TOWARDS CREDIBLE AND EFFICIENT NETWORK SIMULATION EXPERIMENTS

E. Mota*, F. Fitzek, A. Wolisz
Technical University of Berlin, Germany
K. Pawlikowski
University of Canterbury, Christchurch, New Zealand

Published in the Proceedings of High Performance Computing 2000 - Grand Challenges in Computer Simulation, Washington D.C., April $17^{th} - 20^{th}$ 2000

ABSTRACT

Despite of even higher performance computers, quantitative steady-state simulation of even a moderate complex system takes very long time. There is always a necessity of efficient procedures that can control the run length of stochastic simulation experiments. This paper compares two sequential batching-based procedures for estimating steady-state mean under Multiple Replications in Parallel. A second motivation to this investigation is that fixed-sample size procedures can yield inappropriate precision of the results. By using well designed sequential analysis procedures for analyzing data carefully, one can guarantee a better quality of the conclusions drawn from the analysis. A practical application illustrates the feasibility of these procedures.

1 INTRODUCTION

Stochastic simulation is extremely useful in system performance evaluation. Since any statistic from such experiments cannot be guaranteed to give a close estimate for every sample, we must design statistics that will give good results on the average or in the long run. Despite of even higher performance computers, quantitative steady-state simulation of even a moderate complex system takes very long time, since to obtain reasonably stable results, that is, estimates with reasonably variances, very large samples are usually necessary.

Natural effort toward the reduction of sample size have been mainly concerned with application of parallel processing enhancements. However, more computing power cannot replace the need for reliable statistical methods of analysis of the output time series of observations arising from simulation experiments.

To control the precision of steady-state estimators, the final estimate of an analyzed parameter should be determined with its confidence interval. We are interested in methods for the automatic generation of confidence intervals, under *Multiple Replications in Parallel* – MRIP scenario, of pre-specified precision, by controlling the length of a single run, in such a way that a wide population of experimenters who have little knowledge or interest in simulation output analysis can take advantage of this enterprise.

The relative width of an estimated confidence interval can be controlled by the use of an appropriate sequential stopping rule This paper will compare two sequential techniques for estimating steady-state mean under Multiple Replications in Parallel.

2 LOOKING AT THE DATA SEQUENTIALLY

Typically, the run length of a stochastic simulation experiment is determined either by assigning the amount of simulation time before initiating the experiment or by letting the simulation run until a prescribed condition occurs. The first approach is known as fixed-sample size procedure, and suffers from the possibility of inappropriate precision of the results.The second approach is generally known as sequential and is the subject of this work.

*Partially supported by ISAE/Fundação Getulio Vargas

Sequential procedures gather observations to investigate a certain parameter of interest, and a decision has to be made to stop the sampling if a predefined condition is achieved, or to continue the sampling and periodically repeat both steps above while necessary. It is evident that the number of observations required to terminate the experiment is a random variable since it depends on observations. One can see that a sequential procedure can be economical in the sense that we may reach a decision earlier compared to fixed-sample-sized experiments, but can be onerous if one wishes a tight precision.

The importance of sequential procedure is widely recognized as the only effective method for controlling the precision of simulation results. Two important issues are : (i) the method of analysis of precision, and (ii) determination of checkpoints.

In order to assess the error done by estimating $\bar{X}$, a steady-state estimator of the mean μ of a performance parameter characterizing a system, one constructs a confidence interval given by $[\bar{X} \pm H]$, where H is the half-width of the confidence interval taken at an assumed confidence level $(1-\alpha)$. However, $\{X_i : i = 1, ...n\}$ are usually positive correlated and, thus, classical statistics can not be applied directly.

Positive correlation denotes negative bias and, thus, the final confidence interval half-width can be underestimated which can lead frequently to a real (experimental) confidence level much smaller than the nominal theoretical confidence level.

Several confidence interval procedures (CIPs) have been suggested in the literature to get valid confidence intervals from simulation output data. Some CIPs try to reduce autocorrelation of output data by means of grouping the observations (Batch Means approach), other take into account the correlation for the computation of variance necessary to construct the confidence interval (Spectral Analysis approach). The quality of the first type of CIPs and their applicability in parallel multiprocessors environment such that of MRIP will be discussed in the next sections.

3 BATCHING TECHNIQUES

Methods based on *Batch Means* (**BM**) are popular and very frequently applied in simulation practice. We have investigated their performance and applicability under *Multiple Replications In Parallel* (**MRIP**), by means of some sequential versions of **BM** that can be applied concurrently on workstations connected via a network (Mota et al. 1999). Considering that each copy of the model is initiated using an independent, non-overlapping parallel time-stream of random numbers, and generates output data independently from others.

MRIP takes into account that generating data independently, and using an asynchronous communication among the processors (to diminish the possibility of dead-locks) one can get a sound speedup by reducing the time needed for generation of the number of output data constituted statistically representative sample. Theoretical justification can be found in (Raj and Khamis 1958) and (Raj 1968), who showed that in sampling with replacement the average over distinct units possesses lower variance than the average over the entire sample including repetitions. Moreover, by using well designed sequential analysis procedures for analyzing data carefully, one can also guarantee a better quality of the conclusions drawn from the analysis.

3.1 Nonoverlapping Batch Means - NOBM

The well-known and exhaustively analyzed classical approach is to divide a series of steady-state observations of length N into B adjacent nonoverlapping batches of size M (N=M.B), and test correlation coefficients of lag k (k=1, ..., N/10) against correlation. Refer to (Pawlikowski 1990) for a thorough treatment of this and related issues). Each failure in the test leads to a choice of a larger batch size and repetition of the test. Having found the "optimal" batch size, observations are reorganized into few batches ($10 \leq B \leq 30$), and new batches of observations are collected. At predefined checkpoints, each replication generates an estimate of the parameter being analyzed and send them to a global analyzer responsible for stopping the simulation when the desired precision is achieved.

3.2 Overlapping Batch Means - OBM

(Meketon and Schmeiser 1982) proposed making a better use of the collected data. Namely, they proposed that after collecting N=M.B observations, one computes so many batch means as possible, each observation initiating a new (overlapped) batch. To save storage, one can maintain in memory just the last batch. As a new observation arrives it is appended at the end of the batch while the first observation of the batch is deleted. A counter advises when the degree of overlapping is achieved, which means that a batch mean has to be calculated. The procedure goes on until the stopping rule is met.

In light of that, one can perceive that OBM is more trickier than NOBM, as the higher the number of degree of overlapping, the longer the amount of computation of batch means is needed. On the other hand, the higher the degree of overlapping, the more values are used in the estimation. Statistically, it implies lower variance and, better coverage, the main criterion of CIP performance comparison.

4 PERFORMANCE ISSUES

In order to compare the performance of different sequential CIPs, one should consider a confidence region at an assumed confidence level η $R(\eta, \mathbf{X})$, associated with each CIP. Such region is obtained by making some assumptions about the random properties of the data. If the assumptions are satisfied, the region $R(\eta, \mathbf{X})$ contains the unknown estimated parameter μ with probability

$$P(\mu \in R(\eta, \mathbf{X})) = \eta$$

When the procedure is correct, the observed coverage is equal to the desired confidence level. In robustness studies one is interested in situations where assumptions fail to hold $F_{\eta^*} \neq \eta$

To avoid a *hit-or-miss* analysis, Schruben suggests the construction of a coverage function for a range of confidence levels and not only for a single value. We have applied the sequential coverage analysis proposed by (Pawlikowski et al. 1998), and studied the coverage function, besides the average number of observations needed for stopping the simulation, E[O], and the variability of the final confidence intervals given by the coefficient of variation of its half-width CoV{H}, to infer the performance of the CIPs under MRIP

4.1 Degree of overlapping

Periodically, a sequential procedure collects estimates to check whether the stopping rule has been satisfied. Let N observations are available, distributed in B contiguous, nonoverlapped batches of size M (N=M.B), be available at a given checkpoint.

If one divides each non-overlapped batch into two equally sized parts, each one with M/2 observations, in such a way that each part initiates a new (overlapped) batch of size M, one can form 2(B-1)+1 overlapped batches. By dividing each batch in four equal sized parts one obtains 4(B-1)+1 batches, and so on (Table 1).

Degree of overlap	Number of batches	degrees of freedom
M/2	2(B-1)+1	1.33
M/4	4(B-1)+1	1.45
M/8	4(B-1)+1	1.48
...	...	...
M/M	M(B-1)+1	1.50

Table 1: Degrees of freedom for complete and partial overlapping

The last case (M/M) is known as *complete overlap*, that is each observation begins a new (overlapped) batch. (Meketon and Schmeiser 1984) suggested an asymptotic number of degree of freedom equal $1.5(B-1)$. The other cases are generally known as *partial overlap* (Welch 1987). However, as pointed out in (Sargent et al. 1992), to get a 1/3 variance reduction (in comparison with the variance of the classical NOBM), Meketon and Schmeiser let $b \to \infty$, which means that probably those differences in degrees of freedom are practically irrelevant in terms of sequential procedure, especially for very high levels of traffic intensity.

We have simulated an M/M/1 queue to estimate the mean waiting time of a customer, when the system load is $\rho = 95\%$. A confidence level was constructed at a confidence level $\eta = 95\%$. Results for P=1 shown in Table 2 [1] supports this statement. If we look at E[O], one could say that Welch's finding is still valid to a sequential analysis of the output simulation data, as with lower degree of overlapping one can get a reasonable variance reduction, The same we can say about the variability of the simulation results measured by CoV{H}.

Degree of overlap	ρ	η	cov±H	E[O]	CoV{H}
M/2	95	95	94.9±1.1	2145911	0.0257
M/4	95	95	92.8±1.4	2129494	0.0270
M/8	95	95	94.0±1.2	2131604	0.0268
M/M	95	95	94.9±1.0	2212278	0.0255

Table 2: OBM performance according to different degrees of overlapping

[1] In this experiment we collected 150 bad confidence intervals to initiate the sequential coverage analysis. See (Pawlikowski 1998)

4.2 Degree of parallelization

Since the batch size selection phase takes very long time, once can think of taking advantage of the parallel power of networked computers and relax to some extent the precision requirements and compensating it by adding more processors (simulation engines), for reducing the length of this phase.

We have investigated this idea with both sequential procedures, (**OBM** and **NOBM**). Instead of calculating the correlations coefficients for lags 1 to L, where L is 10% of the sample size, as recommended in (Pawlikowski 1990), we calculated just correlations of lag 1. Moreover, we used a smaller number of batches for testing correlations. This should degrade the coverage of results if simulation is run on a single processor.

Indeed, coverage of **OBM** has now a worst value, but as we add more processors it improves. The same cannot be said concerning **NOBM**, as the improvement obtained by adding more processors was not enough to yield valid coverage, which means that this method is not suitable to be run under MRIP when traffic intensity is very high. One can conclude that MRIP itself can be a variance reductor and compensates certain lack of fullfilment of assumptions of the methods of analysis. Fig. 1 depicts our empirical results.

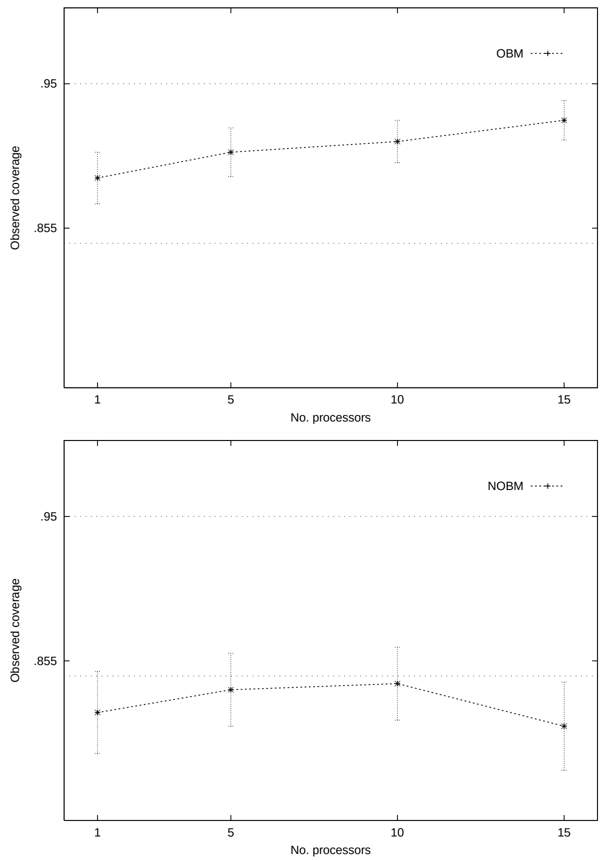

Fig. 1: Compensating batch size imprecision by higher degree of parallelization

5 SUITABILITY TO PARALLEL PROCESSING

Under **MRIP**, another criterion of fundamental importance is the relationship between the length of preprocessing of output data, required by a given CIP and the total length of the simulation. By preprocessing we mean the extension of preliminary tasks carried out by the procedure until it begins delivering estimates to the central analyzer. In terms of Amdahl's law, it constitutes the fraction of the simulation that cannot be parallelized.

Concerning all batching-methods, the preprocessing corresponds to the batch size determination (BSD) phase, and the length of transient phase. Both, the length of BSD phase and the total run length N are conveniently measured in terms of number of observations, as the workstations are being shared by other applications and measures in terms of time can be somewhat deceiving.

It is worthwhile to emphasize that, as we used a very highly loaded ($\rho = 95\%$) M/M/1 queuing system and the stopping rule was a confidence interval at 95% of confidence level with a 5-%relative precision, BSD is very long and that's why **NOBM** offers so poor performance, while the actual implementation of **OBM** requires fewer observations to stopping the simulation, the relation BSD/N suggests that this method is promising under **MRIP**. When the underlying requirements are not completely fullfiled, one can try to reduce BSD to improve even more its suitability.

6 A COMPARATIVE CASE STUDY

After a careful investigation of the above sequential procedures on different queuing systems under increasing traffic intensity and high confidence level, we decided to assess the MRIP in practice, by means

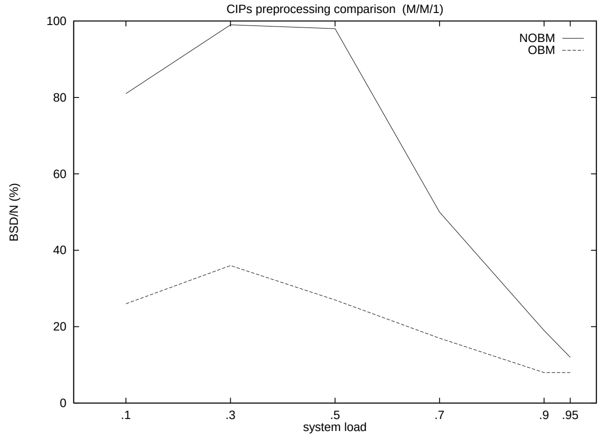

Fig. 2: CIPs under not so very highly loaded system

of a simulation of a real time-consuming network problem.

We consider a CDMA based mobile communication system with a specific number of **W**ireless **T**erminals *(WTs)*. All WTs communicate with one central **B**ase **S**tation *(BS)*, which coverage defines the cell boundaries (see figure 3).

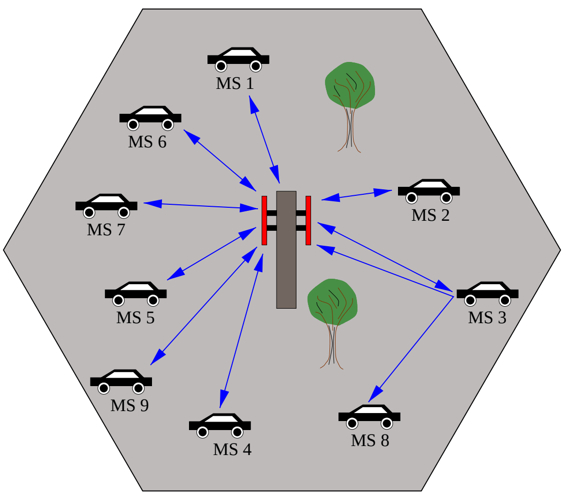

Fig. 3: Nine Wireless Terminals Communicating With One Base Station

The mobile communication system supports a number of codes much higher than the number of active WTs. All WTs are sending asynchronous as well on bit level as on chip level. The wireless link is considered to be unreliable with a varying ***Bit Error Probability*** *(BEP)*. The value of the BEP depends on the number of active channels k. For the chosen scenario we assume an **A**dditive **W**hite **G**aussian **N**oise (AWGN) channel with **B**inary **P**hase **S**hift **K**eying (BPSK). Codes will be assigned before the connection is established. The total number of codes per mobile is set by the QoS requirements of the mobile.

The simulations have been performed using the `Ptolemy` simulation tool and full parallelization, statistical evaluation and run length control we have used Akaroa with the `Ptolemy` interface akstars (see 4). It is worthwhile to say, that no additional effort was required from the analyst, and we could say that the framework can be still considered transparent from the user point of view.

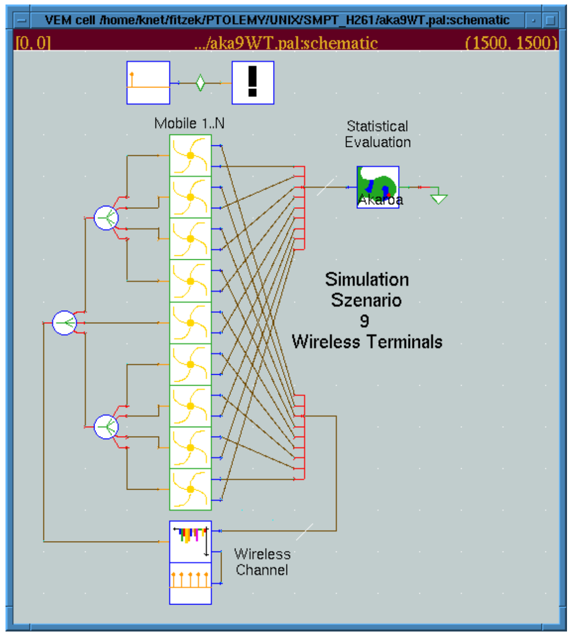

Fig. 4: Ptolemy Structure

We formed a communication system with 9 WTs and one BS. Outer-Cell interference was not taken under consideration. The channel between the WT and the BS is modeled with a multilayered Markov chain, considering two channel states (*bad* and *good*) and the impact of used channels on the BEP on bit level. The main parts of the simulation model are the protocol implementation of the WT and BS.

We investigated the influenced jitter with high accuracy. Because of the nature of the wireless link, which is influenced by the user's mobility and the active mobiles using the wireless channel, is very difficult to predict how many observations values are necessary. To overcome this problem we use the sequential analysis proposed in section 3. Moreover the simulation of the wireless channel on bit level is very time consuming. Therefore a fast simulation strategy is desirable.

7 SUMMARY AND FINAL CONCLUSIONS

We wanted to assess accurate estimates of the jitter. Running the described simulation model using only one processor, the simulation took 188 hours, and by using 10 homogeneous processors under Akaroa-2, an MRIP implementation developed at the University of Canterbury, Christchurch, New Zealand, the run length dropped to 10 hours, a considerable speedup considering that no additional modification in the model was required. All experiments required no special priorities and ran together with other applications at the Telecommunication Network Group of the Technical University of Berlin..

Reliability of the results can be based on our empirical investigations on different queuing systems, which are commonly used to model communication systems. We have adopted the complete overlapping because it has been proved to be more efficient besides relieving the user of giving another parameter for the statistical method of analysis.

References

[1] Frank H.P. Fitzek, A.Wolisz,"QoS Support In Wireless Networks Using Simultaneous MAC Packet Transmission (SMP T)",Advanced Simulation Technologies Conference (ASTC 1999),April,1999

[2] Heidelberger, P., *Discrete event simulations and parallel processing: statistical properties*, SIAM Journal Stat. Comput., vol.9, pp.1114–1132, November 1988.

[3] ETSI TC-RES,"Radio Equipment and Systems (RES); High Performance Radio Local Area Network (HIPERLAN); Type 1",Functional Specification,France,Dec.,1996

[4] McNickle, D., Pawlikowski, K. and Ewing, G., *Experimental evaluation of confidence interval procedures in sequential steady-state simulation*, Proc. 1996 Winter Simulation Conference, pp. 382—389, 1996.

[5] Meketon, M.S. and Schmeiser, B., *Overlapping batch means: something for nothing ?*, Proc. 1984 Winter Simulation Conference, 227–230, 1984.

[6] Mota, E., Wolisz, A. and Pawlikowski. K., *Sequential batch means techniques for mean value analysis in distributed simulation*, Proc. 13th European Simulation Multiconference Warsaw, pp. 129–134, 1999.

[7] Pawlikowski, K., *Steady-state simulation of queueing processes: a survey of problems and solutions*, ACM Computing Surveys, vol.22, pp. 123–170, 1990.

[8] Pawlikowski, K., Yau, V. and McNickle, D., *Distributed stochastic discrete-event simulation in parallel time streams*, Proc. of the 1994 Winter Simulation Conference, pp. 723–730, 1994.

[9] Pawlikowski, K., McNickle, D. and Ewing. G., *Coverage of confidence intervals in steady-state simulation*, Journal Simulation Practice and Theory, 6(1998), pp. 255–267.

[10] Raj, Des and Khamis, Salem H., *Some remarks on sampling with replacement*, Ann. Math. Statistics, Vol.29, pp. 550–557, 1958.

[11] Raj, Des, Sampling Theory. McGraw-Hill, New york, 1968.

[12] Sargent, R.G.m Kang, K. and Goldsman, D., *An investigation of finite- sample behavior of confidence interval estimators*, Operations Research Vol. 40, No. 5, pp. 898–913, September-October 1992.

[13] Schmeiser, B., *Batch size effects in the analysis of simulation output*, Operations Research, Vol.30, no.3, 556–598, May-June 1982.

[14] Schruben, L., *A coverage function for interval estimators of simulation response*, Management Science 26, 18–27, 1980.

[15] Schruben, L.W., *Detecting initialization bias in simulation output*, Oper. Res., Vol.30, pp. 569–590, 1982.

[16] Welch, P., *On the relatioship between batch means and overlapping batch means*, Proceedings of the 1987 Winter Simulation Conference, 320–323, 1987.