# Analysis of Parallel Replicated Simulations Under a Completion Time Constraint

PETER W. GLYNN
Stanford University
and
PHILIP HEIDELBERGER
IBM Thomas J. Watson Research Center, Hawthorne

We analyze properties associated with a simple yet effective way to exploit parallel processors in discrete event simulations: averaging the results of multiple, independent replications that are run, in parallel, on multiple processors. We focus on estimating expectations from terminating simulations, or steady state parameters from regenerative simulations. We assume that there is a CPU time constraint, $t$, on each of $P$ processors. Unless the replication lengths are bounded, one must be willing to simulate beyond any fixed, finite time $t$ on at least some processors in order to always obtain a strongly consistent estimator (as the number of processors increases). We therefore consider simulation experiments in which $t$ is viewed as either being a strict constraint, or a guideline, in which case simulation beyond time $t$ is permitted. The statistical properties, including strong laws, central limit theorems, bias expansions, and completion time distributions of a variety of estimators obtainable from such an experiment are derived. We propose an unbiased estimator for a simple mean value. This estimator requires preselecting a fraction of the processors. Simulation beyond time $t$ may be required on a preselected processor, but only if no replications have yet been completed on that processor.

Categories and Subject Descriptors: G.1. **[Numerical Analysis]**: General—*parallel algorithms*; G.3 **[Probability and Statistics]**—*probabilistic algorithms (including Monte Carlo)*; I.6.1 **[Simulation and Modeling (G.3)]**: Simulation Theory—*types of simulation (discrete)*

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Bias, discrete event simulation, estimation, multiple replications, parallel processing, renewal theory

## 1. INTRODUCTION

Discrete event simulations often require large amounts of computer time in order to produce statistically accurate estimates. This is particularly true of queueing network models of manufacturing, communications, and computer

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.
Authors' addresses: Peter W. Glynn, Department of Operations Research, Stanford University, Stanford, CA 94305; Philip Heidelberger, IBM Thomas J. Watson Research Center, Hawthorne, P.O. Box 704, Yorktown Heights, NY 10598.
Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its data appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
© 1991 ACM 1049-3301/91/0300-0067 1.50

systems. Such simulations therefore represent an important potential application for parallel processors. Distributed simulation, or the execution of a single realization of a stochastic process on multiple cooperating processors, has recently been the focus of a good deal of research. Fujimoto [11] contains an excellent introduction to this topic, including a discussion of why distributed simulation is difficult, a description of a variety of synchronization techniques, and a literature review. For further surveys and a representative sample of research in this area see [28, 30, 32, 36], or [37]. While significant speedups have been achieved in distributed simulations of specific, specially structured queueing systems (see, e.g., [10, 18, 25, 29], or [39]), effective use of distributed simulation for the type of complex models that often arise in practice currently remains an area of research.

However, a simple alternative to distributed simulation easily takes advantage of parallel processing technology: running multiple independent replications of the model, in parallel, on multiple processors and averaging the results of at the end of the runs. The method can potentially be applied to any model and does not require advanced parallel processing hardware, for instance, it can be used on a collection of workstations attached to a local area network. Heidelberger [19] developed a simple model to compare the statistical efficiency (specifically the mean squared errors) of these two approaches for estimating so-called steady state quantities. This analysis shows, qualitatively, that the parallel replications approach is statistically more efficient than distributed simulation, provided:

(1) the model's memory requirements are small enough so that it can reasonably fit into the memory of a single processor;
(2) the model can be run long enough on a single processor so that initialization bias is not significant (compared to the standard deviation); and
(3) a main reason why the model must be run for long periods of time is the slow rate at which the standard deviation decreases.

We believe that these conditions are satisfied for many queueing models, for example, networks in moderate to heavy traffic with, say, up to hundreds of queues: such systems are difficult to simulate primarilly because the standard deviations of the point estimates are typically large (see, e.g., [38]). As technology advances and processors become faster and memories larger, we expect the class of models suitable for the parallel replications approach to become ever larger. Further statistical properties associated with this approach for steady state estimation are considered in [13, 14] and [17].

When one considers the estimation of quantities arising from so-called transient, or terminating, simulations, the parallel replications approach appears to be even more attractive. Examples of such quantities are the following:

(1) the expected time until a queue length first exceeds some level (given a prespecified set of initial conditions);
(2) the expected number of customers that can be served in a fixed time interval (again, given a prespecified set of initial conditions);

(3) the mean time to failure in a reliability model; and
(4) the expectation of an integral, or sum, over a cycle in a regenerative process (see [34]), for instance, the integral of a queue size. In this case a replication is associated with a regenerative cycle and the parallel replications approach for estimating transient quantities can be thought of as a parallel regenerative method (see, e.g., [7]) for estimating steady state quantities.

Intuitively, one should be able to just turn on the parallel processor for some period of time, say $t$, and average the resulting observations at the end of the run. For a large number of processors, one should be able to make $t$ small, thereby running only a few replications on each processor. Thus, highly accurate estimates should be obtained in a very short period of time.

However, there are some potentially serious *statistical* problems with the parallel replications approach, especially for a large number of processors. These problems arise mainly because of the sampling bias associated with the fixed completion time $t$. First of all, what should one do with the replications that are in progress at time $t$? Should they be discarded, or allowed to complete? Second, how should one average the resulting observations? There are several ways this can be done. Does it make a difference? What are the statistical properties of the resulting estimators? This paper studies these and related questions.

In the case of a single processor, these issues were investigated by Meketon and Heidelberger [26], who showed that under certain circumstances it is better to complete the replication in progress at time $t$. Specifically, if $t$ is measured in units of simulated time, then in the case of ratio estimation in regenerative simulation, the bias gets reduced from order $1/t$ to $1/t^2$ by completing the regenerative cycle (replication) in progress at time $t$. In the parallel processing setting, these (and other) issues were addressed by Heidelberger [20], who showed that some of the most obvious estimates obtainable from parallel replication schemes are guaranteed to produce incorrect results, in the sense that they converge to the wrong quantity with probability one as $P$, the number of processors, increases. In [20], other estimates with correct convergence properties were proposed and analyzed. Associated with these estimators is a completion time penalty that arises because all, or some, of the incomplete replications must be allowed to finish in order to reduce or remove the bias. A subsequent paper by Glynn and Heidelberger [16] revisited the single processor case, obtaining finer bias expansions for a variety of estimators and relating these expansions to the bias-reducing technique in [26]. Other issues related to parallel replication schemes have also been analyzed by Bhavsar and Isaac [1].

The present paper explores the parallel processing implications of the results in [16]. We generalize and improve upon the estimators suggested in [20]. The generalization permits more than just ratio estimation, and the improvements include new estimators with shorter completion times. In addition, whereas [20] considers asymptotic behavior as either $t \to \infty$ or $P \to \infty$, we analyze situations in which both $t$ and $P$ approach $\infty$ simultaneously. This allows us to determine, for example, the relative rates at which $t$

and $P$ must grow in order to obtain valid confidence intervals when one discards all replications in progress at time $t$. Since, in reality, we never actually have an infinite number of processors, these results should be interpreted as determining how large $t$ needs to be qualitatively for a given, large, number of processors.

The paper is organized as follows. In Section 2, we introduce notation and, in the interest of keeping the paper self-contained, review the most relevant results from [16] and [20]. Section 3 considers the estimation of a simple mean value, while Section 4 considers the estimation of a nonlinear function of a vector of simple means (e.g., a ratio of two mean values). Completion time results associated with the various estimators are derived in Section 5, and the results are summarized in Section 6.

## 2. NOTATION AND REVIEW

We let $P$ denote the number of processors. We assume that processors are identical and that, simultaneously, each processor runs multiple replications of the simulation. The output of replication $j$ on processor $i$ is a random variable (r.v.) $X_{ij}$. The goal of the simulation is to estimate $\mu = \mathrm{E}[X_{ij}]$. We let $\tau_{ij}$ denote the (random) amount of (computer or real) time required to run replication $j$ on processor $i$ and let $S_i(n) = \tau_{i1} + \cdots + \tau_{in}$ be the time that it takes processor $i$ to complete $n$ replications (also let $S_i(0) = 0$). In a simulation of length $t$, processor $i$ can complete $N_i(t)$ replications where $N_i(t) = \sup\{n \geq 0: S_i(n) \leq t\}$. Throughout the paper, we make the reasonable assumption that $\{(X_{ij}, \tau_{ij}),\ i = 1, \ldots, P,\ j \geq 1\}$ are i.i.d. (independent and identically distributed) r.v.s. Under this assumption $\{N_i(t),\ t \geq 0\}$, $i = 1, \ldots, P$ are i.i.d. renewal processes (see [5] or [35]).

Notice that there are many possible ways to estimate $\mu$ from such an experiment. One could estimate $\mu$ based on simulating a fixed total number of replications, or based on a completion time constraint $t$ as in the above setting. While other stopping rules are also possible, we only consider estimators based on a completion time constraint, which represents a realistic and practical method for running such parallel replication schemes. In [20], a variety of such estimators for $\mu$ are considered and analyzed. The first, and perhaps the most obvious, thing to do is simply average all of the observations that have completed by time $t$. This results in the following estimate:

$$\mu_1(P, t) \equiv \frac{\sum_{i=1}^{P} \sum_{j=1}^{N_i(t)} X_{ij}}{\sum_{i=1}^{P} N_i(t)}. \tag{2.1}$$

In [20], it is shown that while $\lim_{t\to\infty} \hat{\mu}_1(P, t) = \mu$ almost surely (a.s.), $\lim_{P\to\infty} \hat{\mu}_1(P, t)$ is typically not equal to $\mu$, in fact $\lim_{P\to\infty} \hat{\mu}_1(P, t) = \mu + O(1/t)$ a.s. In other words, if one attempts to estimate $\mu$ by running a very large number of processors for a short amount of time, then the estimate need not converge to $\mu$. On the other hand, if one completes all the replications in

progress at time $t$, then the estimate

$$\hat{\mu}_2(P,t) \equiv \frac{\sum_{i=1}^{P} \sum_{j=1}^{N_i(t)+1} X_{ij}}{\sum_{i=1}^{P} N_i(t) + 1} \tag{2.2}$$

has the property that $\lim_{t\to\infty} \hat{\mu}_2(P,t) = \lim_{P\to\infty} \hat{\mu}_2(P,t) = \mu$ a.s. The difference in asymptotic behavior (as $P \to \infty$) between $\hat{\mu}_1(P,t)$ and $\hat{\mu}_2(P,t)$ is due to the ratio form of the estimates and the fact that $(N_i(t) + 1)$ is a stopping time and, therefore, by Wald's equation (see p. 186 of [21]), the limit of $1/P$ times the numerator of Eq. 2.2 converges to $\mathrm{E}[N_i(t) + 1]\mathrm{E}[X_{ij}]$ while $1/P$ times the denominator converges to $\mathrm{E}[N_i(t) + 1]$. Thus the limiting ratio is independent of $t$ and produces the desired result. Since $N_i(t)$ is not a stopping time, this argument cannot be applied to $\hat{\mu}_1(P,t)$. The price to be paid for this consistent estimation of $\mu$ is an increased completion time, which was shown to grow as order $\ln(P)$ in [20] under a variety of distributional assumptions on $\tau_{ij}$.

This discussion shows that, for the estimators described above, one must simulate beyond time $t$ in order to obtain strong consistency as $P \to \infty$. We next show that no matter what estimator is used, one cannot expect to always get the right answer by simply setting a fixed, finite completion time $t$ and "throwing processors" at the problem. More specifically, suppose $t$ is given. Let $\hat{\theta}_X(P,t)$ be any estimator for $\mathrm{E}[X_{ij}]$ that can be constructed from information obtained in the interval $[0,t]$, that is, $\hat{\theta}_X(P,t)$ is a function of $\{(X_{ij}, \tau_{ij}),\ j \le N_i(t),\ i = 1,\ldots,P\}$. We require $\hat{\theta}_X(P,t)$ to be a universally valid estimator in the sense that $\lim_{P\to\infty} \hat{\theta}_X(P,t) = \mathrm{E}[X_{ij}]$ a.s., regardless of the distributions of $X_{ij}$ and $\tau_{ij}$. Suppose now that $\mathrm{P}\{\tau_{ij} > t\} > 0$ and define $Y_{ij} = X_{ij}$ if $\tau_{ij} \le t$ and $Y_{ij} = X_{ij} + 1$ if $\tau_{ij} > t$. Then, $\mathrm{E}[X_{ij}] \ne \mathrm{E}[Y_{ij}]$, however $\hat{\theta}_X(P,t) = \hat{\theta}_Y(P,t)$, since

$$\{(X_{ij}, \tau_{ij}),\ j \le N_i(t),\ i = 1,\ldots,P\} = \{(Y_{ij}, \tau_{ij}),\ j \le N_i(t),\ i = 1,\ldots,P\}. \tag{2.3}$$

Therefore,

$$\mathrm{E}[X_{ij}] = \lim_{P\to\infty} \hat{\theta}_X(P,t) = \lim_{P\to\infty} \hat{\theta}_Y(P,t) \ne \mathrm{E}[Y_{ij}], \tag{2.4}$$

that is, $\hat{\theta}_Y(P,t)$ is not consistent for $(Y_{ij}, \tau_{ij})$. Thus, no such universal estimator exists, and one must be willing to simulate beyond time $t$ on at least some processors in order to obtain a universally valid estimator. This paper will define and analyze the properties of such a class of estimators.

Before defining these estimators, we need to review some results from [16] for the case of a single processor. To prevent introducing new notation, we keep the processor subscript $i$, even though it is not needed in the rest of this section. Define $\bar{X}_i(0) \equiv 0$ and for any $n \ge 1$, define $\bar{X}_i(n) \equiv \sum_{j=1}^{n} X_{ij}/n$. For the case of a single processor, the properties of $\bar{X}_i(N_i(t))$ were studied in

detail in [16]. The basis for determining these properties is the relationship

$$\mathrm{E}[\,\bar{X}_i(N_i(t))] = \mathrm{E}[\,X_{i1};\,\tau_{i1} \le t] = \mu - \mathrm{E}[\,X_{i1};\,\tau_{i1} > t], \tag{2.5}$$

where, for a real-valued r.v. $Y$, $\mathrm{E}[\,Y;\,A]$ denotes $\mathrm{E}[\,YI(A)]$ and $I(A)$ denotes the indicator of the event $A$. This relationship depends on the fact that, given $N_i(t) = k$, $(X_{i1}, \ldots, X_{ik})$ are exchangeable r.v.s. Equation 2.5 has appeared in Pathak [31] and Kremers [22] in the context of survey sampling from a finite population. Kremers also states the result for the so-called infinite population case which corresponds to our probabilistic setting (for a single processor). A special case of Eq. 2.5 when $X_{ij} = \tau_{ij}$ appears on page 93 of Ross [33]. From Equation 2.5, bias expansions can be obtained. For example (Corollary 5 of [16]), if $\mathrm{E}[\,|\,X_{ij}\,|\,] < \infty$ and $\mathrm{E}[\,|\,X_{ij}\tau_{ij}^{p}\,|\,] < \infty$ for some $p > 0$, then $\mathrm{E}[\,\bar{X}_i(N_i(t))] = \mu + o(t^{-p})$, that is, $\lim_{t\to\infty} t^{p}\,|\mathrm{E}[\,\bar{X}_i(N_i(t))] - \mu\,| = 0$. Similarly (Corollary 6 or [16]), if $\mathrm{E}[\,|\,X_{ij}e^{\theta\tau_{ij}}\,|\,] < \infty$ for some $\theta > 0$, then $\mathrm{E}[\,\bar{X}_i(N_i(t))] = \mu + o(e^{-\theta t})$.

Equation 2.5 also suggests an unbiased estimate of $\mu$, as follows. Defining $\tilde{N}_i(t) \equiv \max(1, N_i(t))$, then $\mathrm{E}[\,\bar{X}_i(\tilde{N}_i(t))] = \mu$. In order to form this estimator, one must complete the replication in progress at time $t$ only if no replications have yet been completed (i.e., if $N_i(t) = 0$, or equivalently, $\tau_{i1} > t$). Strong laws and central limit theorems for both $\bar{X}_i(N_i(t))$ and $\bar{X}_i(\tilde{N}_i(t))$ follow directly from standard results in probability and renewal theory: For example,

$$\sqrt{t}\,(\bar{X}_i(\tilde{N}_i(t)) - \mu) \Rightarrow \sigma\mathrm{E}[\,\tau_{ij}]^{1/2}N(0,1), \tag{2.6}$$

as $t \to \infty$ where $\Rightarrow$ denotes convergence in distribution, $N(a, b)$ denotes a normally distributed r.v. with mean $a$ and variance $b$, and $\sigma^2$ is the variance of $X_{ij}$ (assumed finite). Drawing on the results of Chow et al. [3], uniform integrability and moment convergence associated with these central limit theorems are given in [16]. For example, Theorem 6 of [16] states that if $\mathrm{E}[\,|\,X_{ij}\,|^{2p+1+\delta}] < \infty$ and $\mathrm{E}[\tau_{ij}^{2p+\delta}] < \infty$ for some $p > 0$ and $\delta > 0$, then

$$\lim_{t\to\infty} \mathrm{E}\Big[\big|\sqrt{t}\,(\bar{X}_i(\tilde{N}_i(t)) - \mu)\big|^{p}\Big] = \mathrm{E}\Big[\big|N\big(0, \sigma^2\mathrm{E}[\,\tau_{ij}]\big)\big|^{p}\Big]. \tag{2.7}$$

In addition, multidimensional versions of the central limit theorem in Eq. 2.6 are also valid. These can be combined with Taylor series expansions and the uniform integrability of Eq. 2.7 to obtain central limit theorems and bias expansions for nonlinear function estimation.

## 3. PARALLEL ESTIMATORS FOR A SIMPLE MEAN

We now build on the results described in the previous section to derive and analyze three alternative parallel processing point estimators for a simple mean. The first estimator, $\bar{\mu}(P, t)$, has the property that, like $\hat{\mu}_1(P, t)$, it can be formed at exactly time $t$. We define

$$\bar{\mu}(P, t) \equiv \frac{1}{P}\sum_{i=1}^{P} \bar{X}_i(N_i(t)). \tag{3.1}$$

Notice that while $\bar{\mu}(P, t)$ and $\hat{\mu}_1(P, t)$ both make use of the same underlying observations, these observations are combined differently. Our first result concerns the expected value and convergence properties of $\bar{\mu}(P, t)$.

PROPOSITION 3.1. *If* $0 < \tau_{ij} < \infty$ *a.s. and if* $\mathrm{E}[\,|X_{ij}|\,] < \infty$, *then*

(1) $\lim_{t\to\infty} \bar{\mu}(P, t) = \mu$ *a.s.*

(2) *If there exists a finite constant* $B$ *such that* $\tau_{ij} \le B$ *a.s. and if* $B \le t$, *then* $\mathrm{E}[\bar{\mu}(P, t)] = \mu$ *and* $\lim_{P\to\infty} \bar{\mu}(P, t) = \mu$ *a.s.*

(3) *If* $\mathrm{E}[\,|X_{ij}|\,\tau_{ij}^k] < \infty$ *for some* $k > 0$, *then* $\mathrm{E}[\bar{\mu}(P, t)] = \mu + o(t^{-k})$ *and* $\lim_{P\to\infty} \bar{\mu}(P, t) = \mu + o(t^{-k})$ *a.s., i.e.,* $\lim_{t\to\infty} t^k \lim_{P\to\infty} |\bar{\mu}(P, t) - \mu| = 0$ *a.s.*

(4) *If* $\mathrm{E}[\,|X_{ij}|\,e^{\theta\tau_{ij}}] < \infty$ *for some* $\theta > 0$, *then* $\mathrm{E}[\bar{\mu}(P, t)] = \mu + o(e^{-\theta t})$ *and* $\lim_{P\to\infty} \bar{\mu}(P, t) = \mu + o(e^{-\theta t})$ *a.s.*

(5) *If* $\mathrm{E}[\,|X_{ij}|^{1+\delta}] < \infty$ *for some* $\delta > 0$, *and if* $\lim_{P\to\infty} t_P = \infty$, *then* $\bar{\mu}(P, t_P) \Rightarrow \mu$ *as* $P \to \infty$.

PROOF. Result (1) follows by ordinary strong laws for cumulative processes, since $P$ is fixed. Results (2), (3), and (4) essentially follow from the strong law of large numbers and Eq. 2.5, see Corollaries 1, 5, and 6 of [16], respectively. For (5), by Chebyshev's inequality $\mathrm{P}\{|\bar{\mu}(P, t_P) - \mu| > \epsilon\} \le \mathrm{E}[\,|\bar{X}_i(N_i(t_P)) - \mu|\,]/\epsilon$, which converges to zero by Corollary 10 of [16]. □

Proposition 3.1 shows that if $t$ is fixed and $P$ gets large, then $\bar{\mu}(P, t)$ need not converge to $\mu$. However, if $P$ is fixed and $t$ gets large, then $\bar{\mu}(P, t)$ does converge to $\mu$. Since $\bar{\mu}(P, t)$ is biased for finite $t$, we next define an unbiased estimate of $\mu$ in the parallel setting to be

$$\tilde{\mu}_1(P, t) \equiv \frac{1}{P} \sum_{i=1}^{P} \bar{X}_i(\tilde{N}_i(t)). \tag{3.2}$$

This estimator may require simulating past time $t$, since it requires completing at least one replication on each processor. Specifically, one must complete the replication in progress on processor $i$ if and only if $N_i(t) = 0$. The convergence properties of this estimator are stated next. The proof of Proposition 3.2 is not given since it is basically the same as that of Proposition 3.1.

PROPOSITION 3.2. *If* $0 < \tau_{ij} < \infty$ *a.s. and if* $\mathrm{E}[\,|X_{ij}|\,] < \infty$, *then*

(1) $\mathrm{E}[\tilde{\mu}_1(P, t)] = \mu$,

(2) $\lim_{t\to\infty} \tilde{\mu}_1(P, t) = \lim_{P\to\infty} \tilde{\mu}_1(P, t) = \mu$ *a.s.*

(3) *If* $\mathrm{E}[\,|X_{ij}|^{1+\delta}] < \infty$ *for some* $\delta > 0$, *and if* $\lim_{P\to\infty} t_P = \infty$, *then* $\tilde{\mu}_1(P, t_P) \Rightarrow \mu$ *as* $P \to \infty$.

The third estimator we consider is also unbiased, but has a shorter completion time than that associated with $\tilde{\mu}_1(P, t)$. Since $\mathrm{E}[\bar{\mu}(P, t)] = \mathrm{E}[X_{i1} I(\tau_{i1} \le t)]$, in order to obtain an unbiased estimate of $\mu = \mathrm{E}[X_{ij}]$, we need only estimate the remainder term $\mathrm{E}[X_{i1} I(\tau_{i1} > t)]$. Instead of using all $P$ processors (as $\tilde{\mu}_1(P, t)$ does), we use $P_1$ preselected processors to estimate the remainder term. Thus, rather than requiring that at least one replication be

completed on all processors, we only require at least one replication on the $P_1$ preselected processors. We assume that these preselected processors are labeled $i = 1, \ldots, P_1$. Formally, this estimator can be written as

$$\tilde{\mu}_2(P, P_1, t) \equiv \bar{\mu}(P, t) + \frac{1}{P_1} \sum_{i=1}^{P_1} X_{i1} I(\tau_{i1} > t). \tag{3.3}$$

Notice that if $P_1 = P$, then $\tilde{\mu}_2(P, P_1, t) = \tilde{\mu}_1(P, t)$, whereas if $P_1 = 0$, then (by convention) $\tilde{\mu}_2(P, P_1, t) = \bar{\mu}(P, t)$. Proposition 3.3 describes the convergence properties of $\tilde{\mu}_2(P, P_1, t)$. Note that the $P_1$ processors must be preselected: an unbiased estimate would not result, for instance, by taking the first $P_1$ uncompleted replications that actually do complete.

PROPOSITION 3.3. *If* $0 < \tau_{ij} < \infty$ *a.s. and if* $\mathrm{E}[\,|X_{ij}|\,] < \infty$, *then*

(1) *If* $P_1 > 0$, *then* $\mathrm{E}[\tilde{\mu}_2(P, P_1, t)] = \mu$,
(2) $\lim_{t\to\infty} \tilde{\mu}_2(P, P_1, t) = \mu$ *a.s.*
(3) *If* $P \to \infty$ *and* $P_1 \to \infty$, *then* $\tilde{\mu}_2(P, P_1, t) \to \mu$ *a.s.*
(4) *If* $\mathrm{E}[\,|X_{ij}|^{1+\delta}] < \infty$ *for some* $\delta > 0$, $\lim_{P\to\infty} t_P = \infty$, *and* $\lim_{P\to\infty} P_1 = \infty$, *then* $\tilde{\mu}_2(P, P_1, t_P) \Rightarrow \mu$ *as* $P \to \infty$.

We next turn to central limit theorems for these estimators. Define $\mu_t \equiv \mathrm{E}[X_{i1} I(\tau_{i1} \le t)]$, $\sigma^2(t) \equiv \mathrm{Var}[\bar{X}_i(N_i(t))]$, and $\tilde{\sigma}^2 \equiv \mathrm{Var}[\bar{X}_i(\tilde{N}_i(t))]$. We begin with central limit theorems for $\bar{\mu}(P, t)$ and $\tilde{\mu}_1(P, t)$ as either $P \to \infty$ or $t \to \infty$. These are applications of well-known results in applied probability.

PROPOSITION 3.4. *If* $0 < \tau_{ij} < \infty$ *a.s. and if* $\sigma^2 < \infty$, *then*

(1) *If* $\sigma^2(t) < \infty$ *then*, $\sqrt{P}(\bar{\mu}(P, t) - \mu_t) \Rightarrow \sigma(t)N(0, 1)$ *as* $P \to \infty$,
(2) *If* $\tilde{\sigma}^2(t) < \infty$ *then*, $\sqrt{P}(\tilde{\mu}_1(P, t) - \mu) \Rightarrow \tilde{\sigma}(t)N(0, 1)$ *as* $P \to \infty$,
(3) *If* $\mathrm{E}[\tau_{ij}] < \infty$ *then*, $\sqrt{Pt}(\bar{\mu}(P, t) - \mu) \Rightarrow \sigma\,\mathrm{E}[\tau_{ij}]^{1/2}N(0, 1)$ *as* $t \to \infty$,
(4) *If* $\mathrm{E}[\tau_{ij}] < \infty$ *then*, $\sqrt{Pt}(\tilde{\mu}_1(P, t) - \mu) \Rightarrow \sigma\,\mathrm{E}[\tau_{ij}]^{1/2}N(0, 1)$ *as* $t \to \infty$.

Because of its inherent bias, $\bar{\mu}(P, t)$ and its associated central limit theorem (part (1)) cannot be the basis of a valid confidence interval for $\mu$ using a fixed amount of computing time and a very large numbers of processors. On the other hand, for a fixed number of processors and a large amount of time, the bias goes to zero so that the central limit theorem for $\bar{\mu}(P, t)$ can be used to form confidence intervals for $\mu$ (part (3)). Since $\tilde{\mu}_1(P, t)$ is unbiased, its central limit theorems can be used to form confidence intervals for $\mu$ for either a large number of processors (part (2)), or a large amount of time (part (4)).

In order to obtain central limit theorems as both $t \to \infty$ and $P \to \infty$ simultaneously (triangular array central limit theorems), we first note that if $\mathrm{E}[\,|X_{ij}|^{5+\delta}] < \infty$ and $\mathrm{E}[\tau_{ij}^{4+\delta}] < \infty$ for some $\delta > 0$, then by Theorem 6 of [16] (essentially Eq. 2.7):

$$\lim_{t\to\infty} t\sigma^2(t) = \lim_{t\to\infty} t\tilde{\sigma}^2(t) = \mathrm{E}[\tau_{ij}]\sigma^2. \tag{3.4}$$

PROPOSITION 3.5. *If* $0 < \tau_{ij} < \infty$ *a.s.*, $\sigma^2 > 0$, $\mathrm{E}[\,|\,X_{ij}\,|^{5+\delta}] < \infty$, $\mathrm{E}[\tau_{ij}^{4+\delta}] < \infty$ *for some* $\delta > 0$, *and* $\lim_{P\to\infty} t_P = \infty$, *then*

(1) $\sqrt{Pt_P}(\bar{\mu}(P, t_P) - \mu_{t_P}) \Rightarrow \sigma \mathrm{E}[\tau_{ij}]^{1/2} N(0,1)$ *as* $P \to \infty$,

(2) $\sqrt{Pt_P}(\tilde{\mu}_1(P, t_P) - \mu) \Rightarrow \sigma \mathrm{E}[\tau_{ij}]^{1/2} N(0,1)$ *as* $P \to \infty$.

PROOF. We show that the conditions of Lyapounov's theorem (Theorem 7.3 in [2]) are satisfied. For $\tilde{\mu}_1(P, t_P)$, we first show that $\sqrt{P}(\tilde{\mu}_1(P, t_P) - \mu)/\tilde{\sigma}(t_P) \Rightarrow N(0,1)$, from which the result follows by Eq. 3.4. In our case, Lyapounov's condition reduces to showing that

$$\frac{\mathrm{E}\left[\left|\,\overline{X}_i\left(\tilde{N}_i(t_P)\right) - \mu\right|^{2+\epsilon}\right]}{P^{\epsilon/2}\tilde{\sigma}^{2+\epsilon}(t_P)} \to 0, \tag{3.5}$$

as $P \to \infty$ for some small $\epsilon > 0$. This follows by multiplying the numerator and denominator of Eq. 3.5 by $\sqrt{t_P}^{2+\epsilon}$ and applying Eqs. 2.7 and 3.4. The proof for $\bar{\mu}(P, t_P)$ is similar. However, Lyapounov's theorem applies to sums of r.v.s with means 0, which accounts for the centering term being $\mu_{t_P}$ rather than $\mu$. □

As in Proposition 3.4, the central limit theorem for $\bar{\mu}(P, t_P)$ is not necessarily centered about the desired quantity $\mu$ due to the bias of this estimator. Confidence intervals for $\mu$ based on this central limit theorem are centered about $\mu_{t_P}$ and have width proportional to $1/\sqrt{Pt_P}$. Thus, confidence interval coverage is degraded unless $|\mu - \mu_{t_P}|$ is small compared to $1/\sqrt{Pt_P}$. This is the case provided $P$ is not too large with respect to $t_P$. Thus, in practice, with a given number of processors, to obtain valid confidence intervals for $\mu$ using $\bar{\mu}(P, t_P)$ requires simulating for a relatively long time on each processor. No such restrictions apply to $\tilde{\mu}_1(P, t_P)$ because it is unbiased.

To obtain valid confidence intervals for $\mu$ using $\bar{\mu}(P, t_P)$, we basically need to replace the centering term, $\mu_{t_P}$, in the central limit theorem for $\bar{\mu}(P, t_P)$ by $\mu$. The ability to replace $\mu_{t_P}$ by $\mu$ depends on the relative growth rates of $P$ and $t_P$, as well as moment conditions on $X_{ij}$ and $\tau_{ij}$. As discussed above, if the number of processors $P$ grows too quickly with respect to the time constraint $t_P$, then the residual bias remains significant, and the central limit theorem cannot be used to form confidence intervals for $\mu$. To quantify this effect, we next give precise conditions under which the desired central limit theorem is obtained.

PROPOSITION 3.6. *Under the same conditions as in Proposition* 3.5,

$$\sqrt{Pt_P}\left(\bar{\mu}(P, t_P) - \mu\right) \Rightarrow \sigma \mathrm{E}\left[\tau_{ij}\right]^{1/2} N(0,1) \quad \text{as } P \to \infty,$$

*provided either*:

(1) $\mathrm{E}[\,|\,X_{ij}\,|\,\tau_{ij}^k] < \infty$ *and* $P = \mathrm{O}(t_P^{2k-1})$, *or*

(2) $\mathrm{E}[\,|\,X_{ij}\,|\,e^{\theta\tau_{ij}}] < \infty$ *for some* $\theta > 0$, *and* $P = \mathrm{O}(e^{2\theta t_P}/t_P)$.

PROOF. For part (1), since Proposition 3.5 is valid, by Theorem 4.1 of [2], it suffices to show that $\sqrt{Pt_P}\,|\,\mu_{t_P} - \mu\,| \to 0$. But by (3) of Proposition 3.1,

$t_P^k |\mu_{t_P} - \mu| \to 0$, from which the result follows. The proof of part (2) is similar. □

Note that the maximum allowable number of processors increases (with respect to the time constraint) with increasing moment assumptions on $\tau_{ij}$. For example, by the Cauchy-Schwarz inequality, $E[|X_{ij}|\tau_{ij}^k] \leq E[X_{ij}^2]^{1/2}E[\tau_{ij}^{2k}]^{1/2} < \infty$ for $k = 2$ under our base assumptions in Proposition 3.5. Thus, under these base assumptions, we require $P = O(t_P^3)$, or equivalently, $t_P = \Omega(P^{1/3})$ (a sequence $a_P = \Omega(b_P)$ if there exists constants $C$ and $P_0$ such that $a_P \geq Cb_P$ for all $P \geq P_0$).

We next consider central limit theorems for $\tilde{\mu}_2(P, P_1, t)$. The primary intent of this analysis is to show that the growth restrictions (1) and (2) of Proposition 3.6 can be loosened considerably (since $\tilde{\mu}_2(P, P_1, t)$ is unbiased) even if $P_1/P$ is very small. Define $R_i(t) \equiv X_{i1}I(\tau_{i1} > t)$.

PROPOSITION 3.7

(1) *Under the same conditions as in part* (3) *of Proposition* 3.4, $\sqrt{Pt}(\tilde{\mu}_2(P, P_1, t) - \mu) \Rightarrow \sigma E[\tau_{ij}]^{1/2}N(0, 1)$ *as* $t \to \infty$.

(2) *Let* $P_1$ *and* $P \to \infty$ *in such a way that* $P_1/P = \alpha$ *for some* $\alpha$ $(0 < \alpha \leq 1)$. *Let* $\tilde{\sigma}_2^2(t) \equiv \sigma^2(t) - 2\mu_t E[R_i(t)] + \text{Var}[R_i(t)]/\alpha$. *Under the same conditions of part* (1) *of Proposition* 3.4: $\sqrt{P}(\tilde{\mu}_2(P, P_1, t) - \mu) \Rightarrow \tilde{\sigma}_2(t)N(0, 1)$ *as* $P \to \infty$.

(3) *Let* $t_P$, $P_1$ *and* $P \to \infty$ *in such a way that* $P_1/P = \alpha$ *for some* $\alpha$ $(0 < \alpha \leq 1)$. *Assume the conditions of Proposition* 3.5. *Then,* $\sqrt{Pt_P}(\tilde{\mu}_2(P, P_1, t_P) - \mu) \Rightarrow \sigma E[\tau_{ij}]^{1/2}N(0, 1)$.

(4) *Let* $t_P$, $P_1$ *and* $P \to \infty$ *in such a way that* $P_1/P = \alpha_P \to 0$. *Assume the conditions of Proposition* 3.5. *Then,* $\sqrt{Pt_P}(\tilde{\mu}_2(P, P_1, t_P) - \mu) \Rightarrow \sigma E[\tau_{ij}]^{1/2}N(0, 1)$, *provided either*:

   (a) $E[X_{ij}^2\tau_{ij}^k] < \infty$ *and* $P/P_1 = O(t_P^{k-1})$ *for some* $k > 1$, *or*

   (b) $E[X_{ij}^2 e^{\theta\tau_{ij}}] < \infty$ *and* $P/P_1 = O(e^{\theta t_P}/t_P)$ *for some* $\theta > 0$.

PROOF. For part (1), Because of Eqs. 3.3 and (3) of Proposition 3.4, the result will be true, provided $\sqrt{t}|R_i(t)| \Rightarrow 0$ for $i = 1, \ldots, P_1$. But $P\{\sqrt{t}|R_i(t)| > \epsilon\} \leq P\{\tau_{i1} > t\} \to 0$ as $t \to \infty$. For part (2), define $Y_i = \alpha\bar{X}_i(N_i(t)) + R_i(t)$, $Z_i = (1 - \alpha)\bar{X}_i(N_i(t))$, $\bar{Y} = \sum_{i=1}^{P_1} Y_i/P_1$, and $\bar{Z} = \sum_{i=P_1+1}^{P} Z_i/(P - P_1)$. Then, $\tilde{\mu}_2(P, P_1, t) = \bar{Y} + \bar{Z}$. Let $\mu_Y = E[Y_i]$, $\mu_Z = E[Z_i]$ and notice that $\mu = \mu_Y + \mu_Z$. Thus,

$$\sqrt{P}(\tilde{\mu}_2(P, P_1, t) - \mu) = \sqrt{P}(\bar{Y} - \mu_Y) + \sqrt{P}(\bar{Z} - \mu_Z). \tag{3.6}$$

Let $\sigma_Y^2 = \text{Var}[Y_i]$ and $\sigma_Z^2 = \text{Var}[Z_i] = (1 - \alpha)^2\sigma^2(t)$. By the ordinary central limit theorem, $\sqrt{\alpha P}(\bar{Y} - \mu_Y) \Rightarrow \sigma_Y N(0, 1)$ and $\sqrt{(1 - \alpha)P}(\bar{Z} - \mu_Z) \Rightarrow \sigma_Z N(0, 1)$. Since $\bar{Y}$ and $\bar{Z}$ are independent, the convergence in distribution occurs jointly, and the result follows by Eq. 3.6, provided $\tilde{\sigma}_2^2(t) = \sigma_Y^2/\alpha + \sigma_Z^2/(1 - \alpha)$. But,

$$\sigma_Y^2 = \alpha^2 \text{Var}[\bar{X}_i(N_i(t))] + \text{Var}[R_i(t)] + 2\alpha \text{Cov}[\bar{X}_i(N_i(t)), R_i(t)]. \tag{3.7}$$

Since $R_i(t)\bar{X}_i(N_i(t)) = 0$, $\mathrm{Cov}[\bar{X}_i(N_i(t)), R_i(t)] = -\mu_i \mathrm{E}[R_i(t)]$, and the result then follows by simple calculations. Part (3) similarly follows, since it is easily shown that $t_P\tilde{\sigma}_2^2(t_P) \to \mathrm{E}[\tau_{ij}]\sigma^2$. For part (4), adopting the above notation, since $\alpha_P \to 0$, by part (1) of Proposition 3.5, $\sqrt{Pt_P}(\bar{Z} - \mu_Z) \Rightarrow \sigma \mathrm{E}[\tau_{ij}]^{1/2}N(0, 1)$. The result is then true provided $\sqrt{Pt_P} \mid \bar{Y} - \mu_Y \mid \Rightarrow 0$ which, in turn, is true provided $\mathrm{Var}[\sqrt{Pt_P}(\bar{Y} - \mu_Y)] = Pt_P\sigma_Y^2/P_1 \to 0$. But

$$\frac{Pt_P\sigma_Y^2}{P_1} = \alpha_P t_P \sigma^2(t_P) - 2t_P\mu_{t_P}\mathrm{E}[R_i(t_P)] + (t_P P/P_1)\,\mathrm{Var}[R_i(t_P)]. \quad (3.8)$$

The first term on the right-hand side of Eq. 3.8 converges to 0 by Eq. 3.4. For the second term, $\mu_{t_P} \to \mu$ and $t_P\mathrm{E}[R_i(t_P)] \to 0$ (provided $\mathrm{E}[\mid X_{ij} \mid \tau_{ij}] < \infty$). For the third term, under assumption (a), $t_P^k \mathrm{Var}[R_i(t_P)] \to 0$ so $(t_P P/P_1)$ $\mathrm{Var}[R_i(t_P)] \to 0$ provided $t_P P/P_1 = \mathrm{O}(t_P^k)$, that is, $P/P_1 = \mathrm{O}(t_P^{k-1})$. The result similarly holds under assumption (b). Therefore, $\mathrm{Var}[\sqrt{Pt_P}(\bar{Y} - \mu_Y)] \to 0$, as required. □

In Proposition 3.7, for fixed $P_1$ and $P$, a properly centered central limit theorem is again obtained as $t \to \infty$ (part 1). Note that the central limit theorem for $\tilde{\mu}_1(P, t_P)$ puts no restrictions on the relative values of $P$ and $t_P$. However, in order for $\bar{\mu}(P, t_P)$ to have a properly centered central limit theorem, a minimum growth rate for $t_P$ with respect to $P$ is required. In contrast, the central limit theorem for $\tilde{\mu}_2(P, P_1, t_P)$ places no direct restrictions on the relative values of $P$ and $t_P$, but rather requires a minimum growth rate for $t_P$ with respect to the ratio $P/P_1$. (This is actually a sufficient condition, see, e.g., exercise 15 on page 49 of [4].) Note also that, for fixed $t$, $\tilde{\sigma}_2(t)$ is a decreasing function of $\alpha = P_1/P$. Thus, there is some variance inflation by not preselecting all of the processors ($\alpha = 1$). However, parts (3) and (4) of Proposition 3.7 show that this variance inflation disappears asymptotically provided either $\alpha$ is fixed or $\alpha = \alpha_P$ does not approach zero too quickly with respect to $t_P$.

When one considers estimating a function $g(\mu)$ by, say, $g(\tilde{\mu}_1(P, t_P))$, then some bias may be introduced (i.e., $\mathrm{E}[g(\tilde{\mu}_1(P, t_P))] \neq g(\mu)$) if $g$ is nonlinear even through $\tilde{\mu}_1(P, t_P)$ is unbiased. The standard approach to characterizing this bias is via Taylor series expansions and moment convergence in the central limit theorem of the underlying point estimate. Therefore, in order to obtain such bias expansions, we need to establish uniform integrability and moment convergence of the underlying point estimators. Proposition 3.8 does this by characterizing conditions under which the various point estimates are uniformly integrable.

PROPOSITION 3.8. *Under the same conditions as in Proposition* 3.5,

(1) $(\sqrt{Pt_P}(\tilde{\mu}_1(P, t_P) - \mu))^2$ *is uniformly integrable as* $P \to \infty$, *and* $\lim_{P\to\infty}$ $\mathrm{E}[(\sqrt{Pt_P}(\tilde{\mu}_1(P, t_P) - \mu))^2] = \sigma^2\mathrm{E}[\tau_{ij}]$.

(2) *If, in addition, either conditions* (1) *or* (2) *of Proposition* 3.6 *hold, then* $(\sqrt{Pt_P}(\bar{\mu}(P, t_P) - \mu))^2$ *is uniformly integrable as* $P \to \infty$, *and* $\lim_{P\to\infty}$ $\mathrm{E}[(\sqrt{Pt_P}(\bar{\mu}(P, t_P) - \mu))^2] = \sigma^2\mathrm{E}[\tau_{ij}]$.

(3) *Under the same conditions of Proposition* 3.7, *parts* (3) *or* (4), $(\sqrt{Pt_P}(\tilde{\mu}_2(P, P_1, t_P) - \mu))^2$ *is uniformly integrable as* $P \to \infty$, *and* $\lim_{P\to\infty} \mathrm{E}[(\sqrt{Pt_P}(\tilde{\mu}_2(P, P_1, t_P) - \mu))^2] = \sigma^2 \mathrm{E}[\tau_{1j}]$.

PROOF. For part (1), by Proposition 3.5, $(\sqrt{Pt_P}(\tilde{\mu}_1(P, t_P) - \mu))^2 \Rightarrow \sigma^2 \mathrm{E}[\tau_{1j}] N(0,1)^2$. Furthermore,

$$\mathrm{E}\left[\left(\sqrt{Pt_P}\left(\tilde{\mu}_1(P, t_P) - \mu\right)\right)^2\right] = Pt_P \mathrm{Var}\left[\tilde{\mu}_1(P, t_P)\right] = t_P \tilde{\sigma}^2(t_P) \to \sigma^2 \mathrm{E}[\tau_{1j}] \tag{3.9}$$

by Eq. 3.4. Therefore, the result follows by Theorem 5.4 of [2]. For part (2),

$$Pt_P \mathrm{E}\left[\left(\bar{\mu}(P, t_P) - \mu\right)^2\right] = Pt_P\left(\mathrm{Var}\left[\bar{\mu}(P, t_P)\right] + \left(\mu_{t_P} - \mu\right)^2\right) = t_P \sigma^2(t_P) + Pt_P\left(\mu_{t_P} - \mu\right)^2. \tag{3.10}$$

The first term on the right-hand side of Eq. 3.10 converges to $\sigma^2 \mathrm{E}[\tau_{1j}]$ by Eq. 3.4, while the second term converges to zero by the same argument as in the proof of Proposition 3.6. Combining this with the convergence in distribution of Proposition 3.6 yields the result. The moment convergence for part (3) was basically established in the proof of Proposition 3.7, and the result then follows similarly. □

We next state multidimensional versions of Propositions 3.5, 3.6, and 3.7. These will also be needed in the next section for nonlinear function estimation. These results are simply shown by applying the Cramér–Wold device (Theorem 7.7 of [2]) to Propositions 3.5–3.7. We require some notation. Let $\mathbf{X}_{ij} = (X_{ij}(1), \ldots, X_{ij}(d))$ be a $d$-dimensional vector valued output of replication $j$ on processor $i$ and let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ where $\mu_a = \mathrm{E}[X_{ij}(a)]$ and define $C_{ab} = \mathrm{Cov}[X_{ij}(a), X_{ij}(b)]$. We now define multidimensional analogues $\bar{\boldsymbol{\mu}}(P, t_P)$, $\tilde{\boldsymbol{\mu}}_1(P, t_P)$ and $\tilde{\boldsymbol{\mu}}_2(P, P_1, t_P)$ of $\bar{\mu}(P, t_P)$, $\tilde{\mu}_1(P, t_P)$ and $\tilde{\mu}_2(P, P_1, t_P)$, respectively, as follows. Define $\overline{\mathbf{X}}_i(n) = (\overline{X}_i(n, 1), \ldots, \overline{X}_i(n, d))$ where $\overline{X}_i(n, a) \equiv \sum_{j=1}^{n} X_{ij}(a)/n$. Component $a$ of $\bar{\boldsymbol{\mu}}(P, t_P)$ is then defined to be

$$\bar{\mu}(P, t_P, a) \equiv \frac{1}{P} \sum_{i=1}^{P} \overline{X}_i\left(N_i(t), a\right). \tag{3.11}$$

The vectors $\tilde{\boldsymbol{\mu}}_1(P, t_P)$ and $\tilde{\boldsymbol{\mu}}_2(P, P_1, t_P)$ are defined analogously. Let $\mathbf{N}(\mathbf{0}, \mathbf{A})$ denote a multidimensional normally distributed random vector with means 0 and Variance/Covariance matrix $\mathbf{A}$.

PROPOSITION 3.9. *Assume that the conditions of Proposition* 3.5 *apply to each component of* $\mathbf{X}_{ij}$. *Then*,

$$\sqrt{Pt_P}\left(\tilde{\boldsymbol{\mu}}_1(P, t_P) - \boldsymbol{\mu}\right) \Rightarrow \mathrm{E}[\tau_{1j}]^{1/2} \mathbf{N}(\mathbf{0}, \mathbf{C}) \quad \text{as } P \to \infty.$$

The same multidimensional central limit theorem holds for $\bar{\boldsymbol{\mu}}(P, t_P)$ and $\tilde{\boldsymbol{\mu}}_2(P, P_1, t_P)$, provided the conditions of Propositions 3.6 and 3.7 apply to each component of the respective random vectors.

We conclude this section with a discussion of the formation of confidence intervals for $\mu$. The central limit theorems of Propositions 3.4–3.7 can be used as the basis for such confidence intervals; however, in practice, the variance terms in these limit theorems are not known and must be estimated. As usual, this presents no theoretical obstacles, since the variance terms can be consistently estimated (see, e.g., [7]). However, there are several ways the variance can be estimated and the appropriate estimator depends on whether $t \to \infty$, $P \to \infty$, or $t$ and $P \to \infty$. We therefore outline the appropriate procedures for $\tilde{\mu}_1(P, t)$. Analogous results also hold for $\bar{\mu}(P, t)$ and $\tilde{\mu}_2(P, P_1, t)$.

First consider the case when $t$ remains fixed, but $P \to \infty$. Define

$$\hat{\sigma}_1^2(P, t) \equiv \frac{1}{P-1} \sum_{i=1}^{P} \left( \bar{X}_i(\tilde{N}_i(t)) - \tilde{\mu}_1(P, t) \right)^2. \tag{3.12}$$

Then $\lim_{P\to\infty} \hat{\sigma}_1^2(P, t) = \tilde{\sigma}^2(t)$ a.s. Combining this with the central limit theorem of part (2) of Proposition 3.4, we obtain $\sqrt{P}(\tilde{\mu}_1(P, t) - \mu)/\hat{\sigma}_1(P, t) \Rightarrow N(0, 1)$ as $P \to \infty$ (the above assumes that $0 < \tilde{\sigma}^2(t) < \infty$). From this central limit theorem, an approximate (say) 90 percent confidence interval for $\mu$ is $\tilde{\mu}_1(P, t) \pm 1.645\hat{\sigma}_1(P, t)/\sqrt{P}$. If $P$ is fixed and $t \to \infty$, define

$$\tilde{\tau}(P, t) \equiv \frac{1}{P} \sum_{i=1}^{P} \frac{1}{\tilde{N}_i(t)} \sum_{j=1}^{\tilde{N}_i(t)} \tau_{ij} \tag{3.13}$$

and

$$\hat{\sigma}_2^2(P, t) \equiv \tilde{\tau}(P, t) \left[ \left( \frac{1}{P} \sum_{i=1}^{P} \frac{1}{\tilde{N}_i(t)} \sum_{j=1}^{\tilde{N}_i(t)} X_{ij}^2 \right) - \tilde{\mu}_1(P, t)^2 \right]. \tag{3.14}$$

Then $\lim_{t\to\infty} \hat{\sigma}_2^2(P, t) = \mathrm{E}[\tau_{ij}]\sigma^2$ a.s. (assuming these terms are finite), and therefore $\sqrt{Pt}(\tilde{\mu}_1(P, t) - \mu)/\hat{\sigma}_2(P, t) \Rightarrow N(0, 1)$ as $t \to \infty$. Finally, if $P \to \infty$ and $t_P \to \infty$, then $\hat{\sigma}_2^2(P, t_P) \Rightarrow \mathrm{E}[\tau_{ij}]\sigma^2$ (assuming $\mathrm{E}[|X_{ij}|^{2+\delta}] < \infty$ and $\mathrm{E}[\tau_{ij}^{1+\delta}] < \infty$ by Proposition 3.2). Therefore, $\sqrt{Pt_P}(\tilde{\mu}_1(P, t_P) - \mu)/\hat{\sigma}_2(P, t_P) \Rightarrow N(0, 1)$.

## 4. NONLINEAR FUNCTION ESTIMATION

In this section we consider estimating a real valued nonlinear function $g(\mu)$ by either $g(\bar{\mu}(P, t_P))$, $g(\tilde{\mu}_1(P, t_P))$, or $g(\tilde{\mu}_2(P, P_1, t_P))$. This problem arises in many contexts, for instance, in variance estimation where $\mathbf{X}_{ij} = (X_{ij}, X_{ij}^2)$ (i.e., $X_{ij}(2) = X_{ij}(1)^2$) and $g(\mu) = \mu_2 - \mu_1^2$. Another application is steady state estimation in regenerative simulations, in which one is interested in estimating ratios of the form $g(\mu) = \mu_1/\mu_2$. We only consider situations in which $t_P$, $P_1$, and $P$ all $\to \infty$. First note that since $\bar{\mu}(P, t_P) \Rightarrow \mu$, $\tilde{\mu}_1(P, t_P) \Rightarrow \mu$ and $\tilde{\mu}_2(P, P_1, t_P) \Rightarrow \mu$ (under the minimal moment conditions given in Section 3), then $g(\bar{\mu}(P, t_P)) \Rightarrow g(\mu)$, $g(\tilde{\mu}_1(P, t_P)) \Rightarrow g(\mu)$ and $g(\tilde{\mu}_2(P, P_1, t_P)) \Rightarrow g(\mu)$, provided that $g$ is continuous at $\mu$.

Define $g_a = \partial g/\partial x_a|_{\mathbf{x}=\mu}$, and $G_{ab} = \partial^2 g/\partial x_a \partial x_b|_{\mathbf{x}=\mu}$. Let $C_k(\mu)$ be the set of functions having finite continuous derivatives of order $j$ for $j = 0, \ldots, k$

in a neighborhood of $\mu$. We next use the standard technique of combining central limit theorems with Taylor series expansions to obtain central limit theorems and bias expansions for $g(\bar{\mu}(P, t_P))$, $g(\tilde{\mu}_1(P, t_P))$ and $g(\tilde{\mu}_2(P, P_1, t_P))$ (see, e.g., Chs. 27 and 28 of [6] or Ch. 2 of [24]).

PROPOSITION 4.1. *Let $g \in C_1(\mu)$ and assume the conditions of Proposition 3.5 apply to each component of $\mathbf{X}_{ij}$. Define*

$$\sigma_1^2 \equiv \mathrm{E}[\tau_{ij}] \sum_{a=1}^{d} \sum_{b=1}^{d} g_a C_{ab} g_b$$

*and assume $0 < \sigma_1^2 < \infty$. Let $t_P$ and $P \to \infty$.*

(1) *If $P_1/P = \alpha$ for some $\alpha$ $(0 < \alpha \leq 1)$. $\sqrt{Pt_P}(g(\tilde{\mu}_2(P, P_1, t_P)) - g(\mu)) \Rightarrow \sigma_1 N(0, 1)$.*

(2) *$\sqrt{Pt_P}(g(\bar{\mu}(P, t_P)) - g(\mu_{t_P})) \Rightarrow \sigma_1 N(0, 1)$.*

(3) *If conditions (1) or (2) of Proposition 3.6 apply to each component of $\mathbf{X}_{ij}$, then $\sqrt{Pt_P}(g(\bar{\mu}(P, t_P)) - g(\mu)) \Rightarrow \sigma_1 N(0, 1)$.*

(4) *If $P_1/P = \alpha_P \to 0$ and conditions (a) or (b) of part (4) of Proposition 3.7 apply to each component of $\mathbf{X}_{ij}$, then $\sqrt{Pt_P}(g(\tilde{\mu}_2(P, P_1, t_P)) - g(\mu)) \Rightarrow \sigma_1 N(0, 1)$.*

PROOF. For simplicity of notation, we assume that $d = 1$ (and use the notation of Section 3). We only establish results (2) and (3): the other cases and the multidimensional versions can be shown, without complications, along similar lines. Using a first order Taylor series expansion, write $g(\bar{\mu}(P, t_P)) = g(\mu_{t_P}) + g'(\xi_P)(\bar{\mu}(P, t_P) - \mu_{t_P})$ where $\xi_P$ is on the line segment between $\mu_{t_P}$ and $\bar{\mu}(P, t_P)$. Since $\mu_{t_P} \Rightarrow \mu$ and $\bar{\mu}(P, t_P) \Rightarrow \mu$, $\xi_P \Rightarrow \mu$ and, therefore, $g'(\xi_P) \Rightarrow g'(\mu)$. Therefore, by Proposition 3.5,

$$\begin{aligned} \sqrt{Pt_P}\left(g\left(\bar{\mu}(P, t_P)\right) - g\left(\mu_{t_P}\right)\right) &= g'\left(\xi_P\right)\sqrt{Pt_P}\left(\bar{\mu}(P, t_P) - \mu_{t_P}\right) \\ &\Rightarrow g'(\mu)\sigma\mathrm{E}[\tau_{ij}]^{1/2} N(0, 1). \end{aligned} \tag{4.1}$$

For part (3), do the Taylor series about $\mu$ rather than $\mu_{t_P}$ and apply the central limit theorem of Proposition 3.6 □

Note that part (1) of this Proposition also applies to $\tilde{\mu}_1(P, t_P)$, since $\tilde{\mu}_2(P, P_1, t_P) = \tilde{\mu}_1(P, t_P)$ for $\alpha = 1$ $(P_1 = P)$.

We next turn to bias expansions. These can be established under a broad variety of moment assumptions and regularity conditions on the function $g$. For example, in the case of a single processor, expansions for $\mathrm{E}[g(\bar{\mathbf{X}}(N_t(t)))]$ and $\mathrm{E}[g(\bar{\mathbf{X}}(\tilde{N}_t(t)))]$ were derived in [16] under the assumption that the function $g$ is bounded a.s. (and $g \in C_2(\mu)$, plus certain moment assumptions). In [15], the expansion for $\mathrm{E}[g(\bar{\mathbf{X}}(N_t(t)))]$ was shown to be valid provided that $g$ is bounded by a polynomial of degree $r$ for any $r \geq 0$ (i.e., $|g(\bar{\mathbf{X}}(N_t(t)))| \leq A + B\|\bar{\mathbf{X}}(N_t(t)) - \mu\|^r$, and $g \in C_2(\mu)$, plus somewhat different moment conditions). This is true, provided $g$ has bounded partial derivatives of order $r$, for example. For the simple case of a function of a mean of a deterministic

number of i.i.d. r.v.s, Chapter 2 of [24] contains such bias expansions, provided $g$ has bounded partial derivatives of order $r$ ($r \geq 3$) and the $r$-th moments of the underlying r.v.s are finite. In the interest of simplicity, we state the results under conditions that make the proof both simple and transparent. Define

$$B \equiv \frac{\mathrm{E}[\tau_{ij}]}{2} \sum_{a=1}^{d} \sum_{b=1}^{d} G_{ab} C_{ab}. \tag{4.2}$$

PROPOSITION 4.2. *Let $g \in C_2(\mu)$ and assume that all of $g$'s partial derivatives of order 2 are bounded everywhere. Suppose that $t_P$ and $P_1 = P_1(P) \to \infty$ as $P \to \infty$.*

(1) *If the conditions of part* (1) *of Proposition* 3.8 *apply to each component of* $\mathbf{X}_{ij}$, *then* $\mathrm{E}[g(\tilde{\mu}_1(P, t_P))] = g(\mu) + B/(Pt_P) + o(1/Pt_P)$.

(2) *If the conditions of part* (2) *of Proposition* 3.8 *apply to each component of* $\mathbf{X}_{ij}$, *then* $\mathrm{E}[g(\bar{\mu}(P, t_P))] = g(\mu) + B/(Pt_P) + o(1/Pt_P)$ *as* $P \to \infty$, *provided either*:

   (a) $\mathrm{E}[\,|X_{ij}|\,\tau_{ij}^k] < \infty$ *and* $P = O(t_P^{k-1})$, *or*

   (b) $\mathrm{E}[\,|X_{ij}|\,e^{\theta\tau_{ij}}] < \infty$ *for some* $\theta > 0$, *and* $P = O(e^{\theta t_P}/t_P)$.

(3) *If the conditions of part* (3) *of Proposition* 3.8 *apply to each component of* $\mathbf{X}_{ij}$, *then* $\mathrm{E}[g(\tilde{\mu}_2(P, P_1, t_P))] = g(\mu) + B/(Pt_P) + o(1/Pt_P)$.

PROOF. We again assume that $d = 1$ and show the result for $g(\bar{\mu}(P, t_P))$. Using a second order Taylor series expansion, we have

$$Pt_P\big(g(\bar{\mu}(P, t_P)) - g(\mu)\big) = g'(\mu)Pt_P\big(\bar{\mu}(P, t_P) - \mu\big) + \frac{g''(\xi_P)}{2} Pt_P\big(\bar{\mu}(P, t_P) - \mu\big)^2 \tag{4.3}$$

where $\xi_P$ is on the line segment between $\mu$ and $\bar{\mu}(P, t_P)$. The expectation of the first term on the right-hand side of Eq. 4.3 equals $g'(\mu)Pt_P\mathrm{E}[X_{i1}; \tau_{i1} > t_P]$ which converges to zero under assumptions (a) or (b). Again, $\xi_P \Rightarrow \mu$, so the second term on the right-hand side of Eq. 4.3 converges in distribution to $(1/2)g''(\mu)\sigma^2\mathrm{E}[\tau_{ij}]N(0, 1)^2$ by Proposition 3.5. So we will be done if $g''(\xi_P)Pt_P(\bar{\mu}(P, t_P) - \mu)^2$ is uniformly integrable. But this follows, since there exists a finite constant $M$ such that $|g''(x)| \leq M$ for all $x$. The proofs for $\tilde{\mu}_1(P, t_P)$ and $\tilde{\mu}_2(P, P_1, t_P)$ are similar, except that the expectations of the first order term in the Taylor series expansions are identically equal to zero since $\mathrm{E}[\tilde{\mu}_1(P, t_P)] = \mathrm{E}[\tilde{\mu}_2(P, P_1, t_P)] = \mu$. □

Proposition 4.2 states that (under suitable regularity conditions on $g$ and growth restrictions on $P$, $P_1$, and $t_P$) the bias goes to zero as a constant divided by the total simulation effort $P \times t_P$. Jackknifing [27] is one method that can be used to mitigate bias due to nonlinearity effects. In [15], the jackknife is explored in the setting of a single budget-constrained processor. We intend to study the budget-constrained jackknife estimator in the multiprocessor setting in future work.

## 5. COMPLETION TIME ANALYSIS

In this section we analyze properties of the random completion time associated with $\tilde{\mu}_2(P, P_1, t)$ (or with $P = P_1$, $\tilde{\mu}_1(P, t)$). Let $F(x) = \mathrm{P}\{\tau_{ij} \leq x\}$, $\bar{F}(x) = 1 - F(x)$, and $M_n = \max(\tau_{i1}, \ldots, \tau_{n1})$. Let $T_i(t) = \max(\tau_{i1}, t)$: $T_i(t)$ is the completion time on processor $i$ $(1 \leq i \leq P_1)$ given the budget constraint $t$ (actually, budget guideline is a better term). Then

$$M(t, P_1) \equiv \max_{1 \leq i \leq P_1} (T_i(t)) = \max(t, M_{P_1}) \tag{5.1}$$

is the completion time of the simulation experiment. We are also interested in the total processing time on the preselected processors,

$$T(t, P_1) \equiv \sum_{i=1}^{P_1} T_i(t), \tag{5.2}$$

and in the number of "active" processors at time $t$,

$$A(t, P_1) \equiv \sum_{i=1}^{P_1} I(\tau_{i1} > t), \tag{5.3}$$

that is, $A(t, P_1)$ is the number of processors required to simulate beyond time $t$.

Since $T(t, P_1)$ and $A(t, P_1)$ are just sums of i.i.d. r.v.s, their limiting behavior (as $P_1 \to \infty$) can be described by standard strong laws and central limit theorems. Similarly, since $M(t, P_1)$ is basically a maximum of i.i.d. r.v.s, its limiting behavior can be derived from results in extreme value theory (see, e.g., [23]). We give a sampling of such results. Note that the properties of these random variables can be determined directly from the values of $t$ and $P_1$. If $t$ and $P_1$ are viewed as functions of $P$, then these properties depend on $P$ only indirectly as expressed by the relationships $t = t_P$ and $P_1 = P_1(P)$.

Consider first the number of active processors $A(t, P_1)$. Note that $A(t, P_1)$ is binomially distributed with parameters $\bar{F}(t)$ and $P_1$. Therefore, if $t$ remains fixed and $P_1 \to \infty$, then $A(t, P_1)/P_1$ obeys a strong law and is approximately normally distributed with mean $\bar{F}(t)$ and variance $F(t)\bar{F}(t)/P_1$. However, if $P$, $P_1$ and $t_P \to \infty$ in such a way that $P_1\bar{F}(t_P) \to \alpha$ $(0 < \alpha < \infty)$, then $A(t_P, P_1)$ converges in distribution to a Poisson r.v. with parameter $\alpha$ (see Section V.5 of [9]). For example, if $\tau_{ij}$ is exponentially distributed with rate $\lambda$ $(\bar{F}(t) = e^{-\lambda t})$ and $t_P = (1/\lambda)\ln(P_1/\alpha)$, then the Poisson convergence is obtained.

Turning now to the total processing time on preselected processors, by Eq. 5.2, if $t$ remains fixed and $P_1 \to \infty$, then $T(t, P_1)/P_1$ obeys both a strong law and central limit theorem. Now consider the behavior as $t \to \infty$. Notice that $T(t, P_1)/P_1 t$ is the ratio of the actual computing time to the planned computing time (on the preselected processors) and $(T(t, P_1) - P_1 t)/P_1$ is the average excess (unplanned) computing time per processor.

PROPOSITION 5.1. *If* $\mathrm{E}[\tau_{ij}] < \infty$, *then as* $t \to \infty$,

(1) $|T(t, P_1) - P_1 t|/P_1 \Rightarrow 0$,

(2) $T(t, P_1)/(tP_1) \Rightarrow 1$.

PROOF. For part (1), since $|T(t, P_1) - P_1 t| \le \sum_{i=1}^{P_1} |T_i(t) - t|$, $\mathbf{P}\{|T(t, P_1) - P_1 t|/P_1 > \epsilon\} \le \mathrm{E}[|T(t, P_1) - P_1 t|]/(P_1\epsilon) \le \mathrm{E}[|T_i(t) - t|]/\epsilon$

$$= \int_{\{x>t\}} (x - t)\, dF(x)/\epsilon \le \int_{\{x>t\}} x\, dF(x)/\epsilon. \tag{5.4}$$

But the right-hand side of Eq. 5.4 converges to 0 since $\mathrm{E}[\tau_{ij}] < \infty$. The proof of part (2) is similar. □

Notice that the convergence in Proposition 5.1 is obtained even if $P_1 \to \infty$ along with $t$. Thus, for large $t$, no matter how many processors there are, the ratio of actual computing to planned computing converges (in probability) to one.

We now turn to the completion time. We assume that $F$ is in the (maximum) domain of attraction of a (finite) extreme value distribution, that is, there exist constants $a_n$ and $b_n$ and a finite r.v. $X^*$ such that $a_n(M_n - b_n) \Rightarrow X^*$. For example, for the exponential distribution with parameter $\lambda$, $a_n = \lambda$, $b_n = \ln(n)/\lambda$ and $P\{X^* \le x\} = \exp(-e^{-x})$. In addition, in this case there is a well-known closed form expression for $\mathrm{E}[M_n]$: $\mathrm{E}[M_n] = (1/\lambda)H_n$ where $H_n = \sum_{i=1}^{n}(1/i) \approx \ln(n)$.

PROPOSITION 5.2. *If* $a_n(M_n - b_n) \Rightarrow X^*$ *and* $\lim_{P\to\infty} P_1 = \infty$, *then*

(1) *If* $\lim_{P\to\infty} a_{P_1}(t_P - b_{P_1}) = -\infty$, *then* $a_{P_1}(M(t_P, P_1) - b_{P_1}) \Rightarrow X^*$.

(2) *If* $\lim_{P\to\infty} a_{P_1}(t_P - b_{P_1}) = \alpha(-\infty < \alpha < \infty)$, *then* $a_{P_1}(M(t_P, P_1) - b_{P_1}) \Rightarrow max(\alpha, X^*)$.

(3) *If* $\lim_{P\to\infty} a_{P_1}(t_P - b_{P_1}) = +\infty$, *then* $\mathrm{P}\{M(t_P, P_1) = t_P\} \to 1$.

PROOF. For part (1), we show $a_{P_1}(M(t_P, P_1) - M_{P_1}) \Rightarrow 0$.

$$\begin{aligned} \mathrm{P}\{|a_{P_1}(M(t_P, P_1) - M_{P_1})| > \epsilon\} &\le \mathrm{P}\{M_{P_1} \le t_P\} \\ &= \mathrm{P}\{a_{P_1}(M_{P_1} - b_{P_1}) \le a_{P_1}(t_{P_1} - b_{P_1})\} \to 0. \end{aligned} \tag{5.5}$$

Part (2) follows from the continuity of the maximum operator. For part (3),

$$\begin{aligned} \mathbf{P}\{M(t_P, P_1) = t_P\} &= \mathbf{P}\{M_{P_1} \le t_P\} \\ &= \mathbf{P}\{a_{P_1}(M_{P_1} - b_{P_1}) \le a_{P_1}(t_{P_1} - b_{P_1})\} \to 1. \quad \square \end{aligned} \tag{5.6}$$

Note that part (1) of Proposition 5.2 will typically apply if $t_P$ remains fixed, in which case $M(t_P, P_1)$ inherits the limiting distribution of $M_{P_1}$.

We next consider the combined implications of Propositions 3.7 and 5.2 in a particular case. Suppose $\tau_{ij}$ has an exponential distribution with mean 1. Let $t_P = \beta \ln(P_1)$. For $\beta < 1$, part (1) of Proposition 5.2 applies, so $\mathrm{E}[M(t_P, P_1)] \approx \ln(P_1)$. For $\beta > 1$, part (3) of Proposition 5.2 applies, so $\mathrm{E}[M(t_P, P_1)] \approx t_P = \beta \ln(P_1)$. Now if $X_{ij}$ is bounded, then the moment condition of part (4) of Proposition 3.7 is satisfied for $\theta < 1$. Thus, in order for

$\tilde{\mu}_2(P, P_1, t_P)$ to obey the proper central limit theorem, we must have $P/P_1 = O(e^{\theta t_P}/t_P)$, that is, $P \le BP_1^{1+\theta\beta}/(\beta \ln(P_1))$ for some constant $B$. Now let $P = P_1^{1+\theta\beta}/(\beta \ln(P_1))$ and suppose we were to use $\tilde{\mu}_1(P, t_P)$ (i.e., we insist on at least one replication on all processors). The expected completion time associated with $\tilde{\mu}_1(P, t_P)$ then grows like $(1 + \theta\beta)\ln(P_1) - \ln(\beta \ln(P_1))$. For $\theta \approx \beta \approx 1$, the expected completion time using $\tilde{\mu}_1(P, t_P)$ is then nearly twice that when using $\tilde{\mu}_2(P, P_1, t_P)$. However, for $\theta \approx 1$ and $\beta > 1$, the ratio of the expected completion times is approximately $(1 + \beta)/\beta \approx 1$ for large $\beta$. Note that in this case, the difference in expected completion times basically grows like $\ln(P_1)$.

The above analysis has considered nondegenerate limits for $A(t_P, P_1)$, $T(t_P, P_1)$, and $M(t_P, P_1)$. In the exponential example, these nondegenerate limits are obtained if $t_P \sim \ln(P_1)$. Clearly, if $t_P \to \infty$ very quickly, then $A(t_P, P_1) = 0$, $T(t_P, P_1) = t_P P_1$, and $M(t_P, P_1) = t_P$ with high probability (see, e.g., part (3) of Proposition 5.2). We conclude this section by analyzing the convergence rates of these r.v.s for large $t_P$ in more detail.

PROPOSITION 5.3. *Let* $E[\tau_{ij}^{k+m+2}] < \infty$ *for some* $k > 0$ *and* $m \ge 0$ *and let* $P_1 = O(t_P^k)$. *If* $\lim_{P\to\infty} P_1 = \infty$ *and* $\lim_{P\to\infty} t_P = \infty$, *then*

(1) $\lim_{P\to\infty} t_P^{m+2} E[A(t_P, P_1)] = 0$.
(2) $\lim_{P\to\infty} t_P^{m+1} E[T(t_P, P_1) - t_P P_1] = 0$.
(3) $\lim_{P\to\infty} t_P^m \, Var[T(t_P, P_1)] = 0$.
(4) $\lim_{P\to\infty} t_P^{m/2} E[M(t_P, P_1) - t_P] = 0$.

PROOF. For part (1),

$$t_P^{m+2} E\left[A(t_P, P_1)\right] = t_P^{m+2} P_1 \bar{F}(t_P) \le C t_P^{m+k+2} \bar{F}(t_P) \to 0 \tag{5.7}$$

by exercise 15 on page 46 of [4]. For part (2), arguing similarly, as in Proposition 5.1,

$$\begin{aligned} t_P^{m+1} E\left[\left|T(t_P, P_1) - P_1 t_P\right|\right] &\le t_P^{m+1} P_1 E\left[\left|T_i(t_P) - t_P\right|\right] \\ &\le C t_P^{m+1+k} E\left[\left|T_i(t_P) - t_P\right|\right] \to 0 \end{aligned} \tag{5.8}$$

by Corollary 8 of [16]. For part (3), $\mathrm{Var}[T(t_P, P_1)] = P_1 \mathrm{Var}[T_i(t_P)]$, and the result follows along similar lines by showing that $t_P^{m+k} \mathrm{Var}[T_i(t_P)] \to 0$. For part (4),

$$t_P^{m/2} E\left[M(t_P, P_1) - t_P\right] \le t_P^{m/2}\left[E\left[T_i(t_P) - t_P\right] + \frac{P_1 - 1}{\sqrt{2P_1 - 1}} \mathrm{Var}\left[T_i(t_P)\right]^{1/2}\right] \tag{5.9}$$

by the global bound (Eq. 4.2.6) on page 59 of [8]). The result follows similarly. □

The analog of Proposition 5.3, under moment generating type assumptions on $\tau_{ij}$, is stated below. Its proof is essentially identical to that of Proposition 5.3.

PROPOSITION 5.4. *Let* $\mathrm{E}[\tau_{ij}^2 e^{\theta \tau_{ij}}] < \infty$ *for some* $\theta > 0$ *and let* $P_1 = \mathrm{O}(e^{\theta_1 t_P})$ *where* $0 < \theta_1 \leq \theta$. *Let* $\theta_2 = \theta - \theta_1$. *If* $lim_{P\to\infty}\ P_1 = \infty$ *and* $lim_{P\to\infty}\ t_P = \infty$, *then*

(1) $lim_{P\to\infty}\ t_P^2 e^{\theta_2 t_P} \mathrm{E}[\,A(t_P, P_1)] = 0.$
(2) $lim_{P\to\infty}\ t_P e^{\theta_2 t_P} \mathrm{E}[T(t_P, P_1) - t_P P_1] = 0.$
(3) $lim_{P\to\infty}\ e^{\theta_2 t_P}\ Var[T(t_P, P_1)] = 0.$
(4) $lim_{P\to\infty}\ e^{\theta_2 t_P/2} \mathrm{E}[\,M(t_P, P_1) - t_P] = 0.$

Propositions 5.3 and 5.4 state that if the number of preselected processors $P_1$ is not too large with respect to the computer time $t_P$, then the expected number of processors that must continue simulating beyond time $t_P$ is very close to zero. Thus the total unplanned computing time is also extremely close to zero, as is the expected waiting time for the last processor to complete simulating.

## 6. SUMMARY

This paper has analyzed properties associated with a simple yet effective way to exploit parallel processors in discrete event simulations: averaging the results of multiple, independent replications that are run in parallel on multiple processors. We assumed that there is a CPU time constraint $t$ on each of $P$ processors. However, we showed that, unless the replication lengths are bounded, one must be willing to simulate beyond any fixed, finite time $t$ on at least some processors in order to always get the right answer. The statistical properties of a variety of estimators were then explored. Limit theorems were obtained for these estimators when either the number of processors or the CPU time constraint approaches infinity. In addition, central limit theorems and bias expansions were obtained when both of these parameters simultaneously get large. In this case, relative growth rates for $P$ and $t$ were determined in order for the estimators to have properly centered central limit theorems. For example, if one insists on never simulating beyond time $t$ (and using the estimator $\bar{\mu}(P, t)$), then $P$ must grow rather slowly with respect to $t$. On the other hand, one can preselect $P_1$ $(0 < P_1 \leq P)$ of the processors and simulate beyond time $t$ on a preselected processor if and only if no replications have yet been completed on that processor. This results in the unbiased estimator $\tilde{\mu}_2(P, P_1, t)$. While one can preselect an asymptotically negligible number of processors (i.e., $P_1/P \to 0$), this places restrictions on the relative values of $t$ and $P_1/P$ (to obtain the desired central limit theorem, as described in Proposition 3.7). However, in practical applications, since the appropriate moment conditions may be difficult to identify and the constants subsumed in the O( ) notation are unknown, correct implementation of this sampling plan may be a problem.

A sensible, practical approach is to preselect a fixed fraction $\alpha$ of the processors. While there is some variance inflation for large $P$ and finite values of $t$ (as opposed to preselecting all the processors), this inflation will be modest provided that $t$ is not too small with respect to the distribution of $\tau_{ij}$ and $\alpha$ is not too close to 0. As $t \to \infty$, there is no (asymptotic) variance

inflation. In addition, for large values of $t$, provided $P_1$ is not too large, $(T(t, P_1) - P_1 t)$, the total amount of computer time used beyond time $t$, is negligible in the sense that $\mathrm{Var}[T(t, P_1)] \to 0$ quite rapidly. The waiting time for the last processor to complete simulating, $(M(t, P_1) - t)$, is also negligible in the sense that $\mathrm{E}[M(t, P_1) - t] \to 0$ very rapidly.

REFERENCES


1. Bhavsar, B. C., and Isaac, J. R. Design and analysis of parallel Monte Carlo algorithms. *SIAM J. Sci. Stat. Comput. 8* (1987), s73-s95.
2. Billingsley, P. *Convergence of Probability Measures* Wiley, New York, 1968.
3 Chow, Y. S., Hsiung, C. A , and Lai, T. L. Extended renewal theory and moment convergence in Anscombe's theorem. *Ann Probability 7* (1979), 304-318.
4. Chung, K L. *A Course in Probability Theory* 2nd ed., Academic Press, New York, 1974.
5 Cox, D. R. *Renewal Theory*. Methuen, London, 1962.
6. Cramér, H. *Mathematical Methods of Statistics* Princeton University Press, Princeton, N.J., 1946.
7. Crane, M. A , and Iglehart, D L. Simulating stable stochastic systems, III: Regenerative processes and discrete event simulations. *Oper. Res. 23* (1975), 33-45
8. David, H A. *Order Statistics*. 2nd ed , Wiley, New York, 1981.
9. Feller, W. *An Introduction to Probability Theory and Its Applications, Vol. I* 3rd ed., Wiley, New York, 1968.
10. Fujimoto, R. M. Time warp on a shared memory multiprocessor In *Proceedings of the 1989 International Conference on Parallel Processing III*. The Pennsylvania State University Press, 1989, 242-249.
11. Fujimoto, R. M. Parallel discrete event simulation. *Commun. ACM 33*, 10 (Oct. 1990), 30-53.
12. Glynn, P W A low bias steady-state estimator for equilibrium processes. Tech., Rep. Dept. of Industrial Engineering, Univ. of Wisconsin, Madison, 1987
13. Glynn, P. W , and Heidelberger, P. Analysis of initial transient deletion for replicated steady-state simulations IBM Res Rep. RC 15259. Yorktown Heights, N.Y , 1989.
14. Glynn, P. W., and Heidelberger, P Analysis of initial transient deletion for parallel steady-state simulations. IBM Res. Rep. RC 15260. Yorktown Heights, N.Y , 1989.
15. Glynn, P. W , and Heidelberger, P. Jackknifing under a budget constraint. IBM Res. Rep RC 15261 Yorktown Heights, N.Y , 1989.
16. Glynn, P. W., and Heidelberger, P. Bias properties of budget constrained simulations *Oper. Res. 38* (1990), 801-814.
17. Glynn, P. W., and Heidelberger, P Experiments with initial transient deletion for parallel, replicated steady-state simulations. IBM Res Rep RC 15770 Yorktown Heights, N Y., 1990.
18. Goli, P., Heidelberger, P., Towsley, D., and Yu, Q Processor assignment and synchronization in parallel simulation of multistage interconnection networks. In *Distributed Simulation*. D. Nicol, Ed. The Society for Computer Simulation International, 1990, 181-187
19. Heidelberger, P. Statistical analysis of parallel simulations In *1986 Winter Simulation Conference Proceedings* J Wilson and J Henriksen, Eds. IEEE Press, New York, 1986, 290-295.
20. Heidelberger, P Discrete event simulations and parallel processing: Statistical properties *SIAM J. Sci Stat. Comput. 9* (1988), 1114-1132.
21 Karlin, S., and Taylor, H. M. *A First Course in Stochastic Processes*. 2nd ed., Academic Press, New York, 1975.
22. Kremers, W K Completeness and unbiased estimation for sum-quota sampling *J Am. Stat. Assoc. 81* (1986), 1070-1073.
23 Leadbetter, M. R., Lindgren, G , and Rootzén, H . *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag, New York, 1983.
24 Lehmann, E. L. *Theory of Point Estimation*. Wiley, New York, 1983.

25. LUBACHEVSKY, B. D. Efficient distributed event-driven simulations of multiple-loop networks. *Commun. ACM 32* (1989), 111-123.
26. MEKETON, M. S., AND HEIDELBERGER, P. A renewal theoretic approach to bias reduction in regenerative simulations. *Manage. Sci. 28* (1982), 173-181.
27. MILLER, R. G. The jackknife—A review. *Biometrika 61* (1974), 1-15.
28. MISRA, J. Distributed discrete event simulation. *Comput. Surv. 18* (1986), 39-65.
29. NICOL, D. M. Parallel discrete-event simulation of FCFS stochastic queueing networks. In *Proceedings of the ACM/SIGPLAN PPEALS 1988. Parallel Programming: Experience with Applications, Languages and Systems*. ACM, New York, 1988, 124-137.
30. NICOL, D., ED. *Distributed Simulation*. Simulation Series 22, No. 2. The Society for Computer Simulation International, San Diego, Calif., 1990.
31. PATHAK, P. K. Unbiased estimation in fixed cost sequential sampling schemes. *Ann. Stat. 4* (1976), 1012-1017.
32. RIGHTER, R., AND WALRAND, J. C. Distributed simulation of discrete event systems. *Proc. IEEE 77* (1989), 99-113.
33. ROSS, S. M. *Stochastic Processes*. Wiley, New York, 1983.
34. SMITH, W. L. Regenerative stochastic processes. *Proc. Roy. Soc. A. 232* (1955), 6-31.
35. SMITH, W. L. Renewal theory and its ramifications. *J. Roy. Stat. Soc. B. 20* (1958), 243-302.
36. UNGER, B., AND FUJIMOTO, R., EDS. *Distributed Simulation, 1989*. Simulation Series 21, No. 2. The Society for Computer Simulation International, San Diego, Calif., 1989.
37. UNGER, B., AND JEFFERSON, D., EDS. *Distributed Simulation, 1988*. Simulation Series 19, No. 3. The Society for Computer Simulation International, San Diego, Calif., 1988.
38. WHITT, W. Planning queueing simulations. *Manage. Sci. 35* (1989), 1341-1366.
39. YU, Q., TOWSLEY, D., AND HEIDELBERGER, P. 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, Calif , 1989, 191-196.

Received February 1990; revised October 1990; accepted October 1990