# A PERSPECTIVE OF BATCHING METHODS IN A SIMULATION ENVIRONMENT OF MULTIPLE REPLICATIONS IN PARALLEL

Edjair Mota
Department of Computer Science
University of Amazonas, Brazil

Adam Wolisz
Telecommunication Networks Group
Technical University of Berlin, Germany

Krzysztof Pawlikowski
Department of Computer Science
University of Canterbury, Christchuch, New Zealand

## ABSTRACT

Discrete event simulation is frequently time-consuming either because modern dynamic systems, such as telecommunication networks, are becoming increasingly complex and/or a great number of observations is required to yield reasonably accurate results. An interesting approach to reduce the time duration of simulation is that of concurrently running multiple replications in parallel (MRIP) on a number of processors connected via networking and averaging the results adequately. We present the results of our research on the suitability of batch-means-based procedures in such distributed stochastic simulation.

## 1 INTRODUCTION

Analysis of output data from steady-state discrete-event simulation has attracted a considerable attention. A sound methodology can be found in the literature (refer to Pawlikowski (1990) for a thorough review of problems and solutions survey), but there are still open questions that deserve attention, especially concerning the quality of the results produced by a plethora of methods proposed for simulation output data analysis.

Estimation of results during steady-state simulation is a nontrivial problem because the output data are neither independent nor identically distributed, but usually highly correlated. In order to apply the classical statistical analysis, one can (for example) group observations in the output sequence in such a way that means over groups of observations, are nearly uncorrelated. Such transformation of original output sequences is applied in a class of methods of simulation output analysis known as Batch Means.

*Batch Means* techniques are widely used due their conceptual simplicity and intuitive implementation. It is based on the concept that if batches are large enough then their mean values can be practically uncorrelated. The main challenge of this method is the determination of batch size; see e.g. Schmeiser (1992).

The classical nonoverlapping batch means estimator (NOBM) is constructed by dividing a sequence $\{X_i\}$ of $n$ steady-state observations, into $b$ contiguous batches of size $m$. Provided that correlations among batch means can be considered negligible at a certain significance level $\beta$, the interval estimator for a performance parameter $\mu$ is given as

$$P(\bar{X}(n) - H \leq \mu \leq \bar{X}(n) + H) = 1 - \alpha \qquad (1)$$

where $\bar{X}$ is the arithmetic mean of the sample of size n, and H is the half-length of the confidence interval defined as

$$H = t_{df,1-\frac{\alpha}{2}}\hat{\sigma}[\bar{X}]. \qquad (2)$$

$\hat{\sigma}[\bar{X}(n)]$ is the standard deviation of the batch means estimator of $\mu$, and $t_{df,1-\frac{\alpha}{2}}$ is the upper $(1 - \frac{\alpha}{2})$ critical point of the t distribution with $df$ degrees of freedom.

In NOBM, df equals $b$-1. In this paper, we discuss three additional sequential methods based on Batch Methods that can be run under MRIP, and use estimators with more degrees of freedom than in the case of NOBM. This suggest that they could require shorter simulations to get final results with an acceptable statistical error.

In the following sections, our discussion is based on the results from sequential stochastic simulation of an $M/M/1/\infty$, in which the mean waiting time was estimated, with the relative statistical error not greater than 5%, at 95% of confidence level.

The main performance criterion for a confidence interval procedure is its coverage, defined as the frequency of the confidence intervals containing the true parameter $\mu$. Usually, the coverage is assessed on a fixed-sample basis, but we are of the opinion that coverage analysis should be analysed sequentially. Therefore, we have applied the sequential coverage analysis proposed in Pawlikowski et al. (1998). Following this approach, the coverage analysis begins when a minimum number

(we assumed 200) of bad confidence intervals, i.e. confidence intervals that do not contain theoretical value, are collected, and results from too short runs are discarded.; It stops when the relative precision of the confidence intervals for the coverage is less than 5%.

## 2 SEQUENTIAL ANALYSIS

Sequential procedures of simulation output data analysis are widely recognized as the only effective techniques for controlling the final precision of simulation results. In the sequential procedures analysed here we apply sequential versions of batch-means-based techniques together with sequential test for detecting the length of the initialtransient period, proposed by Schruben (1982).

In NOBM, in its version presented in Pawlikowski (1990), for reducing dependence between batch means, correlation coefficients for lag k (k=1, ..., $n$) are tested sequentially. If all these correlations cannot be considered negligible at a significance level $\beta = 0.01$, more observations are collected and the test is repeated.

To improve the quality of this test, we applied a nonparametric method based on jackknife estimators, for calculating the correlation coefficients.

When independence test succeeds, observations are grouped into 25 batches according to Schmeiser's results (1982), and estimation phase initiates. From now on, when NOBM collects a batch of size $m^*$, the optimal batch size found in the previous phase, a checkpoint is reached and, if desired relative precision is detected, simulation stops. This sequential method behaves acceptably when system load $\rho < 90\%$, but as load increases its quality worsens (see Table 1).

This asymptotic failure was expected by Glynn and Whitt (1991), when they showed that "there is no variance estimator based on a fixed number of batch that is consistent". In light of that, we introduced a variant of NOBM to get rid of this *anomaly*, and called it NOBM/GW : at each checkpoint, when the relative precision is not achieved, the number of batches increases by 2. This increment is enough to keep in touch with Schmeiser's findings and, at the same time, to guarantee a better asymptotic behavior. Table 1 shows that NOBM/GW indeed improves somewhat the coverage in very high-loaded systems, when compared to classical NOBM.

As an attempt to weaken the strong correlations, we also designed and implemented a sequential version of a fixed-sample size technique proposed by Fox et al. (1991), in which one discards some observations between consecutive batches. This is the so-called *Spaced Batch Means* (SBM) technique. As in NOBM, the

| NOBM | | | |
|---|---|---|---|
| $\rho$ (%) | cov $\pm H$ | E[O] | CoV{H} |
| 91 | 91.10 ± 1.4 | 545.27 | 0.0390 |
| 92 | 88.90 ± 1.7 | 661.25 | 0.0393 |
| 93 | 87.60 ± 1.9 | 829.99 | 0.0398 |
| 94 | 86.60 ± 2.0 | 1079.23 | 0.0407 |
| 95 | 86.00 ± 2.2 | 1474.18 | 0.0381 |

| NOBM/GW | | | |
|---|---|---|---|
| $\rho$ (%) | cov $\pm H$ | E[O] | CoV{H} |
| 91 | 90.06 ± 1.5 | 557.78 | 0.0297 |
| 92 | 88.80 ± 1.7 | 674.15 | 0.0307 |
| 93 | 88.60 ± 1.7 | 851.09 | 0.0313 |
| 94 | 88.40 ± 1.8 | 1095.49 | 0.0320 |
| 95 | 86.50 ± 2.0 | 1501.23 | 0.0339 |

| SBM | | | |
|---|---|---|---|
| $\rho$ (%) | cov $\pm H$ | E[O] | CoV{H} |
| 91 | 91.20 ± 1.4 | 559.37 | 0.0294 |
| 92 | 89.90 ± 1.5 | 677.01 | 0.0308 |
| 93 | 89.20 ± 1.6 | 858.13 | 0.0313 |
| 94 | 88.10 ± 1.8 | 1101.58 | 0.0318 |
| 95 | 87.60 ± 1.9 | 1505.77 | 0.0335 |

| OBM | | | |
|---|---|---|---|
| $\rho$ (%) | cov $\pm H$ | E[O] | CoV{H} |
| 91 | 94.90 ± 0.8 | 747.76 | 0.0252 |
| 92 | 94.50 ± 0.9 | 925.23 | 0.0257 |
| 93 | 95.20 ± 0.8 | 1191.40 | 0.0254 |
| 94 | 94.70 ± 0.9 | 1573.26 | 0.0259 |
| 95 | 94.20 ± 1.0 | 2210.58 | 0.0258 |

Table 1: Performance of Batching Methods under MRIP. cov $\pm H$ is a confidence interval for the coverage; E[O] is the average sample size, and CoV{H} is the coefficient of variation of the confidence interval half-length for the coverage.

number of batches increases in the estimation phase, if needed.

When the number **s** of discarded observations is **s**=0, we have the classical NOBM. Theoretically, the greater **s**, the better should be the coverage of final results, but that imposes an obvious problem of throwing out many observations. We have adopted a spacing equal 20% of the initial batch size but the best solution would probably be to determine **s** according to the properties of the underlying stochastic process. Table 1 shows that, in the reported cases, SBM is a bit better in the sense of coverage than NOBM and NOBM/GW.

Finally, we implemented a sequential version of *Overlapping Batch Means* (OBM), proposed by Meketon and Schmeiser (1984), which reuses some observations of one batch for constructing the next (overlapping) batch of observations. The original work suggested that after finding $m^*$, each observation should initiates a new (overlapped) batch. Of course, batch means are becoming strongly correlated but the number of batch means is much larger and that compensates the negative effect of correlation. Welch (1987), studied the possibility of achieving the same results with partial overlapping.

One can clearly see in Table 1 that, in terms of coverage, OBM behaves much better than the other methods. Its final confidence intervals are more stable and achieve almost perfect coverage for high values of $\rho$.

## 3   MULTIPLE REPLICATIONS

Research on speeding up execution of simulation models is a challenging problem which has attracted a considerable scientific interest and effort. A simple yet effective way to exploit computer networks for computationally intensive discrete-event simulation is to run multiple independent replications in parallel (Pawlikowski 1994), on multiple processors and to average the results appropriately. Of course, each replication should be run using different, independent sequence of (pseudo) random numbers. This way, by using P processors simulation output data can be generated P times faster. The only communication overhead of MRIP is associated with loading of the model into different processors (at the beginning of simulation) and with transmissions of data to a central analyser whenever a checkpoint is reached, to calculate global estimates. This approach is statistically efficient provided that the initialization bias is not severe (Heidelberger 1986).

MRIP approach offers a quite simple solution concerned with the credibility of the final simulation results. Refer to Pawlikowski et al. (1994), and Mota et al. (1999, 2000) for other investigation issues in these area.

## 4   ACHIEVABLE SPEEDUP

According to Amdahl's law (1967), if a fraction $f$ of a computation is inherently sequential, then the speedup S(P) is bounded above by

$$S(P) = \frac{1}{f + \frac{1-f}{P}} \qquad (3)$$

where P is the number of processors and $f$ is defined to be the ratio of the service demand of sequential parts of the computation to the service demand of the entire computation.

In steady-state simulation, considering that results are analyzed sequentially, MRIP imposes a limit to the average speedup that should be incorporated into above expression.

Let $N_{min}$ be the number of collected observations, sufficient for achieving the required precision of the final results. One could think of a situation in which there are so many processors employed that each one achieves just the first checkpoint only, and the stopping rule is reached. Let this number of processors be $P_{min}$.

Let $N_o$ be the length of the transient phase, $N_1$ be the amount of observations collected until the first checkpoint is achieved, and D be the distance between consecutive checkpoints (Figure 1). A truncated version of Amdahl's law for the MRIP scenario, formulated in (Pawlikowski and McNickle, 2000) states that:

$$S(P) = \begin{cases} \frac{1}{f + \frac{1-f}{P}} & if \;\; P \leq P_{min} \\[3mm] \frac{1}{f + \frac{1-f}{P_{min}}} & otherwise \end{cases} \qquad (4)$$
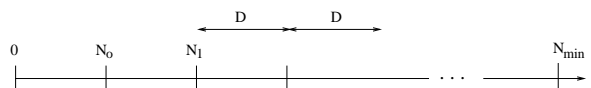


Figure 1: Steady-State Structure

In Batch Means techniques, observations collected during initial transient phase are not used by steady state estimates, thus:

$$f = \frac{N_0}{N_{min}}$$

Note that if D means the number of observations collected between consecutive checkpoints, then $P_{min}$ times D observations is needed to stop the simulation. That is

$$P_{min} = \frac{(1-f)N_{min}}{D} = \frac{N_{min} - N_0}{D} \qquad (5)$$

Considering the classical NOBM, NOBM/GW and SBM, D is at least equal to $m^*$,

Note that, OBM offers greater flexibility as D can be smaller than $m^*$. Namely, in the case of complete overlapping, the distance between checkpoints can be, theoretically, as short as 1. Thus, for $f$ remaining the same as in the previous methods, $P_{min}$ can be considerably greater in OBM.

# 5 EMPIRICAL INVESTIGATION

The following discussion will be based on the results obtained by AKAROA-2, an MRIP implementation developed at the University of Canterbury, New Zealand, and used at the Technical University of Berlin, Germany.

## 5.1 Speedup

To assess the average speedup obtained when the sequential batching techniques considered are applied in *Multiple Replications in Parallel* scenario, we simulated an $M/M/1/\infty$ queuing system, with traffic intensity 95%, and constructed a confidence interval at 95% of confidence level.

A specific feature of sequential techniques is that randomness of data collected at the output of the model being analyzed can fortuitously yield the stopping condition much earlier than it could be expected, and that can lead to wrong results. In light of that, we adopted a rule of thumb proposed by Ruth Lee et al. (1999). Namely, while using P processors we:

1. run the simulation experiment 3 times;

2. accepted the results produced by the longest simulation run only;

3. recorded the average length of the transient phase, measured by the number of transient observations $N_o$ discarded by each of P processors;

4. recorded the average number of observations $N_1$ required to achieve the first checkpoint;

5. recorded the average total number of observations $N_{min}$ when simulation was stopped.

Each time when these steps were followed, we obtained an n-uple $(N_o, N_1, N_{min})$. To improve the accuracy of the results we repeated the above sequence 100 times and averaged the results at the end. We repeated the whole experiment for P=2, 4, 6, 8, 10, 15 and 20 processors.

$P_{min}$, the number of processors that could still give a speedup, was calculated from the truncated Amdahl's law, using the results obtained from simulations on P=1 processors.

| D = $m^*$ | | | | | |
|---|---|---|---|---|---|
| P | $P_{min}$ | $N_o$ | $N_1$ | $N_{min}$ | S(P) |
| 1 | 832 | 626 | 85300 | 710110 | 1.000 |
| 2 | 832 | 707 | 78100 | 698875 | 1.998 |
| 4 | 832 | 602 | 76658 | 715507 | 3.989 |
| 8 | 832 | 651 | 64322 | 665697 | 7.946 |
| 10 | 832 | 648 | 63628 | 666770 | 9.913 |
| 15 | 832 | 641 | 62459 | 656010 | 14.798 |
| 20 | 732 | 638 | 63496 | 659464 | 19.639 |
| D = $m^*/10$ | | | | | |
| P | $P_{min}$ | $N_o$ | $N_1$ | $N_{min}$ | S(P) |
| 1 | 8655 | 629 | 81800 | 708609 | 1.000 |
| 2 | 8655 | 666 | 72550 | 709144 | 1.998 |
| 4 | 8655 | 617 | 78913 | 732444 | 3.990 |
| 8 | 8655 | 656 | 71787 | 725743 | 7.950 |
| 10 | 8655 | 647 | 67763 | 698836 | 9.917 |
| 15 | 8655 | 673 | 65203 | 682858 | 14.796 |
| 20 | 8655 | 639 | 66727 | 686178 | 19.652 |
| D = $m^*/50$ | | | | | |
| P | $P_{min}$ | $N_o$ | $N_1$ | $N_{min}$ | S(P) |
| 1 | 43551 | 617 | 81600 | 711375 | 1.000 |
| 2 | 43551 | 661 | 81650 | 722998 | 1.998 |
| 4 | 43551 | 647 | 77000 | 743142 | 3.990 |
| 8 | 43551 | 671 | 76001 | 749429 | 7.950 |
| 10 | 43551 | 660 | 69405 | 721601 | 9.918 |
| 15 | 43551 | 664 | 69427 | 708131 | 14.806 |
| 20 | 43551 | 635 | 60686 | 629966 | 19.624 |

## 5.2 Granularity

In AKAROA-2, each simulation engine collects a number of observations before calculating an estimate that is sent to a global analyzer. In the case of methods based on *batching*, one should wait until a number of batches are collected since only then an estimate can be obtained.

The complete overlapping version of **OBM** offers an attractive alternative in terms of batching methods under **MRIP**, as each new observation can be used to form a new overlapped batch, and an estimate can be obtained.

To investigate the complete overlapping version of **OBM**, by varying the distance D between checkpoints, we have simulated an $M/M/1/\infty$ queuing system, with traffic intensity 95%, stopping the simulation when the relative precision reached 5% or less. Figures 2, 3 and 4 summarizes this experiment.

Figure 2 shows the rate within which a sequential confidence interval procedure based on OBM converges to the desired relative precision. Clearly, this convergence becomes slower as the distance D between consecutive checkpoints decreases (D is measured in the
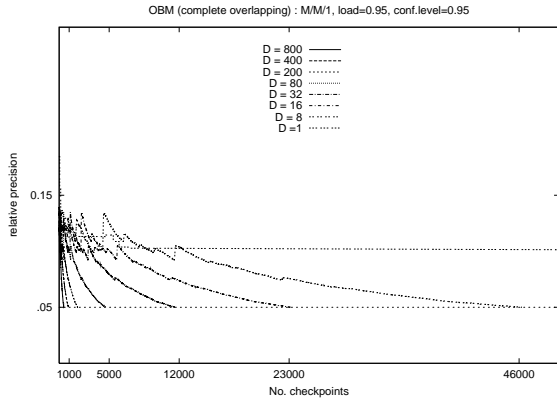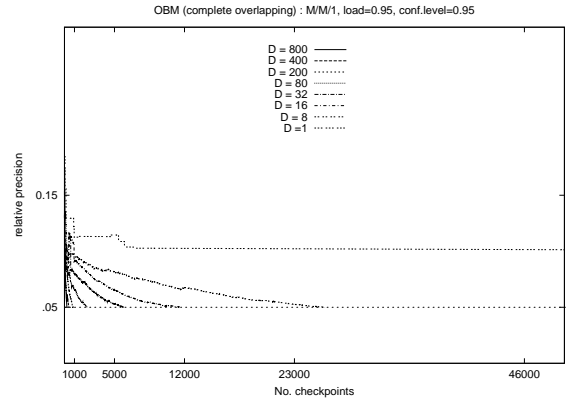
Figure 2: Convergence of Relative Precision : P = 1
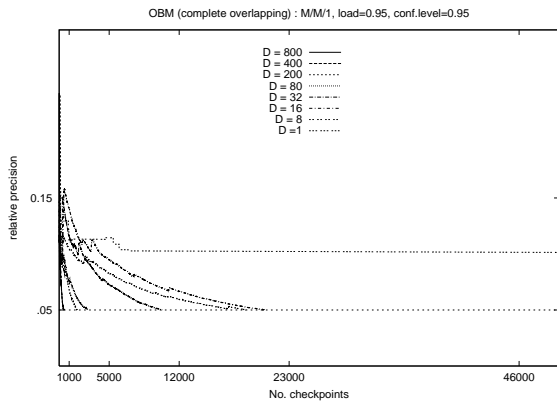


Figure 3: Convergence of Relative Precision : P = 2



Figure 4: Convergence of Relative Precision: P = 4

number of observations). Either the variance increases as we try to reduce the spacing between consecutive checkpoints, or the variance reduces very slowly. Figures 3 and 4 show the same effect but the degree of parallelization was increased, and convergence was somewhat faster than for P=2 and P=4, respectively.

## 6 FINAL REMARKS

Our investigation of methods based on the concept of Batch Means show that OBM is very robust in the sense that it yields confidence intervals with probability close to the nominal confidence level, especially when traffic intensity is very high. An analysis of the performance of this method applied to queuing networks can be found in Fitzek et al. (2000).

MRIP is a promising, though somewhat unpopular, approach for speeding up stochastic simulation experiments of complex dynamic systems. We have applied this methodology on simulation of modern telecommunication problems, and we have gotten very impressive results in terms of both speedup and quality of final results. The additional effort required from the analyst was minimal, since MRIP implementation in AKAROA.2 appeared to be very user-friendly. Currently, we are investigating other methods of simulation output data analysis, including those based on the concepts of *standardized time series*.

## REFERENCES

Fitzek, F. Mota, E., Ewers. E., Wolisz, A. and Pawlikowski, K., *An eeficient approach for speeding up simulation of wireless networks*, "Accepted to the 2001 Western Multiconference 2001.

Fox, B.L., Goldsman, D. and Swain, J.J., *Spaced Batch Means*, Operations Research, Vol.10, 255-263, 1991.

Glynn, P.W. and Whitt, W. *Estimating the Asymptotic Variance with batch means*, Operations Research Letters, 431–435, 1991.

Heildelberger,P., *Statistical analysis of parallel simulations*, Proc. 1986 Winter Simulation Conference, 290–295, 1986.

Lee, R., McNickle, D. abd Pawlikowski. K., *Sequential steady-state simulation : Rules of thumb for improving accuracy of the final results*, Proc. European Simulation Symposium, pp. 618–822, Erlangen 1999.

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

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.

E. Mota, F. Fitzek, A. Wolisz and K. Pawlikowski, *Increasing the accuracy of network simulation experiments*, Proc. 2000 Advanced Simulation Technologies Conference, pp. 356–361, Washington D.C., April 2000.

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

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.

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.

Pawlikowski, K., and McNickle, D. *Speeding up quantitative stochastic simulation". 2000. Submitted*

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

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

Welch, P., *On the Relationship Between Batch Means and Overlapping Batch Means*, Proceedings of the 1987 Winter Simulation Conference, 320-323, 1987.

## AUTHOR BIOGRAPHIES

**EDJAIR MOTA** is an assistant professor in the Department of Computer Science at University of Amazonas. He received a B.S. degree in Electrical Engineering from University of Amazonas; and M.S. in Computer Science from University of Paraiba, in Brazil. Actually, he is a Ph.D. candidate in Computer Science from Technical University of Berlin. His email is `<edjair@dcc.fua.br>`

**ADAM WOLISZ** is a chaired Professor of Electrical Engineering and Computer Science at the Technical University Berlin, where he is leading the Telecommunication Networks Group. He obtained his Dr.-Ing. degree in Computer Engineering) respectively in 1983 at the Silesian Technical University in Gliwice, Poland. His research interests are in architectures and protocols of communication networks as well as protocol engineering with impact on performance and Quality of Service aspects.

**KRZYS PAWLIKOWSKI** is an Assistant Professor of Computer Science at the University of Canterbury. His research interests include quantitative stochastic simulation, and performance modelling and evaluation of telecommunication networks. He received a Ph.D. in Computer Engineering from the Technical University of Gdanski, Poland, in 1975. He is a senior member of IEEE.