Journal of Simulation (2008) 2, 69–80

www.palgrave-journals.com/jos

Variance-based sampling for simulating cycle time—throughput curves using simulation-based estimates

JW Fowler[1]*, SE Leach[1†], GT Mackulak[1] and BL Nelson[2]

[1]*Arizona State University, Tempe, AZ, USA;* [2]*Northwestern University, Evanston, IL, USA*

Simulation at several different traffic intensities is required when generating a cycle time-throughput (CT-TH) curve. Previous work has shown that a variance-based allocation using an asymptotic variance approximation results in more precise confidence intervals than those achieved through naïve sampling in which an equal amount of effort is allocated to each traffic intensity being investigated. Many systems for which these CT-TH curves are desired are too complex for an asymptotic variance approximation to be easily determined. This paper presents a fixed-budget variance-based sampling allocation procedure for the simulation of CT-TH curves using variance estimates of the sample mean calculated from pilot simulation runs. The proposed allocation procedure significantly improved the range of precision over the naïve allocation.
Journal of Simulation (2008) **2,** 69–80. doi:10.1057/palgrave.jos.4250033

Keywords: simulation; queueing systems; confidence intervals; and variance

1. Introduction

1.1. Motivation

A cycle time-throughput (CT-TH) curve displays the projected average cycle time plotted against throughput rate, or start rate, with cycle time defined as the time from entering to leaving the system. CT-TH curves are often employed as decision-making tools in manufacturing settings (Brown *et al*, 1997). For other than the simplest of systems, simulation is commonly used to generate various points along the curve. The systems for which CT-TH curves are desired are typically complex, requiring long simulation run lengths and extensive output analysis. In most manufacturing settings, the time and budget available for simulation activities is limited, thus increasing the importance of sampling decisions.

The most straightforward sampling method, and the one most commonly used by practitioners, is to allocate an equal amount of simulation effort to each throughput rate being simulated, referred to as naïve sampling. In the case of a CT-TH curve where cycle time variance is known to increase rapidly as throughput approaches capacity, naïve sampling is likely to lead to widely varying precision at the throughput rates simulated.

*Correspondence: JW Fowler, Department of Industrial Engineering, Arizona State University, PO Box 875906, Tempe, AZ 85287-5906, USA. E-mail: john.fowler@asu.edu

†Current address: Department of Systems and Engineering Management, Air Force Institute of Technology, 2950 Hobson Way, Wright-Patterson AFB, OH 45433-7765, USA . Tel: +1 937 255 3636 x7390, Fax: +1 937 656 4699, E-mail: sonia.leach@afit.edu

Figure 1 presents an example of a CT-TH curve with widely varying precision. Since CT-TH curves support start rate-decisions, it is reasonable to assume that widely varying precision along the curve is undesirable.

1.2. Statement of the problem

Several different traffic intensities, that is design points, must be simulated in order to generate a CT-TH curve. While other papers address methods for generating a CT-TH curve (see Fowler *et al*, 2001; Park *et al*, 2002; Yang *et al*, 2007), the objective of this paper is to determine an allocation of simulation effort to the design points of the CT-TH curve being simulated. In this paper, we consider an approach using simulation-based sample means and variance estimates of the sample means obtained during pilot simulation runs to determine a fixed-budget sampling strategy that strives to achieve nearly equal absolute or relative precision.

2. Background and theory

Leach *et al* (2005) presents an allocation-of-effort procedure that aims to produce relatively equal precision along the CT-TH curve. The range of precision along a CT-TH curve is determined by the difference between the largest and smallest precision generated at the design points simulated. Absolute precision at design point h, $AP(h)$, is measured by confidence interval half-width at that design point, and the range of absolute precision, designated as $RangeAP$, is measured by the difference between the largest and smallest

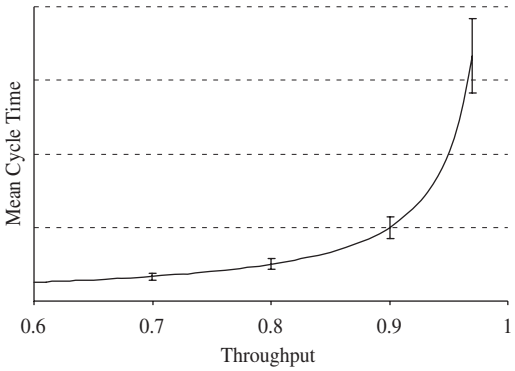

Figure 1 Widely varying precision along a CT-TH curve using equal allocation.

absolute precision of the design points simulated and calculated using:

$$RangeAP = \max_{h \in H} \{AP_h\} - \min_{h \in H} \{AP_h\} \quad (1)$$

where H is the set of all design points h that are simulated. Relative precision at design point h, $RP(h)$, and the range of relative precision, designated as $RangeRP$, are similarly measured using the ratio of the confidence interval half-width to the estimate of the mean response. The procedure in Leach *et al* (2005) uses asymptotic variance (Whitt, 1989) to determine the allocation of simulation effort to each design point prior to expending any simulation effort. Asymptotic variance, though, can be difficult to approximate for complex systems.

Pilot simulation runs are easily and commonly performed to collect and analyse a small portion of simulation generated data before expending the larger remaining portion of the simulation effort (Law and Kelton, 2000; Banks *et al*, 2005). Pilot simulation runs are generally required for methods to achieve a desired precision at a single design point. Extension of the single design point procedures to systems requiring simulation at multiple traffic intensities, though, is more complex. A detailed discussion of these issues is presented in Leach *et al* (2005).

This paper extends the previous work by providing a way to minimize the range of precision calculated for the design points simulated. The focus is the allocation of a fixed budget of available simulation effort to specific design points using the simulation-based sample mean and an estimate of variance of the sample means obtained from performing pilot simulation runs. With this capability, the design points investigated along a CT-TH curve can be simulated to produce precision that is more nearly equal.

3. Proposed fixed-budget method

To determine an allocation of simulation effort such that precision at each design point is nearly equal, a variance estimate of the mean response estimates, $S^2_{\bar{X}}$, is needed for each design point. These estimates will be generated from a portion of the total simulation budget available expended as pilot simulation replications. The variance estimate of the mean response estimates, $S^2_{\bar{X}}$, is used to calculate the confidence interval half-width about the grand mean response estimate, $\overline{\overline{X}}$; this confidence interval half-width being our measure of absolute precision. The confidence interval around the grand mean response estimate, $\overline{\overline{X}}$, is estimated by:

$$\overline{\overline{X}} \pm t_{1-\alpha/2,n-1}\sqrt{\frac{S^2_{\bar{X}}}{n}} \quad (2)$$

where n is the number of replications performed, and $t_{1-\alpha/2,n-1}$ is the $1-\alpha/2$ quantile of the Student's t distribution with $n-1$ degrees of freedom.

Assume it takes one unit of computer time to generate and process one simulated elementary observation. Let H be the set of all design points investigated in the simulation experiment, $H = \{h | h$ is a design point being simulated$\}$, $n(h)$ be the number of replications and $m(h)$ be the number of observations per replication for design point h. The total effort of the simulation, T, measured in number of elementary observations is given by:

$$T = \sum_{h \in H} n(h)m(h) \quad (3)$$

There are two options for arriving at an allocation of effort. The first option requires the simulation practitioner to set the number of observations per replication, or run length, at each design point and allows the methodology to determine the number of replications to perform, while the second option requires the simulation practitioner to set the number of replications to perform at each design point and allows the methodology to determine the run length. Methodologies for both options are presented.

3.1. Method for a fixed number of observations per replication

It follows from Equation (2) that the number of replications needing to be performed at design point h to achieve a fixed precision $\varepsilon(h)$ is:

$$n(h) \geqslant (t_{1-\alpha/2,n(h)-1})^2 \frac{S^2_{\bar{X}}(h)}{\varepsilon^2(h)} \quad (4)$$

The expression of the minimum total effort arrived at by substituting the equality in Equation (4) into Equation (3)

is given by:

$$T = \sum_{h \in H} \left((t_{1-\alpha/2, n(h)-1})^2 \frac{S^2_{\bar{X}}(h) m(h)}{\varepsilon^2(h)} \right) \tag{5}$$

From the discussions above, it follows that the portion of a fixed budget of total effort to be allocated at each design point is based upon the precision desired, $\varepsilon(h)$. Our objective is relatively equal precision at each design point investigated, either in absolute or relative terms. Given the length of the replications at each traffic intensity, $m(h)$, chosen by the simulation practitioner, equal absolute precision at each traffic intensity (all $\varepsilon(h) = \varepsilon$) corresponds to the total effort, T, given in Equation (6):

$$T = \frac{1}{\varepsilon^2} \sum_{k \in H} ((t_{1-\alpha/2, n(k)-1})^2 S^2_{\bar{X}}(k) m(k)) \tag{6}$$

The percent effort for each traffic intensity h, $\pi(h)$, is given in Equation (7):

$$\pi(h) = \frac{(t_{1-\alpha/2, n(h)-1})^2 S^2_{\bar{X}}(h) m(h)}{\sum_{k \in H} ((t_{1-\alpha/2, n(k)-1})^2 S^2_{\bar{X}}(k) m(k))} \tag{7}$$

If equal relative precision is desired (all $\varepsilon(h) = \gamma E[X(h)]$), the total effort is given by Equation (8):

$$T = \frac{1}{\gamma^2} \sum_{k \in H} \left((t_{1-\alpha/2, n(k)-1})^2 \frac{S^2_{\bar{X}}(k) m(k)}{[\bar{X}(k)]^2} \right) \tag{8}$$

The effort given each traffic intensity h for relative precision is given in Equation (9):

$$\pi(h) = \frac{(t_{1-\alpha/2, n(h)-1})^2 \frac{S^2_{\bar{X}}(h) m(h)}{[\bar{X}(h)]^2}}{\sum_{k \in H} \left((t_{1-\alpha/2, n(k)-1})^2 \frac{S^2_{\bar{X}}(k) m(k)}{[\bar{X}(k)]^2} \right)} \tag{9}$$

In either the absolute or relative precision cases, the number of replications to be accomplished at traffic intensity h is equal to:

$$n^*(h) = \frac{\pi(h) T}{m(h)} \tag{10}$$

Because implementation in a simulation requires an integer number of replications at each design point, $n(h)$ is rounded up to the next largest integer:

$$n(h) = \lceil n^*(h) \rceil = \left\lceil \frac{\pi(h) T}{m(h)} \right\rceil \tag{11}$$

As the highest traffic intensity investigated approaches system capacity, the nonlinearly increasing nature of cycle time variance and thus, the variance of the sample means, combined with a constrained budget produces allocations in which the highest traffic intensity claims nearly all of the effort, starving the lower traffic intensities. Therefore, a minimum number of runs at each design point must be designated. For example, the responsible simulation practitioner would not be satisfied with a simulation-generated estimate based on a single observed data point. Although guidelines for choosing a minimum number of replications have appeared in the literature, it is generally left to the practitioner to factor in conditions of the specific system being simulated. This paper adopts a minimum of five data points and four degrees of freedom established in Leach *et al* (2005) and supported in Law and Kelton (2000).

Imposing lower bounds and rounding up will most likely result in a total effort larger than the original budget. In cases in which additional effort is suggested by the sampling strategy but not available, the experimenter will need to reduce the effort to meet the given budget. The elimination of replications should be accomplished at the lowest traffic intensity first, provided the lower bound on the number of replications for that design point is maintained. If the reduction of replications forces all of the allocations to the lower bound, the practitioner may need to reconsider replication run lengths, $m(h)$, and total available budget, T, or possibly consider changing the experiment design (see Kelton (1986) and Schmeiser (1982) for related issues).

The following algorithm is provided for clarity.

Algorithm I

Given the following parameters:

T:	the total effort budgeted, measured in number of elementary cycle-time observations
T_R:	the total effort recommended, measured in number of elementary cycle time observations
$\boldsymbol{H}$:	the set of design points, h, to be investigated
k:	the iteration variable
$S^2_{\bar{X}}(h)$:	pilot simulation estimate of the variance of sample mean cycle time of design point h
$\bar{X}(h)$:	pilot simulation estimate of the mean for design point h
$\bar{X}_i(h)$:	estimate of the mean from pilot run i for design point h
$t_{1-\alpha/2, df}$:	the $1-\alpha/2$ quantile of the Student's t-distribution with df degrees of freedom
$n_k(h)$:	the number of replications at design point h during iteration k
$n_0(h)$:	the number of pilot simulation replications at design point h
$m(h)$:	the number of observations per replication at design point h
$nlb(h)$:	the lower bound on the number of replications needed at design point h
$p_k(h)$:	numerator of Equations (7, 9) for design point h during iteration k
$\pi_k(h)$:	percentage of effort allocated to design point h during iteration k
$\rho(h)$:	the traffic intensity associated with design point h

$AP_k(h)$: the absolute precision associated with design point h during iteration k
$RP_k(h)$: the relative precision associated with design point h during iteration k
$\boldsymbol{N}_k$: the vector of replications allocated to design points during iteration k
α: the confidence level desired
δ_a: the range of absolute precision desired
δ_r: the range of relative precision desired

1. Perform the $n_0(h)$ pilot simulation runs. For each design point h, calculate the sample mean, $\bar{X}(h)$, and the variance estimate of the sample mean, $S^2_{\bar{X}}(h)$:

$$\bar{X}(h) = \frac{\sum_{i=1}^{n_0(h)} \bar{X}_i(h)}{n_0(h)}$$

$$S^2_{\bar{X}}(h) = \frac{\sum_{i=1}^{n_0(h)} (\bar{X}_i(h) - \bar{X}(h))^2}{n_0(h) - 1}$$

2. Initialize the iteration variable:
   Set $k = 1$.

3. For each traffic intensity being considered:
   A. Calculate the contributions to the allocation equation—numerator of Equation (7) for absolute precision and numerator of Equation (9) for relative precision

$$\text{For absolute precision, } p_k(h) = (t_{1-\alpha/2, n_{k-1}(h)-1})^2 S^2_{\bar{X}}(h) m(h)$$

$$\text{For relative precision, } p_k(h) = \frac{(t_{1-\alpha/2, n_{k-1}(h)-1})^2 S^2_{\bar{X}}(h) m(h)}{[\bar{X}(h)]^2}$$

   B. Calculate the percentage of effort:

$$\pi_k(h) = \frac{p_k(h)}{\sum_{h \in H} p_k(h)}$$

   C. Calculate the replications required:

$$n^*_k(h) = \frac{\pi_k(h) T}{m(h)}$$

   D. Round the number of replications to an integer value:

$$n_k(h) = \lceil n^*_k(h) \rceil$$

   If $n_k(h) < nlb(h)$, then $n_k(h) = nlb(h)$.

   E. Calculate the theoretical precision:

$$\text{For absolute precision : } AP_k(h) = t_{1-\alpha/2, n_k(h)-1} \sqrt{\frac{S^2_{\bar{X}}(h)}{n_k(h)}}$$

$$\text{For relative precision : } RP_k(h) = \frac{t_{1-\alpha/2, n_k(h)-1}}{[\bar{X}(h)]^2} \sqrt{\frac{S^2_{\bar{X}}(h)}{n_k(h)}}$$

   F. Summarize allocations in vector format:

$$\boldsymbol{N}_k = [n_{1,k+1}, n_{2,k+1}, \ldots, n_{|H|,k+1}]$$

4. Determine if any one of the stopping criterion have been met:
   A. Acceptable range of precision if $RangeAP_k \leqslant \delta_a$ for absolute precision or if $RangeAP_k \leqslant \delta_r$ for relative precision, or
   B. Convergence of the allocation vector if the current vector of allocations equals a vector of allocations from a previous iteration, $\boldsymbol{N}_k = \boldsymbol{N}_a$ for some $a < k$, then let $n(h) = n_k(h)$ for all $h \in H$.

5. If either of the two stopping criteria are met, then let $T_R = \sum_{h \in H} n(h) m(h)$ and STOP. Otherwise, set $k = k + 1$ and return to step 3.

End of Algorithm I.

If upon completion of Algorithm I the amount of effort recommended for allocation, T_R, is greater than the original budget, the number of replications performed will need to be reduced, as discussed above. Algorithm II performs the necessary reductions and is provided for clarity.

Algorithm II
Define the additional parameter:

ΔT: the reserve budget beyond T, measured in number of elementary observations, where $\Delta T \geqslant 0$.

A. Calculate the effort required by the allocation suggested using Algorithm I.

$$T_R = \sum_{h \in H} n(h) m(h)$$

If $T_R - T \leqslant \Delta T$, STOP. Otherwise, continue.

B. Determine which design points can be reduced in effort:

Let F be the set of design points h such that $n(h) \geqslant nlb(h) + 1$. If $F = \varnothing$, STOP. The lower bound and total effort criteria cannot both be met. Reassess

lower bound and total effort choices and repeat Algorithm I. Otherwise, continue.

C. Choose eligible design point with lowest traffic intensity:

Choose design point j from F such that $\rho(j) < \rho(i)$, $i \neq j$, i, $j \in F$.

D. Reduce the effort at this design point:

Let $n(j) = n(j) - 1$.

E. Calculate new total effort required:

$$T_R = \sum_{h \in H} n(h)m(h)$$

F. Assess stopping criteria—can budget be met with revised allocation:

If $T_R - T \leqslant \Delta T$, STOP. Otherwise, return to step B.

End of Algorithm II.

3.2. Method for a fixed number of replications

For situations in which the number of replications performed at each design point, $n(h)$, is fixed, the pilot run-based allocation methodology is used to determine the run length, or observations per replication, $m(h)$, at each design point. In order to apply the methodology for a fixed number of replications at each design point, the number and length of the pilot simulation runs needed to obtain the sample mean and variance estimate of the sample means must be specified. Let $m_p(h)$ designate the number of observations per replication performed at design point h during pilot simulation runs.

From the derivations in Leach *et al* (2005) and using $m_p(h)S^2_{\bar{X}}$ as an estimate of asymptotic variance for $m_p(h)$ not too small, the number of replications at each design point is represented by:

$$n(h) \geqslant (t_{1-\alpha/2,\, n(h)-1})^2 \frac{S^2_{\bar{X}}(h)m_p(h)}{\varepsilon^2(h)} \tag{12}$$

Substituting (12) into (3) results in:

$$T = \sum_{h \in H} \left((t_{1-\alpha/2,\, n(h)-1})^2 \frac{S^2_{\bar{X}}(h)m_p(h)m(h)}{\varepsilon^2(h)} \right) \tag{13}$$

Requiring the run length of the pilot simulation runs to be the same at each design point (all $m_p(h) = m_p$), and equal absolute precision at each traffic intensity (all $\varepsilon(h) = \varepsilon$), the expression of total effort is given as Equation (14):

$$T = \frac{m_p}{\varepsilon^2} \sum_{h \in H} ((t_{1-\alpha/2,\, n(h)-1})^2 S^2_{\bar{X}}(h)m(h)) \tag{14}$$

The portion of total effort allocated to each design point is given as Equation (15):

$$\pi(h) = \frac{(t_{1-\alpha/2,\, n(h)-1})^2 S^2_{\bar{X}}(h)m(h)}{\sum_{k \in H} ((t_{1-\alpha/2,\, n(k)-1})^2 S^2_{\bar{X}}(k)m(k))} \tag{15}$$

When interested in equal relative precision, the total effort and portion of effort allocated to each design point are arrived at using the same arguments as above. The total effort for relative precision is calculated by:

$$T = \frac{m_p}{\gamma^2} \sum_{k \in H} \left((t_{1-\alpha/2,\, n(k)-1})^2 \frac{S^2_{\bar{X}}(k)m(k)}{E^2[\bar{X}(k)]} \right) \tag{16}$$

The portion of effort allocated to each design point is given by:

$$\pi(h) = \frac{(t_{1-\alpha/2,\, n(h)-1})^2 \frac{S^2_{\bar{X}}(h)m(h)}{E^2[\bar{X}(h)]}}{\sum_{k \in H} \left((t_{1-\alpha/2,\, n(k)-1})^2 \frac{S^2_{\bar{X}}(k)m(k)}{E^2[\bar{X}(k)]} \right)} \tag{17}$$

The run length of the $n(h)$ replications performed at a given design point would be calculated by:

$$m(h) = \left\lceil \frac{\pi(h)T}{n(h)} \right\rceil \tag{18}$$

The situation incurred by following this procedure is that the pre-specified run length of the pilot simulation runs is very likely to differ from the run length recommended by the variance-based pilot simulation allocation method. The mismatching of run lengths of the pilot simulation replications and the remaining allocation of replications leads to two options for proceeding with this methodology. These two options differ by the manner in which the simulations are performed and data are collected.

The first fixed replications option involves performing the simulations as independent replications and capturing replication means and variance estimates of the replication means as a basis for the pilot simulation estimates. If the pilot simulation run lengths differ from the run lengths performed for the remainder of the simulation replications, the overall response mean estimate and precision calculation would be difficult to interpret. In this case, the pilot simulation runs representing the smaller portion of the total simulation budget expended would be discarded and the overall response mean and precision calculation would be based only upon the remaining complement of simulation replication data. The removal of simulation runs in a

constrained budget environment is a significant limitation in proceeding with this option.

The second fixed replication option involves performing the simulation as one continuous replication analysed using non-overlapping batch means. In this option, $m(h)$ is the number of elementary observations collected at each design point. This option is more procedurally intensive, requiring two potentially cumbersome data collection issues: recording and storing every elementary observation; and recording and storing the state space and position in the random number streams at the conclusion of the pilot simulation activity. In the replication means option presented above, each replication is initiated from the empty and idle state and only the mean of the elementary observations of each replication is required to be stored. In addition to the data collection issues, this option requires the determination of number and size of batches for the pilot simulation data and number of batches for the total complement of simulation data. Issues surrounding batch number and size selection include high variability if too few batches are chosen and diminishing returns for precision and increased impact of the initial transient if too many batches are chosen (Schmeiser, 1982). The advantage of this option, though, is that all of the simulation data, including the pilot simulation activity data, would remain available for the calculation of the overall mean estimate and precision.

Algorithms I and II can be used to determine the allocations given a situation of fixed number of replications at each design point with a few modifications. In Algorithm I, Step 3D must be changed to round the number of observations per replication to an integer value, such that $m_k(h) = \lceil m_k^*(h) \rceil$. If $m_k(h) < mlb(h)$, where $mlb(h)$ is the lower bound established for the run length at design point h, then $m_k(h) = mlb(h)$. Step 3F would be changed to reflect a vector of run lengths and the corresponding stopping condition would be the convergence of the vector of run lengths. For Algorithm II, Steps B and D would involve reducing the run length at the chosen design point until the effort available is met.

4. Experimentation and results

Experimentation was conducted using a discrete event simulator implementing the random number generator suggested in Marse and Roberts (1983) and recommended in Law and Kelton (2000). As suggested in previously published literature, a crude truncation of 20% is employed (Fowler *et al*, 2001). Mean cycle time estimates are calculated indirectly from mean delay time estimates and expected time in service (Law, 1975; Carson and Law, 1980).

4.1. Experimentation and results for a fixed number of observations per replication

M/M/1 queueing model. Experimentation of the proposed sampling method was conducted on an M/M/1 queueing system operating under a first-in, first-out (FIFO) policy with a service rate of $\mu = 1$. Given the assumptions of system throughput being equal to the arrival rate of $\lambda < 1$ and a yield of one, the service rate of $\mu = 1$ allows traffic intensity, $\rho = \lambda/\mu$, to be used as the system throughput. Since the region of the CT-TH curve of primary interest is as the system approaches capacity, the four design points chosen for experimentation have the traffic intensities of $\rho(1) = 0.70$, $\rho(2) = 0.80$, $\rho(3) = 0.90$ and $\rho(4) = 0.97$ (Fowler *et al*, 2001), where $\rho = 1.00$ represents system capacity.

The results of the proposed sampling method are compared to a naïve sampling in which the observations per replication are the same for each design point, as suggested in Fowler *et al* (2001). A total effort of $T = 40$ million observations was budgeted for the total simulation effort of each M/M/1 CT-TH curve, and run length for each replication performed was set to 250 000 observations.

Table 1 presents the sampling allocations resulting from Algorithms I and II and range of precision results for experimentation repeated five times. Four allocation methods are compared. The first is the naïve method, described earlier, which presents the allocation in which an equal number of replications are performed at each design point. The second method is the asymptotic variance-based method that presents the allocation resulting from the sampling procedure of Leach *et al* (2005). The third and fourth methods of allocation represent the method proposed in this paper. These two methods differ in the manner in which the pilot runs are allocated to the design points. The third method allocates the pilot simulation effort to each design point naïvely, and the fourth method uses the asymptotic variance-based allocation procedure from Leach *et al* (2005). The fourth method is presented for comparison purposes to investigate whether having even a crude asymptotic variance approximation to allocate pilot simulation effort offers an advantage in identifying an allocation of total effort resulting in improved range of precision. In the table of results, these final two methods are identified by the pilot run allocation of replications for the four design points. For example, an allocation identified by '5/6/7/8' would signify that five pilot replications were performed at the lowest traffic intensity design point, six pilot replications at the second lowest traffic intensity design point, and so on. Between 12.5 and 25% of the total effort available was allocated to the pilot simulation runs, corresponding to naïve allocations of five and ten pilot replications per design point. Five pilot replications corresponds to the lower bound on the number of replications at each design point and 10 pilot replications corresponds to the maximum recommended 25% of the total simulation budget (Law and Kelton, 2000).

In the absolute precision cases, the smallest range of precision corresponds to five pilot simulation replications, as indicated by the bold and underlined values in the table. This is the same allocation produced by the asymptotic

Table 1 Allocations and range of precision ($\alpha = 0.05$) for the M/M/1 queueing system

Replication allocations	*Pilot run allocation (%)*	*Absolute precision results*						*Relative Precision Results*					
		Traffic intensity (ρ)				*RangeAP*		*Traffic Intensity (ρ)*				*RangeRP (%)*	
		0.70	*0.80*	*0.90*	*0.97*	*Mean*	*SD*	*0.70*	*0.80*	*0.90*	*0.97*	*Mean*	*SD*
Obs/rep, *m*(*h*)		250K	250K	250K	250K			250K	250K	250K	250K		
Naïve	0.0	40	40	40	40	1.3823	0.1533	40	40	40	40	3.7989	0.4765
Asy. var based*,†	0.0	5	5	5	145	**0.7089**	0.0382	5	5	12	138	**1.3501**	0.3821
5/5/5/5*	12.5	5	5	5	145	**0.7089**	0.0382	5	5	7	143	1.7994	0.7372
6/6/6/6	15.0	6	6	6	142	0.7182	0.0249	6	6	6	142	2.0961	0.7345
5/5/5/9*	15.0	5	5	5	145	**0.7089**	0.0382	5	5	5	145	2.6915	1.1142
8/8/8/8	20.0	8	8	8	136	0.7450	0.0270	8	8	8	136	1.8435	0.4447
5/5/5/17*	20.0	5	5	5	145	**0.7089**	0.0382	5	5	5	145	2.6915	1.1142
10/10/10/10	25.0	10	10	10	130	0.7731	0.0284	10	10	10	130	1.7676	0.0666
5/5/5/25*	25.0	5	5	5	145	**0.7089**	0.0382	5	5	5	145	2.6915	1.1142

*denotes best range of absolute precision.
†denotes best range of relative precision.

variance-based allocation method from Leach *et al* (2005) as well as the pilot simulation runs allocated according to the asymptotic variance-based allocation. These results indicate that as the number of pilot simulation replications allocated at the lower traffic intensities increases, the reduced effort at the higher throughputs increases variance. This in turn increases the range of precision across all design points. In the relative precision results, the use of 25% of the total effort toward pilot simulation replications produced the smallest pilot run-based range of precision, though the asymptotic variance-based allocation had the best relative range of precision overall.

This same experimentation was repeated for varying run lengths for the four design points chosen. The three additional run-length scenarios investigated were: scenario 1 with $m(1) = m(2) = 125\,000$, $m(3) = 250\,000$ and $m(4) = 500\,000$; scenario 2 with $m(1) = 100\,000$, $m(2) = 0$, $m(3) = 300\,000$ and $m(4) = 600\,000$; and, scenario 3 with $m(1) = 100\,000$, $m(2) = 300\,000$, $m(3) = 0$ and $m(4) = 600\,000$. Scenario 1 investigates the effect of increasing run lengths for higher variance throughput levels, and scenarios 2 and 3 investigate the same effect coupled with the omission of an interior design point. For absolute precision, the smallest ranges of precision correspond to five pilot simulation replications in all three additional scenarios. This is the same allocation produced by the asymptotic variance-based allocation method from Leach *et al* (2005) as well as the pilot simulation runs allocated according to the asymptotic variance-based allocation. For relative precision, the best range of precision resulted from the asymptotic variance-based allocation. For scenario 1, the best range of precision resulted from the pilot run-based allocation using 20.0% of the total effort and allocated to the design points naïvely. In general, though, the allocations determined by the variance-based pilot simulation allocation method produced smaller ranges of both absolute and relative precision than the corresponding naïve allocations. Detailed discussion and results for these investigations appears in Leach (2005).

Additional 4-design point allocations of the M/M/1 queueing model. The concentration of effort at the highest traffic intensities exhibited by both the asymptotic variance-based and pilot simulation-based allocation methods is related to the distance (gap) between design points (0.10). Three additional 4-design point experiments were conducted capturing a smaller gap size of 0.05 starting with traffic intensities of $\rho(1) = 0.70$, 0.75 and 0.80 to evaluate the procedure. The results for this set of experimentation is detailed in Leach (2005). A large portion of effort is still allocated to the largest traffic intensity in the absolute precision case, but fewer of the lower traffic intensity allocations are at the lower bound. This is more obviously the case for relative precision. Concentration of effort begins to creep up to the highest traffic intensity for both absolute and relative precision when using design points of $\rho(h) = 0.80$, 0.85, 0.90 and 0.95. The best range of precision for either absolute or relative precision occurred using the asymptotic variance-based allocation. Overall, though, the range of precision results reveals that the proposed variance-based pilot simulation allocation method produces smaller ranges than the naïve method.

Five station Jackson network. Experimentation was also conducted on a 5-station Jackson network model (Fowler *et al*, 2001). This network offers complexity while still allowing reasonable analytical queueing calculations for comparison. The system description and analytical queueing results are shown in Table 2, where w_i is the analytical cycle time for station i. Station 3, indicated in bold, is identified as the bottleneck station and is used to determine

Table 2 Analytical queueing results for the 5-station Jackson network queueing model

Station	*Number of servers*	*Service rate*	*Arrival rate (λ)*							
			0.6		*0.7*		*0.8*		*0.9*	
i	c_i	μ_i	ρ_i	w_i	ρ_i	w_i	ρ_i	w_i	ρ_i	w_i
1	1	1.0100	0.59	2.4390	0.69	3.2258	0.79	4.7619	0.89	9.0909
2	5	0.2100	0.57	5.2114	0.67	5.6952	0.76	6.6940	0.86	9.2248
3	**4**	**0.2500**	**0.60**	**4.7176**	**0.70**	**5.4288**	**0.80**	**6.9822**	**0.90**	**11.8775**
4	4	0.2525	0.59	4.6421	0.69	5.3087	0.79	6.7330	0.89	10.9587
5	3	0.3500	0.57	3.5659	0.67	4.1270	0.76	5.2024	0.86	7.8075
			Total:	20.5760		23.7856		30.3735		48.9595

Table 3 Allocations and range of precision ($\alpha = 0.05$) for the 5-station Jackson network queueing system

Replication allocations	*Pilot run allocation (%)*	*Absolute precision results*						*Relative precision results*					
		Traffic intensity (ρ)				*RangeAP*		*Traffic intensity (ρ)*				*RangeRP (%)*	
		0.60	*0.70*	*0.80*	*0.90*	*Mean*	*SD*	*0.60*	*0.70*	*0.80*	*0.90*	*Mean*	*SD*
Obs/rep, $m(h)$		750K	750K	750K	750K			750K	750K	750K	750K		
Naïve	0.0	40	40	40	40	0.1845	0.0160	40	40	40	40	0.3483	0.0317
Asy. var based	0.0	5	5	9	141	0.0887	0.0300	5	11	26	118	0.1004	0.0540
5/5/5/5*	12.5	5	5	24	126	**0.0623**	0.0162	5	15	47	93	0.1116	0.0448
6/6/6/6	15.0	6	6	31	117	0.0657	0.0093	6	14	59	81	0.1449	0.0337
5/5/5/9	15.0	5	5	5	145	0.1743	0.0447	5	5	23	127	0.1817	0.0721
8/8/8/8	20.0	8	8	13	131	0.0889	0.0100	8	8	43	101	0.1451	0.0674
5/5/5/17	20.0	5	5	8	142	0.1084	0.0407	5	8	28	119	0.1065	0.0926
10/10/10/10	25.0	10	10	10	130	0.1023	0.0207	10	10	19	121	0.1617	0.0377
5/5/5/25†	25.0	5	5	11	139	0.0757	0.0255	5	9	31	115	**0.0948**	0.0691

*denotes best range of absolute precision.
†denotes best range of relative precision.

the traffic intensity of the entire system. This results in $\rho(h) = 0.60, 0.70, 0.80$ *and* 0.90 as the traffic intensities investigated. The observations per replication scenarios provided in Table 1 were investigated for this model, with a total of $T = 120$ million observations budgeted for each CT-TH curve. Run length for each replication performed was set to 750 000 observations.

Table 3 provides the allocations, determined using Algorithms I and II, and range of precision results for the 5-station Jackson network queueing system, and as in Table 1, the bold and underlined values represent the best performances. These results confirm the success of the proposed variance-based pilot simulation allocation method with respect to producing smaller ranges of precision for the design points simulated. In the absolute precision scenarios, the smallest range of precision corresponded to the 12.5% naïve pilot simulation allocation. As was observed in the M/M/1 queueing model results, these results indicate that the additional effort at the lower traffic intensities adversely affects the range of absolute precision. In the relative precision results, the best performing pilot replication allocation produces a smaller range of relative precision than the naïve allocation. These results indicate great promise for the proposed variance-based pilot simulation allocation method in that the 5-station Jackson network is more representative of true systems than an M/M/1 system.

Similar to the M/M/1 experimentation of Section 4.1 above, this same experimentation was repeated for varying run lengths for the four design points chosen. The three additional run-length scenarios for the 5-station Jackson network investigated were: scenario 1 with m(1) = m(2) = 375 000, m(3) = 750 000 and m(4) = 1 500 000; scenario 2 with m(1) = 300 000, m(2) = 0, m(3) = 900 000 and m(4) = 1 800 000; and, scenario 3 with m(1) = 300 000, m(2) = 900 000, m(3) = 0 and m(4) = 1 800 000. For absolute precision, the 15% pilot run allocation of total effort produced the best results, outperforming the asymptotic variance-based results. For relative precision, the results across scenarios varied. For scenario 1, the asymptotic variance-based allocation produced the smallest range of precision, while for scenarios 2 and 3, the 20 and 25% naïve allocations produced the smallest ranges of precision, respectively. Detailed discussion and results for these investigations appear in Leach (2005).

4.2. Experimentation and results for a fixed number of replications

Independent replication implementation for a fixed number of replications. The independent replication implementation for a fixed number of replications was investigated for the M/M/1 queueing network model using five pilot replications of 100 000 elementary observations per replication performed at each design point, which is equal to 5% of the total effort available. For consistency and comparison to previous experimentation, $n(h) = 40$ was chosen for each design point. It was determined that the run length for the subsequent replications beyond the pilot runs would be no shorter than the length of the pilot runs, therefore the minimum acceptable run length was also set to 100 000 elementary observations per replication. In both the absolute and relative precision cases, the pilot run-based allocation method allocated the minimum acceptable run length to each of the three lowest traffic intensities, allowing the pilot simulation replications at those design points to be incorporated into the final calculations of sample mean and precision. At the highest traffic intensity, though, the allocation of the subsequent runs did differ from the pilot runs requiring that the pilot runs for that design point be discarded and not incorporated into the final calculations of sample mean and precision. The resulting run length allocation and precision results are provided in Table 4 with the best range of precision results indicated as bold and underlined. In both the absolute and relative precision cases, the pilot run-based allocation using a fixed number of replications at each design point produced ranges of precision more than 50% smaller than the naïve allocation. These results indicate that the pilot run-based allocation procedure is successful at producing smaller ranges of precision under conditions of fixed replications at each design point.

The same experimentation was conducted on the 5-station Jackson queueing network model, with five pilot runs of 300 000 elementary observations per replication at each design point. Similar to the M/M/1 experimentation above, the resulting allocations recommended in the absolute precision case for the three lowest traffic intensities was equal to run length of the pilot replications allowing those runs to be used in the final analysis and only the pilot runs at the highest traffic intensity needing to be discarded. In the relative precision case, the pilot runs at the two highest traffic intensities were required to be discarded. The results for the 5-station Jackson queueing model experimentation are presented in Table 5, revealing similar improvements in range of precision of the pilot run-based allocations over the naïve allocations.

Batch means implementation for fixed number of replications. The batch means implementation for a fixed number of replications was conducted on the M/M/1 queueing network model and investigated for 5, 10, 20 and 40 batches. In each batching case, a total of 1 million elementary observations were collected during pilot simulation at each of the four design points, equating to 10.0% of the total budget. As with the independent replication implementation, the minimum allowable batch length was set to the length of the pilot simulation batch length producing the estimates for the allocation procedure. The associated naïve allocation to each batching case allocated one-quarter of the total budget available, 10 million elementary observations, to each design point and was analysed according to the number of batches indicated. The results for each batching case are presented in Table 6 with the best range of precision results for each case given in bold.

The absolute precision results for the batch means experimentation show that the variance-based pilot simulation allocation using fixed $n(h)$ outperforms the naïve allocation, with the best range of absolute precision result occurring with 40 batch means. For the relative precision cases, similar results were produced in all but the five batch means cases, with the best range of relative precision result occurring with twenty batch means. These results are consistent with the findings of Schmeiser (1982) in which fewer than 10 batches are not recommended because of high

Table 4 Allocations and range of precision ($\alpha = 0.05$) comparison for the M/M/1 queueing system fixed $n(h)$ experimentation

	Absolute precision results						*Relative precision results*					
	Traffic intensity (ρ)				*RangeAP*		*Traffic intensity (ρ)*				*RangeRP*	
	0.70	*0.80*	*0.90*	*0.97*	*Mean*	*SD*	*0.70*	*0.80*	*0.90*	*0.97*	*Mean*	*SD*
Naïve												
Obs/rep, $m(h)$	250K	250K	250K	250K			250K	250K	250K	250K		
Reps, $n(h)$	40	40	40	40	1.3426	0.1837	40	40	40	40	3.7020	0.5103
Variance based												
Obs/rep, $m(h)$	100K	100K	100K	690K			100K	100K	100K	690K		
Reps, $n(h)$	40	40	40	40	**0.6914**	0.0667	40	40	40	40	**1.6058**	0.1036

Table 5 Allocations and range of precision ($\alpha = 0.05$) comparison for the five station Jackson queueing system fixed $n(h)$ experimentation

	Absolute precision results						*Relative precision results*					
	Traffic intensity (ρ)				*RangeAP*		*Traffic intensity (ρ)*				*RangeRP (%)*	
	0.60	*0.70*	*0.80*	*0.90*	*Mean*	*SD*	*0.60*	*0.70*	*0.80*	*0.90*	*Mean*	*SD*
Naïve												
Obs/rep, $m(h)$	750K	750K	750K	750K			750K	750K	750K	750K		
Reps, $n(h)$	40	40	40	40	0.1845	0.0160	40	40	40	40	0.3814	0.0449
Variance based												
Obs/rep, $m(h)$	300K	300K	300K	2060K			300K	300K	570K	1755K		
Reps, $n(h)$	40	40	40	40	**0.0912**	0.0221	40	40	40	40	**0.1361**	0.0318

Table 6 Allocations and range of precision ($\alpha = 0.05$) comparison for the M/M/1 queueing system fixed $n(h)$ batch means experimentation

	Absolute precision results						*Relative precision results*					
	Traffic intensity (ρ)				*RangeAP*		*Traffic intensity (ρ)*				*RangeRP*	
	0.70	*0.80*	*0.90*	*0.97*	*Mean*	*SD*	*0.70*	*0.80*	*0.90*	*0.97*	*Mean*	*SD*
5 batches per design point												
Naïve allocation												
Obs/batch, $m(h)$	2000K	2000K	2000K	2000K			2000K	2000K	2000K	2000K		
Precision	0.1215	0.2034	0.5484	1.0188	1.2742	0.4787	3.531	3.9395	5.3326	3.001	**2.7562**	1.6365
Variance based allocation												
Obs/batch, $m(h)$	200K	200K	200K	7400K			200K	200K	1900K	5700K		
Precision	0.0393	0.1433	0.7152	0.2591	**0.9209**	0.2897	1.1771	2.8585	5.5267	0.4717	3.2145	1.5274
10 batches per design point												
Naïve allocation												
Obs/batch, $m(h)$	1000K	1000K	1000K	1000K			1000K	1000K	1000K	1000K		
Precision	0.0667	0.1237	0.3151	1.3203	1.3616	0.2925	1.9396	2.3961	3.0638	3.8892	2.6319	0.9546
Variance based allocation												
Obs/batch, $m(h)$	100K	100K	100K	3700K			100K	100K	290K	3510K		
Precision	0.0541	0.1175	0.4737	0.1967	**0.6088**	0.2373	1.6191	2.3442	2.6468	1.2785	**1.6677**	0.7784
20 batches per design point												
Naïve allocation												
Obs/batch, $m(h)$	500K	500K	500K	500K			500K	500K	500K	500K		
Precision	0.0437	0.0827	0.2337	1.1545	1.2796	0.1490	1.2692	1.602	2.2718	3.4006	2.7709	0.6961
Variance based allocation												
Obs/batch, $m(h)$	50K	50K	50K	1850K			50K	50K	260K	1640K		
Precision	0.0475	0.1084	0.4319	0.3912	**0.5527**	0.1793	1.4218	2.1631	1.803	1.7806	**0.9500**	0.2060
40 batches per design point												
Naïve allocation												
Obs/batch, $m(h)$	250K	250K	250K	250K			250K	250K	250K	250K		
Precision	0.0312	0.0609	0.1809	1.2787	1.3293	0.1006	0.9054	1.1798	1.759	3.7665	3.1788	0.4234
Variance based allocation												
Obs/batch, $m(h)$	25K	25K	25K	925K			25K	25K	120K	830K		
Precision	0.0413	0.092	0.3886	0.4893	**0.5230**	0.1300	1.2308	1.8295	2.0478	2.1228	**1.0027**	0.1368

variability. These results confirm the conclusions from the fixed $n(h)$ independent replication experimentation above that the pilot run-based allocation produces smaller ranges of precision than a naïve allocation in cases where the number of replications or batches are fixed.

Batch means implementation for 0.05 gap between design points. In the absolute precision batch means experimentation in the previous section, the minimum amount of effort was allocated at the three lowest traffic intensities with the remaining effort primarily going to the highest

traffic intensity. In the relative precision cases, this same allocation occurred for the two lowest traffic intensities, but the majority of the total effort was still allocated to the highest traffic intensity. As with the fixed $m(h)$ experimentation presented above, this resulting allocation of effort is primarily due to the large difference in the variance estimates of the sample mean between the largest and smallest traffic intensities investigated.

This finding is further investigated by experimentation on the M/M/1 queueing network model for cases in which the gap between the design point traffic intensities is 0.05. Three additional 4-design point experiments were conducted capturing a smaller gap size of 0.05 starting with traffic intensities of $\rho(1) = 0.70$, 0.75 and 0.80 to evaluate the procedure, similar to the M/M/1 experimentation described in Section 'Additional 4-design point allocations of the M/M/1 queueing model'. Each case was investigated for 20 batches, and detailed in Leach (2005). In all three cases, the variance-based pilot simulation allocation using batch means for implementing a fixed number of replications produced smaller ranges of both absolute and relative precision. These results confirm the improvement in range of precision using the pilot run-based allocations in the absolute and relative precision cases.

5. Findings and future work

The proposed method presented in this paper is based upon the sample mean and variance estimates of the sample mean generated from a portion of simulation budget expended as pilot simulation runs. The proposed method was successful in reducing the range of both absolute and relative precision for the CT-TH curves investigated for varying amounts of effort expended during pilot simulation activity.

For the fixed run length methodology, the results revealed that the pilot run-based allocations using the smallest amount of pilot simulation effort produced results just as good, if not better, than pilot run-based allocations using larger amounts of pilot simulation effort. These results follow from the premise that using a small amount of effort toward pilot simulation activity leaves a large amount of effort to be allocated appropriately according to the allocation methodology. In addition, the results revealed that allocating the pilot simulation effort using an asymptotic variance-based method did not produce better results in all cases. From these results, we conclude that implementation of the pilot run-based allocation method should be accomplished by naively allocating the minimum amount of effort allowed toward pilot simulation activity. For the fixed numbers of replications methodology, results revealed that the pilot run-based allocation methodology also produced smaller ranges of both absolute and relative precision than a naïve allocation. Implementation of this method by either independent replications or batch means proved to be equally successful.

The variance-based pilot simulation allocation methodology presented in this paper, for either a fixed number of replications or a fixed number of observations per replication, established a framework from which to achieve improved queueing simulation results simply by changing the allocation of simulation effort for different traffic intensities with respect to response variance. The two methods, asymptotic variance based and pilot run based, complement each other by providing two allocation options. The asymptotic variance-based method is appropriate for systems in which the practitioner has the time and expertise available to approximate the asymptotic variance. The pilot run-based method is applicable in situations in which the system is too complex for asymptotic variance to be well approximated, the time or expertise to determine the asymptotic variance is unavailable, or when pilot simulation runs are already being executed for some other purpose. Improvement in the range of precision of the results are expected regardless of which allocation method is employed.

Future research for this topic area will continue to focus on methods for determining the allocation of simulation effort for more complex systems. The asymptotic variance-based allocation method will also be formulated and solved as a mathematical optimization allowing both the number and length of the replications to vary freely.

Acknowledgements—This research was partially supported by grants DMII 0140441/0140385 of the National Science Foundation and grants 04-FJ-1224/04-FJ-1225 which are jointly funded by the Semiconductor Research Corporation and the International SEMATECH Manufacturing Initiative. Additionally, the authors would like to thank Professor Bruce Ankenman from Northwestern University for his contribution to this research. The views expressed in this article are those of the author and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the United States Government.

References

Banks J, Carson II JS, Nelson BL and Nichol DM (2005). *Discrete Event System Simulation*, 4th edn, Prentice-Hall: New Jersey.

Brown S, Chance F, Fowler JW and Robinson JK (1997). A centralized approach to factory simulation. *Future Fab Int* **1**(3): 83–86.

Carson JS and Law AM (1980). Conservation equations and variance reduction in queueing simulations. *Opns Res* **28**(3): 535–546.

Fowler JW, Park S, Mackulak GT and Shunk DL (2001). Efficient cycle time-throughput curve generation using fixed sample size procedure. *Int J Prod Res* **39**(12): 2595–2613.

Kelton WD (1986). Replication splitting and variance for simulating discrete-parameter stochastic processes. *Opns Res Lett* **4**(6): 275–279.

Law AM (1975). Efficient estimators for simulated queueing systems. *Mngt Sci* **22**(1): 30–41.

Law AM and Kelton WD (2000). *Simulation Modeling and Analysis*, 3rd edn, McGraw-Hill, Inc: Boston.

Leach SE (2005). *Variance-based sampling allocation method for simulating multiple traffic intensities of a cycle time-throughput curve*, Doctoral Dissertation. Arizona State University, Tempe, AZ.

Leach SE, Fowler JW, Mackulak GT, Nelson BL and Marquis JL (2005). *Asymptotic variance-based sampling for simulating cycle time-throughput curves*, Working Paper ASUIE-ORPS-2005-003, Arizona State University, Tempe.

Marse K and Roberts SD (1983). Implementing a portable FORTRAN uniform (0,1) generator. *Simulation* **41**: 135–139.

Park S, Fowler JW, Mackulak GT, Keats JB and Carlyle WM (2002). D-optimal sequential experiments for generating a simulation-based cycle time-throughput curve. *Opns Res* **50**(6): 981–990.

Schmeiser BW (1982). Batch size effects in the analysis of simulation output. *Opns Res* **30**(3): 556–568.

Whitt W (1989). Planning queueing simulations. *Mngt Sci* **35**(11): 1341–1366.

Yang F, Ankenman BE and Nelson BL (2007). Efficient generation of cycle time-throughput curves through simulation and metamodeling. *Nav Res Logist* **54**: 78–93.

Received 16 February 2007;
accepted 3 October 2007 after one revision