. \end{aligned} \tag{6.4}$$

To handle the first term on the right-hand side of (6.4), we use Glynn (1989b) to show that

$$\frac{1}{\lambda pc(p)} E\left[\sum_{i=1}^{p} S_i(c(p))\right] = \frac{a}{c(p)} + o\left(\frac{1}{c(p)}\right) \tag{6.5}$$

as $p \to \infty$.

To deal with the second term on the right-hand side of (6.4), we bound it by

$$2E\left[\sum_{i=1}^{p} \frac{S_i(c(p))}{pc(p)}\right]^2 + 2E\left[\sum_{i=1}^{p} \frac{(T_i(c(p)) - \lambda c(p))}{\xi(p)^2 pc(p)}\right]^2. \tag{6.6}$$

The second term in (6.6) is dominated by

$$\begin{aligned} &2M^2 E\left[\sum_{i=1}^{p} \frac{(T_i(c(p)) - \lambda c(p))}{pc(p)}\right]^2 \\ &= \frac{2M^2}{c(p)^2}(E[T_1(c(p) - \lambda c(p))])^2 + \frac{2M^2}{pc(p)^2}\mathrm{var}(T_1(c(p)) - \lambda c(p)) \\ &= o(1/c(p)) + O(1/pc(p)), \end{aligned} \tag{6.7}$$

using (4.1)'s uniform integrability. The first term in (6.5) can be handled similarly, yielding an estimate of $O(1/c(p))$. Combining (6.4) through (6.6) yields the result for $\alpha_2(p)$.

Proof of Theorem 7. We first consider $\alpha_1(p)$. We note that Theorem 6 implies that

$$\begin{aligned} &\sqrt{pc(p)}(\alpha_1(p) - \alpha) \\ &= p^{-(1/2)} \sum_{i=1}^{p} V_i(p) + \sqrt{pc(p)}(E\alpha_1(p) - \alpha) \\ &= p^{-(1/2)} \sum_{i=1}^{p} V_i(p) + b_1 \sqrt{\frac{p}{c(p)}} + \sqrt{p}o\left(\frac{1}{c(p)}\right), \end{aligned}$$

where

$$V_i(p) = \sqrt{c(p)} \frac{\int_0^{T_i(c(p))} X_i(s)ds}{T_i(c(p)) - E\alpha_1(p)}.$$

By condition (4.1), $\{V_i(p)^2 : p > p_0\}$ is uniformly integrable for some (finite) p_0. We can therefore apply the Lindeberg–Feller theorem (see Chung (1974, p. 205)) to conclude that

$$p^{-(1/2)} \sum_{i=1}^{p} V_i(p) \Rightarrow \lambda^{-(1/2)} \sigma N(0,1)$$

as $p \to \infty$; this proves all three implications for the estimator $\alpha_1(p)$.

The second estimator is handled similarly. We note that

$$\sqrt{pc(p)}(\alpha_2(p) - \alpha) = \frac{pc(p)}{\sum_{i=1}^{p} T_i(c(p))} \cdot \left[p^{-(1/2)} \sum_{i=1}^{p} \tilde{S}_i(p) - \sqrt{\frac{p}{c(p)}} E\tilde{S}_1(p)\right]$$

where $\tilde{S}_i(p) = c(p)^{-(1/2)}[S_i(c(p)) - ES_i(c(p))]$. From (7.2), it is evident that

$$pc(p)/\sum_{i=1}^{p} T_i(c(p)) \Rightarrow \lambda^{-1} \quad \text{as } p \to \infty.$$

A uniform integrability argument similar to that for $\alpha_1(p)$ then shows that

$$p^{-(1/2)} \sum_{i=1}^{p} \tilde{S}_i(p) \Rightarrow \lambda^{(1/2)} \sigma N(0,1).$$

To complete the proof, we refer to Glynn (1989b), where it is shown that $ES_i(c(p)) = \lambda b_2/c(p) + o(1/c(p))$.

Proof of Theorem 8. The argument largely mimics the proof of Theorem 7. The first step is to obtain an expression for the bias of $\alpha_3(p)$. We claim that the bias takes the form

$$E\alpha_c(p) = \alpha + \frac{H_{12}}{c(p)} + o\left(\frac{1}{c(p)}\right) \tag{6.8}$$

as $p \to \infty$. A careful study of Glynn (1989b) shows that the proof there needs to be modified in two respects. First, one needs to argue that

$$\left\{c^{-1}\left(\int_{\gamma(c)}^{T(c)} [X(s) - \alpha]ds\right)^2 : c > c_0\right\}$$

is uniformly integrable for some $c_0 < \infty$. This follows from the observation that

$$\int_{\gamma(c)}^{T(c)} [X(s) - \alpha]^2 ds \le 2\left(\int_0^{T(\chi(c))} [X(s) - \alpha]ds\right)^2 + 2\left(\int_0^{T(c)} [X(s) - \alpha]ds\right)^2;$$

each of the processes on the right-hand side is uniformly integrable as a result of (4.1) and hence the left-hand side is uniformly integrable (Chung (1974, p. 100)). Second, we note that

$$E\int_{\gamma(c)}^{T(c)} [X(s) - \alpha]ds = \int_{\chi(c)}^{c} E\hat{Y}(s)ds,$$

where $\hat{Y}(c) = [X(T(c)) - \alpha]/\chi(T(c))$. Since $\|X\| < \infty$ and χ is bounded away from zero, it follows that $\hat{Y}$ is a bounded process. Hence, the results of Nummelin and Tuominen (1982) apply, showing that $E\hat{Y}(t) \to 0$ exponentially fast. It is then evident that

$$\int_{\chi(c)}^{c} E\hat{Y}(s)ds = o(1)$$

as $c \to \infty$. As a consequence, the term that contributes the quantity $a/c(p)$ to the asymptotic bias of $\alpha_1(p)$ is $o(1/c(p))$ in the current setting. This yields (6.7).

The proof of Theorem 8 is completed in much the same way as Theorem 7. The uniform integrability established above allows us to apply the Lindeberg–Feller theorem once again, finishing the proof.

Proof of Theorem 9. The analogue of (6.8) for the estimator $\alpha_4(p)$ is

$$E\alpha_4(p) = \alpha + o\left(1/c(p)\right),$$

and is established in basically the same way. The result of the proof goes through as in Theorem 8.

7. Summary. This paper has investigated theoretical properties of an attractive method for using parallel processors in discrete event simulations: running independent replications, in parallel, on multiple processors and averaging the results at the end of the runs. In previous papers, we considered the problem of estimating transient quantities, while in this paper we consider the steady-state estimation procedure. While the method of replications with initial transient deletion is conceptually simple to apply, some care needs to be taken in order to obtain estimators with the proper convergence behavior. Specifically, the growth rates for the number of processors (replications), the length of the replications, and the length of the deletion period need to be controlled in order to produce valid confidence intervals for steady-state parameters. In the parallel processing setting, a sampling plan in which the replication lengths are given by limits on computer time is particularly attractive since the completion time of the experiment is deterministic (assuming the machine is dedicated to running the simulation experiment). However, in this case, the leading term in the bias expansion of the straightforward estimator without deletion, $\alpha_1(p)$, is $(a + H_{12})/c(p)$ where $c(p)$ is the computer time per replication, a is due to initialization bias, and H_{12} is due to the fact that the denominator in the ratio estimate is random. Deleting an appropriate portion of each replication removes the initialization bias $a/c(p)$, but does not remove the ratio bias $H_{12}/c(p)$. Thus, when this estimator is used and the replication length is determined by computer time, deletion is essentially useless. On the other hand, the bias expansion of a new estimator, $\alpha_2(p)$, has leading term $a/c(p)$, which is removed entirely by appropriate deletion. Therefore, in practice, we recommend use of the new estimator with deletion, $\alpha_4(p)$.

Experimental results concerning the performance of these estimators in simulations of simple queueing network models are reported in Glynn and Heidelberger (1992). Those experimental results confirm the theoretical results presented here and reinforce our recommendation to use $\alpha_4(p)$ rather than $\alpha_3(p)$. Our experiments showed that $\alpha_4(p)$ outperforms $\alpha_3(p)$, in terms of exhibiting less bias and truer confidence interval coverage, when the number of processors is large and the amount of time per processor is relatively small.

While we have described the results in terms of a computer time constraint on the replication lengths, they remain valid for essentially any other measure of replication length. Examples include computing charges (which may involve costs for CPU, memory, and I/O use), real time (i.e., wall-clock time, which may differ in multiprogrammed environments), the total number of events processed, and the total number of events of a certain type processed (such as departures from a network). In addition, the results are applicable to simulation experiments on a single processor if the replication lengths are determined in the above fashion.

REFERENCES

B. C. Bhavsar, and J. R. Isaac (1987), *Design and analysis of parallel Monte Carlo algorithms*, SIAM J. Sci. Statist. Comput., 8, pp. 73–95.

P. Billingsley (1968), *Convergence of Probability Measures*, John Wiley, New York.

E. Bolthausen (1980), *The Berry–Esseén theorem for functionals of discrete Markov chains*, Z. Wahrsch. verw. Gebiete, 54, pp. 59–73.

——— (1982), *The Berry–Esseén theorem for strongly mixing Harris recurrent Markov chains*, Z. Wahrsch. verw. Gebiete, 60, pp. 283–289.

K. L. Chung (1974), *A Course in Probability Theory*, Academic Press, New York.

M. A. Crane and D. L. Iglehart (1975), *Simulating stable stochastic systems*, III: *Regenerative processes and discrete event simulations*, Oper. Res., 23, pp. 33–45.

S. N. Ethier and T. G. Kurtz (1986), *Markov Processes: Characterization and Convergence*, John Wiley, New York.

W. Feller (1968), *An Introduction to Probability Theory and Its Applications*, John Wiley, New York.

R. M. Fujimoto (1989), *Time warp on a shared memory multiprocessor*, in Proc. 1989 Internat. Conf. Parallel Processing, Vol. III., F. Ris and P. M. Kogge, eds., The Pennsylvania State University Press, State College, PA, pp. 242–249.

——— (1990), *Parallel discrete event simulation*, Comm. ACM, 33, pp. 31–53.

P. W. Glynn (1982), *Regenerative aspects of the steady-state simulation problem for Markov chains*, Tech. Report 17, Department of Operations Research, Stanford University, Stanford, CA.

——— (1987), *Limit theorems for the method of replication*, Stochastic Models, 4, pp. 344–350.

——— (1989a), *A* GSMP *formalism for discrete event systems*, Proc. IEEE, 77, pp. 14–23.

——— (1989b), *A low bias steady-state estimator for equilibrium processes*, Tech. Report 47, Department of Operations Research, Stanford University, Stanford, CA.

P. W. Glynn and P. Heidelberger (1990), *Bias properties of budget constrained simulations*, Oper. Res., 38, pp. 801–814.

——— (1991a), *Analysis of parallel replicated simulations under a completion time constraint*, ACM Trans. Modeling and Computer Simulation, 1, pp. 3–23.

——— (1991b), *Analysis of initial transient deletion for replicated steady-state simulations*, Oper. Res. Lett., 10, pp. 437–443.

——— (1992), *Experiments with initial transient deletion for parallel, replicated steady-state simulations*, Management Sci., 38, pp. 400–418.

P. W. Glynn and W. Whitt (1987), *Sufficient conditions for functional-limit-theorem versions of* $L = \lambda W$, Queueing Systems, Theory, Appl., 1, pp. 279–287.

P. Goli, P. Heidelberger, D. Towsley, and Q. Yu (1990), *Processor assignment and synchronization in parallel simulation of multistage interconnection networks*, in Distributed Simulation, D. Nicol, ed., The Society for Computer Simulation International, San Diego, CA, pp. 181–187.

P. Heidelberger (1986), *Statistical analysis of parallel simulations*, in 1986 Winter Simulation Conference Proceedings, J. Wilson and J. Henriksen, eds., IEEE Press, Piscataway, NJ, pp. 290–295.

——— (1988), *Discrete event simulations and parallel processing: Statistical properties*, SIAM J. Sci. Statist. Comput., 9, pp. 1114–1132.

S. Karlin and H. M. Taylor (1975), *A First Course in Stochastic Processes*, Academic Press, New York.

B. D. Lubachevsky (1989), *Efficient distributed event-driven simulations of multiple-loop networks*, Comm. ACM, 32, pp. 111–123.

N. Maigret (1978), *Thé orè me de limite centrale fonctionnel pour une chaîne de Markov reċurrente au sens de Harris et positive*, Ann. Inst. Henri Poincaré, 14, pp. 425–440.

M. S. MEKETON AND P. HEIDELBERGER (1982), *A renewal theoretic approach to bias reduction in regenerative simulations*, Management Sci., 28, pp. 173–181.

C. M. NEWMAN AND A. L. WRIGHT (1981), *An invariance principle for certain dependent sequences*, Ann. Probab., 9, pp. 671–675.

D. M. NICOL (1988), *Parallel discrete-event simulation of* FCFS *stochastic queueing networks*, in Proc. ACM/SIGPLAN PPEALS 1988, Parallel Programming: Experience with Applications, Languages and Systems, ACM Press, New York, pp. 124–137.

E. NUMMELIN (1984), *General Irreducible Markov Chains and Non-negative Operators*, Cambridge University Press, Cambridge, U.K.

E. NUMMELIN AND P. TUOMINEN (1982), *Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory*, Stochastic Process. Appl., 12, pp. 187–202.

W. L. SMITH (1955), *Regenerative stochastic processes*, Proc. Roy. Soc. London Ser. A., 232, pp. 6–31.

B. UNGER AND D. JEFFERSON, EDS. (1988), *Distributed Simulation*, 1988, Simulation Series 19, No. 3, The Society for Computer Simulation International, San Diego, CA.

B. UNGER AND R. FUJIMOTO, EDS. (1989), *Distributed Simulation*, 1989, Simulation Series 21, No. 2, The Society for Computer Simulation International, San Diego, CA.

W. WHITT (1980), *Some useful functions for functional limit theorems*, Math. Oper. Res., 5, pp. 67–85.

Q. YU, D. TOWSLEY, AND P. HEIDELBERGER (1989), *Time-driven parallel simulation of multistage interconnection networks*, in Distributed Simulation, 1989, B. Unger and R. Fujimoto, eds., The Society for Computer Simulation International, San Diego, CA, pp. 191–196.

SIAM J. SCI. STAT. COMPUT.
Vol. 13, No. 4, pp. 923–947, July 1992

006

AN IMPLEMENTATION OF THE FAST MULTIPOLE METHOD WITHOUT MULTIPOLES*

CHRISTOPHER R. ANDERSON†

Abstract. An implementation is presented of the fast multipole method, which uses approximations based on Poisson's formula. Details for the implementation in both two and three dimensions are given. Also discussed is how the multigrid aspect of the fast multipole method can be exploited to yield efficient programming procedures. The issue of the selection of an appropriate refinement level for the method is addressed. Computational results are given that show the importance of good level selection. An efficient technique that can be used to determine an optimal level to choose for the method is presented.

Key words. Poisson equation, fast summation, point sources

AMS(MOS) subject classifications. 65C99, 35J05, 34B27

1. Introduction.

The purpose of this paper is three-fold. First we will present a method for computing N-body interactions that is similar to the fast multipole method (FMM) as developed by Greengard and Rokhlin [5], [6] and Van Dommelen and Rundensteiner [15], but one that does not use complex power series in two dimensions or spherical harmonic expansions in three dimensions. Our procedure will be based on the use of Poisson's formula for representing solutions of Laplace's equation. While the accuracy and operation count of the resulting method is almost identical to the fast multipole method, the method does offer some advantages. One advantage is that the component operations of the multipole method, such as shifting and combining multipoles, are very easy to formulate for approximations based on Poisson's formula. Another advantage is that the difference between the two- and three-dimensional methods is very slight, and so programming a three-dimensional method is relatively straightforward once a two-dimensional method has been programmed. The second aspect this paper discusses is how multigrid programming strategy can be used to facilitate the programming of our method and others like it (such as the original fast multipole method). Third, we wish to discuss the issue of parameter selection when using these "fast" methods. Essentially, the computational efficiency of these methods depends critically upon the choice of a level of refinement of physical space. A wrong choice can lead to a very inefficient algorithm. We shall present a procedure for obtaining an optimal choice of the refinement level.

The problem of calculating N-body interactions occurs in a wide variety of computational problems—discrete vortex calculations, galaxy simulations, plasma simulations, etc. For each of these computational problems the calculation takes on a slightly different form, but each shares the common feature that the interaction is determined via solutions of Laplace's equation. So, rather than address each different application, we will discuss the following N-body model problem: Given N charged particles at locations x_i with strengths κ_i the goal is to calculate the potential $\phi(x_i)$,

* Received by the editors July 23, 1990; accepted for publication (in revised form) May 16, 1991.
† Department of Mathematics, University of California, Los Angeles, California, 90024. This research was supported by Office of Naval Research contract N00014-86-K-0691, National Science Foundation grant DM586-57663, and IBM fellowship D880908.