INCREASING THE ACCURACY OF NETWORK SIMULATION EXPERIMENTS

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

Keywords: Parallel simulation, multiple replications, batching techniques, computer networks

ABSTRACT

1 Introduction

Statistical techniques are extremely useful in system performance evaluation. Since any statistic cannot be guaranteed to give a close estimate for every sample, we must design statistics that 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 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 a confidence interval, under Multiple Replications in Parallel – MRIP, 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 means under Multiple Replications in Parallel.

2 Looking at the data sequentially

Typically, the runlength 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 procedure and is the subject of this work.

Sequential procedures, also known as non-fixed sample size experiments, gather observations at the output of a simulation model to investigate a certain parameter of interest, and a decision has to been taken to make the required estimation and 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 the outcome of the observations. According to this thought 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 a reasonable method for controlling the precision of simulation results. Two fundamental issues that motivate the design of more efficient sequential procedures is the possibility of specifying the desired precision for the estimated parameter, and the termination rule to conclude the experiment whether the precision has been reached. Ideally, the rule should be computationally as easy as the relative half-width criterion.

In order to assess the error done by estimating \( \bar{X} \), a steady-state estimator of a performance parameter \( \mu \) characterising 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, \( \{X_i : i = 1, \ldots, n\} \) are usually positive correlated and normality is also not always satisfied.

Positive correlation denotes negative bias and, thus, the final confidence interval half-width can be underestimated which leads frequently to a final coverage less than the nominal confidence level. By coverage we mean the frequency that a final confidence interval contains the true value being estimated.

Several confidence interval procedures (CIPs) have been suggested in the literature to get valid confidence intervals from the output of a simulation model. Some CIPs try to reduce this correlation, e.g., by means of grouping the observations (Batch Means approach) or taking 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

Since Batch Means (BM) is popular and very frequently applied in simulation experiments, we have found interesting investigating their performance and applicability under Multiple Replications In Parallel (MRIP), by means of some sequential versions that can be run independently on workstations connected via a network. Considering that each copy of the model is initiated with a non-overlapping parallel time-stream of random number, estimates coming from them can be taken for granted as independent from each other.

MRIP takes into account that generating data independently, and using an asynchronous communication among the processors (which avoids any situation of dead-lock) one can get a sound speedup. Theoretical justification can be found in Raj and Khamis [14] and Raj [15], 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 analysing 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 analysed classical procedure divides a a series of steady-state observations of length \( N \), into B adjacent nonoverlapping batches of size \( M \) (\( N=M/B \), for simplicity), and test correlation coefficients for lag \( k (k=1, \ldots, N/10) \) against correlation (refer to [11] 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 tests. Finding the "optimal" batch size, observations are reorganised into few batches \( (10 \leq B \leq 30) \), and new batches of observations are collected. at predefined checkpoints, each replication generates an estimates of the parameter being analysed and send them to a global analyser that is responsible for stopping the simulation when the desired precision is achieved.

#### 3.2 Overlapping Batch Means - OBM

Meketon and Schmeiser [17] proposed making a better use of the collected data and, 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 maintains 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. This mechanism is called '1-1. 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 exposed above, one can perceive that OBM is more trickier than NOBM, as the higher the number of degree of overlapping, the higher the amount of computation of batch means. On the other hand, the higher the degree of overlapping, the more values are used in the estimation. Statistically, it implies in lower variance and, expected better coverage, the main criterion of CIP
performance comparison.

4 Performance issues

In order to compare the performance of the above sequential CIPs, one should construct a confidence region at an assumed confidence level \( \eta \) \( \mathbf{R}(\eta, \mathbf{X}) \). This region is obtained by making some assumptions about the random characteristic of the data. If these assumptions are satisfied, the region \( \mathbf{R}(\eta, \mathbf{X}) \) contains the unknown parameter \( \mu \) being estimated with probability

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

When a procedure is correct, the observed coverage is equal to the desired coverage, or 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. [13], and this metric, besides the average number of observations \( \mathbb{E}[O] \) an the variability of the final confidence intervals given by the coefficient of variation of its half-width \( \text{CoV}(H) \), will be used to infer the performance of the CIPs under MRIP.

4.1 Degree of overlapping

Periodically, a sequential procedure collects estimates to verify whether the stopping rule has been achieved. At this moment, \( N \) observations are available, distributed in \( B \) contiguous, nonoverlapped batches of size \( M \) (\( N=M.B \)).

If one divides each batch into two equal 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 \( 2(B-1)+1 \) batches. By dividing each batch in four equal sized parts one obtains \( 4(B-1)+1 \) batches, and so on. Schematically:

The last case \( (M/M) \) is known as complete overlap, that is each observation begins a new (overlapped) batch. Meleton and Schmeiser [9] suggest an asymptotic number of freedom equal \( 1.5(B-1) \). The other cases are generally known as partial overlap [20]. However, as pointed out by Sargent et al. [16], to get a \( 1/3 \) variance reduction when compared to the variance of the classical NOBM, Meleton 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.

Results for \( P=1 \) shown in Table 0.31 support this statement, if we consider just the coverage of the results. If we look at the average number of observations \( \mathbb{E}[O] \), 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, although difference here was not so outstanding. The same we can say about the variability of the results measured by the coefficient of variation of the c.i. half-length, \( \text{CoV}(H) \).

<table>
<thead>
<tr>
<th>Degree of overlap</th>
<th>Number of batches</th>
<th>degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M/2 )</td>
<td>2(B-1)+1</td>
<td>1.33</td>
</tr>
<tr>
<td>( M/4 )</td>
<td>4(B-1)+1</td>
<td>1.45</td>
</tr>
<tr>
<td>( M/8 )</td>
<td>4(B-1)+1</td>
<td>1.48</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( M/M )</td>
<td>M(B-1)+1</td>
<td>1.50</td>
</tr>
</tbody>
</table>

Table 1: Degrees of freedom for complete and partial overlapping

<table>
<thead>
<tr>
<th>Degree of overlap</th>
<th>( \rho )</th>
<th>( \eta )</th>
<th>( \text{cov}(H) )</th>
<th>( \mathbb{E}[O] )</th>
<th>( \text{CoV}(H) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M/2 )</td>
<td>95</td>
<td>95</td>
<td>94.9±1.1</td>
<td>2143911</td>
<td>0.0257</td>
</tr>
<tr>
<td>( M/4 )</td>
<td>95</td>
<td>95</td>
<td>92.8±1.4</td>
<td>2129494</td>
<td>0.0270</td>
</tr>
<tr>
<td>( M/8 )</td>
<td>95</td>
<td>95</td>
<td>94.0±1.2</td>
<td>2131604</td>
<td>0.0268</td>
</tr>
<tr>
<td>( M/M )</td>
<td>95</td>
<td>95</td>
<td>94.9±1.0</td>
<td>2212278</td>
<td>0.0255</td>
</tr>
</tbody>
</table>

Table 2: OBM performance according to different degrees of overlapping

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 at our disposal and relax to some extent the precise determination and compensate the inherently negative effects by adding more processors, or equivalently more simulation engines.

We have investigated this idea with both sequential procedures, (OBM) and (NOBM). Instead of calculating the correlations coefficients for lags 1 to

1In this experiment we collected just 150 bad confidence intervals to initiate the sequential coverage analysis, as each experiment takes really very long time
L, where L is 10% of the sample size, we calculated just for lag 1. Moreover, we adopted a smaller number of batches for testing correlations, which should degrade the coverage.

Indeed, coverage that has been usually quite attractive for OBM, has now a worst value, but as we add more processors the coverage can be improved. Classical OBM seems not to take advantage on this important of MRIP, as the coverage does not achieve even the lowest bound for an acceptable coverage, chosen as 10% of the nominal confidence level. One can conclude that MRIP itself can be a variance reducer and compensates certain lack of fulfilment of assumptions of the methods of analysis, provided that they are robust. Figure 1 contains our empirical results.

5 Suitability to parallel processing

Under MRIP, another criterion of fundamental importance is the relationship between the length of preprocessing of each CIP and the total length of the simulation. By preprocessing we mean the extension of preliminary tasks carried on by the procedure until it begins delivering estimates to the central analyser. In terms of Amdahl’s law, it implies the fraction of the simulation that cannot be parallelised.

Concerning batching methods, the preprocessing corresponds to the batch size determination (BSD) (and the length of transient phase as well, when estimating steady-state parameters). 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.

![Fig. 1: OBM - compensating batch size imprecision with higher degree of parallelization](image)

It is worthwhile to emphasize that, as we used an M/M/1 very highly loaded (ρ = 95%) and constructed a confidence interval at 95% of confidence level with a 5%-relative precision of stopping rule, BSD is very long and that’s why NOBM offers so poor performance, while the actual implementation of OBM requires more observations to stopping the simulation, the relation BSD/N suggests that this method is promising to be used under MRIP, and due the fact that it is robust when the underlying requirements are not completely fulfilled, 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 queueing systems under increasing traffic intensity and high confidence level, we decided to assess the MRIP in practice, by means of a simulation of a real time-consuming network problem.

We consider a CDMA based mobile communication system with a specific number of Wireless Terminals (WTs). All WTs communicate with one
central Base Station (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 Additive White Gaussian Noise (AWGN) channel with Binary Phase Shift Keying (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.

6.1 Simultaneous MAC Packet Transmission (SMPT)

Using the simplest ARQ mechanism Send and Wait, like it is discussed in [1] and suggested within the recent wireless LAN standards [7, 3], each erroneous packet is retransmitted while following stored packets have to wait until the packet has been transmitted successfully. So the effective bit rate decreases from $B_{\text{good}}$ to $B_{\text{bad}}$ and simultaneously the jitter of following packets increases. We assume that a resulting bit rate $B_{\text{bad}}$ is not acceptable for the required throughput specified by the QoS parameters. Further the increased jitter is not acceptable for the application.

SMPT is a method that overcomes the problem of unstable bandwidth and high delay variation under the use if parallel channels (see [4]). In contrast to the sequential transmission SMPT performs transmissions in parallel. But the parallel channels will only be used to retransmit corrupted packets. Therefore stored packets will not suffer by the retransmissions on the initial channel. There are different approaches how the parallel channels can be used considering that each additional active channel will cause a lower Signal Noise Ratio [5, 6].

7 The Simulation Model

The simulations have been performed using the Ptolemy simulation tool [22] 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. Each WT generates a stream of transport units (like UDP segments and therefore called segments) with a specific load and pass these segments to the DLC layer via the network layer, where each of them is divided into a group of DLC packets. To each packet a header with length $\zeta$ is added. This header $\zeta$ is used to identify DLC packets in the right order and to assign the DLC packets to the appropriate segment and means for error detection. The frame, which is composed by one DLC packet and the header $\zeta$ is called a Data Link Control Packet Data Unit (DPDU). The length of a DPDU is denoted as $L_DPDU$. All DPDUs are stored in an queue with a fixed length $L_{queue}$ within the DLC layer and will be sent with different ARQ based transmission methods over the wireless link. Packets that are prone to errors only by one bit error will not be decoded successfully on the receiver side and will be counted as loss. The load generation module generates packets with constant bit rate allowing variable as well as fixed segment sizes with a specific load. We neglect the fading effects and assume an optimal power control within the WTs. Nevertheless using the wireless channel each WT will influence other WTs by an increased background noise.

To get a feeling how SMPT will influence positively the QoS for the wireless link we will investigate 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.

8 Results and final considerations

While writing this draft version of our paper we have not got a clear idea of how long this simulation model would take to give us accurate estimates of the jitter, as it is running for some days on a single (though fast) processor. Hopefully, when writing the final version we can give the exact speedup achieved, whereas running the same model on 10 almost homogenous processors under Alarcon-2, an MRI implementation developed at the University of Canterbury, Christchurch, New Zealand, we got results in around 10 hours, and they express reasonably what we expected about the problem.

Reliability of the results can be based on our empirical investigations on different queueing 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


