Regenerative Estimation Variants of Response Times in Closed Networks of Queues

PANAJOTIS KATSAROS CONSTANTINE LAZOS
Department of Informatics
Aristotle University of Thessaloniki
54006 Thessaloniki
GREECE

Abstract: - In this paper, we present a comparison of the possible regenerative estimation variants of response times, in multivariate simulations of closed queuing networks. The underlying stochastic framework of the techniques under study is first described and the applicability of each one of them is discussed. An appropriate sequential control procedure has been selected, in order to produce confidence intervals of the same nominal level and similar width, for the response times of interest. The first experimental results exhibit improved coverage of the corresponding analytic solutions, when a marked job based method is used, instead of an indirect estimation, by simulating a single regeneration sequence of the common number-in-queue process (usually used when estimating other characteristics like throughputs, utilizations, queue lengths etc). This finding clearly implies the simultaneous use of more than one regeneration sequence for multivariate studies that include response time characteristics.

Key-Words: - Queuing networks, Regenerative method, Simulation output analysis, Sequential control, Coverage analysis

1 Introduction

The regenerative method is an estimation procedure applied for analyzing the output of steady-state simulations. It is based on the assumption, that any regenerative process is probabilistically restarted at each consecutive regeneration point. Thus, observations collected from "cycles" of random lengths, determined by successive regenerative instants of the simulated process, are independent to each other and identically distributed (i.i.d.). This allows, the use of the classical central limit theorem for deriving 100(1-α)% confidence intervals for the estimated characteristics.

The i.i.d. nature of regeneration cycles gives this method a firmer theoretical underpinning than is the case for the other simulation output analysis methods. Moreover, since all cycles are probabilistic replicas of each other, the regenerative method is not tied in, with the problem of the initial transient, where the observations from the "beginning" of the simulation are not representative of the system's steady-state behavior. However, the method is not in widespread use, perhaps of the difficulty in identifying the initial regenerative system state.

As regarding queuing network simulations, an important theoretical contribution, towards this direction, was the work published by Shedler [7]. More precisely, the author considers multiclass networks of queues with priorities among job classes and Markovian job routing. An appropriate state vector for such networks is a linear job stack, an enumeration by service center and job class of all jobs. When the service times have exponential distributions (with parameters which may depend on the service center, class of job in service and state of the system), the job stack process is a continuous time Markov chain with finite state space. When the set of recurrent states of the process is irreducible, it is possible to obtain point estimates and confidence intervals for general characteristics of the steady state, by restricting a single simulation of the job stack process to the set of recurrent states. If there is a service center that sees only one job class or it is such that, jobs of the lowest priority are subject to preemption, this can be easily achieved by initializing the system at any state accessible from the state with all jobs at that center. In this case, we say that the common number-in-queue approach is applied.

In open queuing networks, where a steady-state distribution exists, the empty system is possible to be selected as an appropriate regenerative state.

For non-Markovian queuing networks, the transition dynamics of the simulated job stack process is convenient to be described by a generalized semi-Markov process (GSMP). GSMPs are a class of stochastic processes that are characterized by both a set of states and a set of events that can trigger state transitions. Corresponding to each event is a clock.

When a clock runs down to zero, this triggers a state transition. Clocks are reset randomly according to a distribution that depends on the specific event that needs to be scheduled. The next state to be visited by the GSMP is then determined stochastically according to a probability that can depend on both the previous state and the triggering event.

One method for determining regeneration time instants in GSMPs is to identify single-states. A state *s* is a single-state, if there is only one active event, when the GSMP is in state *s*. Thus, a single-state corresponds to a configuration of the job stack, such that, all jobs are at the same center with exactly one job in service. The GSMP, which is restricted to the set of all states accessible from *s*, has been proven [7] to be irreducible and to be characterized by the regenerative property.

Shedler' s job stack process provides an adequate framework for applying the classical or other regenerative estimation procedures for queuing network characteristics like throughputs, utilizations and queue lengths. However, our paper is mainly focused on simulation methods for response times, i.e. the random times for jobs to traverse specific portions of the network. Certainly, one possibility is to exploit the regenerative sequence generated, when applying the described number-in-queue approach, by modifying the selected regenerative estimators according to the Little' s rule. Such an approach would assume the response times of interest to be totally contained within the corresponding regenerative cycles, which is not true in all cases. Moreover, the experimental results reported in [2], exhibit unacceptably low levels of coverage of the corresponding analytic solutions.

In this work, we first review the theoretical background of two different alternatives, namely: the marked job and the labeled job methods. A new comparative analysis is then presented. We are not interested in an in depth study of the efficiency properties of the alternative methods, but on the quality of the results obtained by applying similar accuracy requirements on the characteristics of interest. The first experimental results show improved coverage of the corresponding analytic solutions for all the tested confidence interval widths (of the same nominal level), when the marked job method is used. This finding is clearly implying the simultaneous use of more than one regeneration sequence for multivariate studies that include response time characteristics.

2 Statistical Estimation and Sequential Control Procedures

Let us assume, our aim is to estimate the mean value of a queuing network characteristic (e.g. throughput), which is given, as a real-valued function *f* over the regenerative stochastic process $X=\{X(t)\ ;\ t \geq 0\}$

$$k(f) = \mathrm{E}[f(X)]$$

Let us also call

$$Z_k(f) = \int_{T_{k-1}}^{T_k} f(X(u)) \cdot du$$

the observation produced by the *k*th regenerative cycle. A 100α % confidence interval for *k*(*f*), after the completion of *N* regenerative cycles, is given ([1]), by

$$\left[\hat{k}(N) - \frac{s(N)\cdot F^{-1}\left(\frac{1+a}{2}\right)}{\sqrt{N}\cdot\bar{\tau}(N)}, \hat{k}(N) + \frac{s(N)\cdot F^{-1}\left(\frac{1+a}{2}\right)}{\sqrt{N}\cdot\bar{\tau}(N)}\right] \quad (1)$$

where

$\bar{\tau}(N)$ is the average cycle length and

$$\hat{k}(N) = \frac{\bar{Z}(N)}{\bar{\tau}(N)} \quad (2)$$

$$s^2(N) = s_{11}^2(N) - 2\hat{k}(N)s_{12}^2(N) + (\hat{k}(N))^2 s_{22}^2(N) \quad (3)$$

with

$$s_{11}^2(N) = \frac{1}{N-1}\sum_{k=1}^{N}(Z_k(f) - \bar{Z}(N))^2\ ,$$

$$s_{22}^2(N) = \frac{1}{N-1}\sum_{k=1}^{N}(\tau_k - \bar{\tau}(N))^2$$

$$s_{12}^2(N) = \frac{1}{N-1}\sum_{k=1}^{N}(Z_k(f) - \bar{Z}(N))(\tau_\kappa - \bar{\tau}(N))$$

Estimator' s effectiveness is usually assessed by:

- The bias, which measures the estimator's systematic deviation from the true value,
  $$Bias[\bar{X}(n)] = E[\bar{X}(n) - \mu_x] \quad (4)$$
- The variance, which measures the estimator's mean (squared) deviation from its true value; that is,
  $$\sigma^2[\bar{\mathrm{X}}(n)] = E[\{\bar{X}(n) - E[\bar{X}(n)]\}^2] \quad (5)$$
- The mean square error (MSE), defined as
  $$MSE[\bar{X}(n)] = E\{[\bar{X}(n) - \mu_x]^2\} \quad (6)$$

From (4) and (5), it is easy to derive the following,

$$MSE[\bar{X}(n)] = \{Bias[\bar{X}(n)]\}^2 + \sigma^2[\bar{\mathrm{X}}(n)]$$

The classical regenerative estimator, given in (2), is consistent, which means that it tends to the true mean with probability 1, as $N \to \infty$; however, it is not unbiased. A number of alternatives have been suggested ([1]), in an attempt to reduce the bias introduced by the estimator' s ratio form. The most important of them are:

- the Fieller estimator,

  $$\hat{k}(N)=\frac{\overline{Z}(N)\cdot\overline{\tau}(N)-q\cdot s_{12}}{[\overline{\tau}(N)]^2-q\cdot s_{22}},$$

  where $q=(z_{1-\alpha/2})^2/N$
- the Beale estimator,

  $$\hat{k}(N)=\frac{\overline{Z}(N)}{\overline{\tau}(N)}\cdot\frac{1+s_{12}/N\cdot\overline{Z}(N)\cdot\overline{\tau}(N)}{1+s_{22}/N\cdot(\overline{\tau}(N))^2}$$
- the jackknife estimator

  $$\hat{k}(N)=\frac{1}{N}\cdot\sum_{i=1}^{N}\theta_i\text{, where}$$

  $$\theta_i=N\cdot(\overline{Z}(N)/\overline{\tau}(N))-(N-1)\left(\sum\nolimits_{j\neq i}Z_j\Big/\sum\nolimits_{j\neq i}\tau_j\right)$$
- and the Tin point estimator

  $$\hat{k}(N)=\frac{\overline{Z}(N)}{\overline{\tau}(N)}\cdot\left[1+\left(\frac{s_{12}}{\overline{Z}(N)\cdot\overline{\tau}(N)}-\frac{s_{22}}{(\overline{\tau}(N))^2}\right)\cdot N^{-1}\right]$$

Respectively, different estimation procedures are being applied in deriving confidence intervals for the Fieller and the jackknife cases. Results reported in various studies, indicate that, for long runs, the Fieller - Fieller, jackknife - jackknife, Tin - classical and classical - classical confidence intervals, give accurate coverage of the parameter of interest. However, for short runs, in [1], the jackknife - jackknife approach did best, followed by the Tin - classical and classical - classical approaches, which performed about the same.

As it has been already noted, response time estimations may be easily derived, by incorporating Little' s rule into the selected regenerative estimation procedure. Thus, when the classical regenerative method is used, the estimators of (2) and (3) are modified as follows:

$$\hat{r}(N)=\frac{\overline{L}(N)}{\overline{k}(N)} \tag{7}$$

with

$\overline{L}(N)$, the observed average number of jobs in the studied portion of the network,

$\overline{k}(N)$, the observed throughput and

$$s_{11}^2(N)=\frac{1}{N-1}\sum_{k=1}^{N}(L_k(f)-\overline{L}(N))^2,$$

$$s_{22}^2(N)=\frac{1}{N-1}\sum_{k=1}^{N}(k_k-\overline{k}(N))^2$$

$$s_{12}^2(N)=\frac{1}{N-1}\sum_{k=1}^{N}(L_k(f)-\overline{L}(N))(k_\kappa-\overline{k}(N))$$

This indirect response time estimation is bound to the limitations, described in Section 1, but on the other hand, it does not make use of additional system regeneration sequences. Also, it has been applied in a number of experimental studies, like for example, those reported in [6] and its use is suggested, in other works, as a convenient implementation possibility.

Another problem, equally important for any multivariate steady-state simulation, is the run lengths, on which the parallel estimation procedures will be based. The reason is, that different queuing network characteristics behave in totally different ways and hence require radically different run lengths to generate adequate confidence intervals. A number of sequential control procedures have been suggested ([4]), for dynamically determining the appropriate sampling lengths in the course of the simulation run.

The Lavenberg and Sauer [3] sequential procedure, which has been used in the experimental results reported in this study, is a direct consequence of (1) and determines the number N of the required regenerative cycles as

$$N\geq\left(\frac{F^{-1}\left(\frac{1+a}{2}\right)}{\delta}\right)^2\cdot\left(\frac{s(l)}{\hat{k}(l)\cdot\overline{\tau}(l)}\right)^2 \tag{8}$$

where $s(l),\hat{k}(l),\overline{\tau}(l)$ are the sample estimates, after the lth cycle of the simulation experiment and δ is the half width of the confidence interval to be obtained, expressed as a % percent of the generated point estimate. As a consequence, the whole experiment terminates, when the required number N of regenerative cycles, over all the queuing network characteristics, has been achieved. An interesting finding ([2]) was the fact that (8) was quite often temporarily satisfied after a very small number of cycles, resulting in highly inaccurate results. However, this problem was overcome, by specifying minimum numbers of cycles, as initial requirements for the sampling lengths of the queuing network characteristics to be assessed.

3 Response Time Estimation with the Marked Job Methods

Let us consider a closed, multiclass queuing network with priorities among job classes and Markovian job routing. At every epoch of continuous time, each job is in exactly one class, but jobs may change class as they traverse the network. Upon completion of service at center i a job of class j goes to center k and changes to class l with probability $p_{i,j,k,l}$ where

$$P=\{p_{i,j,k,l}:(i,j),(k,l)\in C\}$$

is a given irreducible Markov matrix and $C \subseteq \{1,2,...,s\} \times \{1,2,...,c\}$ is the set of (center, class) pairs in the network. At each service center, jobs queue and receive service according to a fixed priority scheme among classes. Within a class at a center, jobs receive service according to a specific queue service discipline, e.g., first-come first-served (FCFS). For convenience (although it is not essential), we assume that only one job can receive service at a center at a time, i.e., the centers are single servers. According to a fixed procedure for each center, a job in service may or may not be preempted if another job of higher priority joins the queue at the center. Initially, we assume exponential service times, with parameters, which may depend on the service center, the class of job being served and the state of the entire network. Let $S_i(t)$ denote the class of the job receiving service at center *i* at time *t*. We set $S_i(t) = 0$ if at time *t* there is no job at center *i*. The classes of jobs served at center *i*, when expressed by order of decreasing priority, are $j_1(i),..., j_{k(i)}(i)$, each of them being an element of the set $\{1,2,..,c\}$. Let $C_{j_1}^{(i)}(t),..,C_{j_{k(i)}}^{(i)}(t)$ denote the number of jobs of the various classes served in center *i*, at time *t*. Suppose the *NJ* jobs of the network, ordered in a linear job stack, defined by the vector *Z*(*t*):

$$Z(t) = (C_{j_{k(1)}}^{(1)}(t),...,C_{j_1}^{(1)}(t), S_1(t);...;C_{j_{k(s)}}^{(s)}(t),...,C_{j_1}^{(s)}(t), S_s(t))$$

Within a class of a particular service center, jobs waiting in queue appear in the job stack in the order of their arrival at the center. For any service center that sees only one class of job, *k*(*i*)=1, we can simplify the state space by replacing $C_{j_{k(i)}}^{(i)}(t), S_i(t)$ by $Q_i(t)$, the total number of jobs at center *i*.

As it has been already noted, in Section 1, Shedler proved, that *Z*(*t*) is a continuous time Markov chain with a finite state space. Also, if there is a service center, which sees only one job class or it is such that jobs of the lowest priority class are subject to preemption, then the job stack process has a single irreducible closed set of recurrent states.

However, since our aim is not only to estimate characteristics based on the queue lengths in jobs of various classes, at the network' s service centers, the specified state variables do not suffice. More specifically, to deal with response times, it is necessary to augment the job stack process definition by introducing the concept of the marked job. We shall keep track of the position of the marked job in the network and measure its response times of interest, as the job circulates the network. The augmented job stack process is defined as the vector, *X*(*t*)=(*Z*(*t*), *N*(*t*)), $t \geq 0$, where *N*(*t*) denotes the position of the marked job from the top of the job stack. As a consequence of the assumption of the Markovian job routing and the exponential service times, the augmented job stack process *X*(*t*) is also a continuous time Markov chain with a finite state space, *E*. However, this process is not necessary to possess only one irreducible set of recurrent states.

Response times are formally defined by means of four subsets of *E*. The sets A_1, A_2 (respectively B_1, B_2) jointly define the random times at which response times for the marked job start (respectively terminate). In effect, they determine, when to start and stop the clock measuring a particular response time of the marked job. We define two sequences of random times, $\{S_j : j \geq 0\}$ and $\{T_j : j \geq 1\}$, where S_{j-1} is the start time and T_j is the termination time of the *j*th response time for the marked job. Assuming that the initial state of the process *X* is such that, a response time for the marked job begins at *t*=0, let

$$S_0 = 0,$$

$$S_j = \inf\{\tau_n \geq T_j : X(\tau_n) \in A_2, X(\tau_{n-1}) \in A_1\}, j \geq 1$$

and

$$T_j = \inf\{\tau_n \geq S_{j-1} : X(\tau_n) \in B_2, X(\tau_{n-1}) \in B_1\}, j \geq 1$$

Then, the *j*th response time for the marked job is $P_j = T_j - S_{j-1}, j \geq 1$

For response times that are complete circuits in the network, A_1=B_1, A_2=B_2 and consequently S_j=T_j for all $j \geq 1$.

Let X_n denote the state of the Markov chain *X*(*t*), when the (*n*+1)-st response time starts: X_n=$X(S_n)$, $n \geq 0$. It follows [7], that $\{X_n : n \geq 0\}$ is a discrete time Markov chain with finite state space A_2 and if there is a service center that sees only one job class or it is such that, jobs of the lowest priority are subject to preemption, then it possesses a single irreducible set of recurrent states. Also, it is proved, that the process $\{(X_n, P_{n+1}) : n \geq 0\}$ is characterized by the regenerative property in discrete time and the expected time between regeneration points is finite. Finally ([7]), the sequence of response times for any other job (as well as the sequence of response times, irrespective of job identity, in order of start or termination) converges in distribution to the same random variable as the sequence of response times for the marked job.

Thus, if a regenerative simulation using the marked job method is to be performed, the augmented job stack process is set to a recurrent state *s*, such that a response time for the marked job starts. For each cycle, the response times for the marked job are accumulated and an appropriate statistical estimation procedure, like those described in Section 2, is applied.

The results described so far, have been also extended for networks with multiple job types. In such networks, the type of a job may influence its routing through the network as well as its service requirements at each center. Each job type has its own routing matrix and for each center, there is a priority ordering of the (type, class) pairs served at the center. It has been proved [7], that the set of recurrent states of the job stack process is irreducible, provided that the routing matrix for each job type is irreducible and there is a service center, which sees only one class of jobs. Finally, for a network with at least two service centers, the same conditions ensure that the augmented job stack process has a single irreducible closed set of recurrent states.

The applicability of the marked job method in non-Markovian networks is also proved, by employing the GSMP formalism in an analogous manner.

The labeled jobs method is another regenerative variant based on the simulation of the so-called fully augmented job stack process, which, in addition to the enumeration of all jobs by service center and job class, maintains also the position of each one of the jobs in the linear job stack. This procedure takes into account the response time observations of all jobs, thus resulting in statistically more efficient estimations. However, it is important to note, that the labeled jobs method is not applicable when estimating the complete circuit response time, a case, where only the marked job method is possible to be used.

4 Multivariate Response Time Simulation: experimental results

A closed form queuing network (Figure 1) was selected, for comparing the quality of the results obtained when applying the indirect and the marked job estimation procedures under the same accuracy requirements and nominal levels of confidence.

Figure 4 presents the results obtained in a sample run. It is important to note, that cycles performed, in each case of the marked job estimation, represent radically different lengths, than in the case of the number-in-queue simulation.

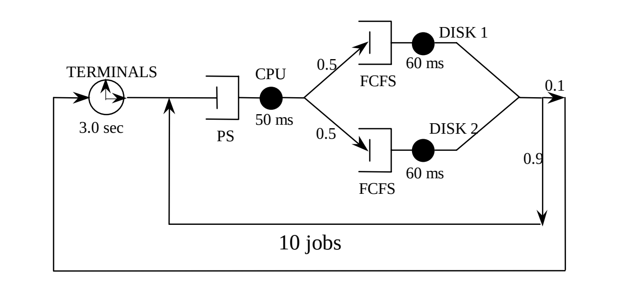

Figure 1 A central server model with terminals

Analytic results (MVA algorithm)

RESOURCE	RESPONSE TIME	THROUGHPUT	QUEUE LENGTH	UTILIZATION
CPU	0,17	17,2	2,91	0,861
DISK1	0,11	8,61	0,96	0,517
DISK2	0,11	8,61	0,96	0,517
TERMINALS	3	1,72	5,17	0,517

Figure 2 Analytic results for the central server model

Experiment description

Random Number Generator:	Mersenne Twister GSFR ([5])
Estimation procedure:	Classical regenerative estimation
Sequential control procedure:	Lavenberg & Sauer ([3])
Minimum number of cycles:	16
Level of confidence:	90%
Regeneration states:	(10, 0, 0, 0) for the number-in-queue process (10, 0, 0, 0, 1) for the TERMINALS response time (marked job estimation) (9, 1, 0, 0, 10) for the CPU response time (marked job estimation) (9, 0, 1, 0, 10) for the DISK1 response time (marked job estimation) (9, 0, 0, 1, 10) for the DISK2 response time (marked job estimation)
Tested half width c.i. cases:	3.5% 3.0% 2.5% 2.0%

Figure 3 Experiment description

RESOURCE	UTILIZATION	THROUGHPUT	QUEUE LENGTH
	TERMINALS		
ReqCIL	2.5 %	2.5 %	2.5 %
ActCIL	+/- 2.5 %	+/- 2.5 %	+/- 2.5 %
CYCLES	60	42	60
LBOUND	0.49654	1.6579	4.9654
MEAN	0.50919	1.7001	5.0919
UBOUND	0.52185	1.7422	5.2185
	CPU		
ReqCIL	2.5 %	2.5 %	2.5 %
ActCIL	+/- 2.4 %	+/- 2.3 %	+/- 2.5 %
CYCLES	22	16	160
LBOUND	0.83062	17.091	2.8729
MEAN	0.85141	17.488	2.9462
UBOUND	0.87219	17.885	3.0195
	DISK1		
ReqCIL	2.5 %	2.5 %	2.5 %
ActCIL	+/- 2.5 %	+/- 2.2 %	+/- 2.4 %
CYCLES	33	33	93
LBOUND	0.50039	8.4319	0.94125
MEAN	0.51303	8.6206	0.96427
UBOUND	0.52567	8.8093	0.98729

	DISK2		
ReqCIL	2.5 %	2.5 %	2.5 %
ActCIL	+/- 2.5 %	+/- 2.3 %	+/- 2.5 %
CYCLES	48	26	171
LBOUND	0.50965	8.4611	0.93476
MEAN	0.52257	8.6588	0.95854
UBOUND	0.53549	8.8564	0.98233

RESOURCE	INDIRECT RES TIME	MARKED JOB RES TIME
	TERMINALS	
ReqCIL	2.5 %	2.5 %
ActCIL	+/- 2.4 %	+/- 2.5 %
CYCLES	60	97
LBOUND	2.9194	2.9529
MEAN	2.9913	3.0284
UBOUND	3.0633	3.104
	CPU	
ReqCIL	2.5 %	2.5 %
ActCIL	+/- 2.5 %	+/- 2.5 %
CYCLES	116	354
LBOUND	0.16693	0.16431
MEAN	0.17118	0.16849
UBOUND	0.17544	0.17268
	DISK1	
ReqCIL	2.5 %	2.5 %
ActCIL	+/- 2.5 %	+/- 2.2 %
CYCLES	32	20
LBOUND	0.10414	0.10878
MEAN	0.10676	0.11118
UBOUND	0.10938	0.11358
	DISK2	
ReqCIL	2.5 %	2.5 %
ActCIL	+/- 2.5 %	+/- 2.3 %
CYCLES	99	32
LBOUND	0.10946	0.10717
MEAN	0.11223	0.10969
UBOUND	0.11501	0.11222

Figure 4 Simulation results (initial seed = 1160)

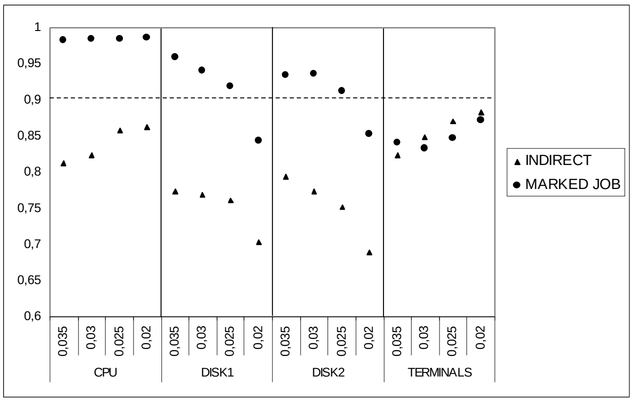

Figure 5 Observed coverage (1000 runs/case)

Figure 5 shows the coverage values obtained from 1000 sample runs for each c.i. half width. It is important to note, that the sequential procedure' s half width requirement is considered to be valid, in respect to the observed variability, if the obtained coverage is close to or higher than the chosen nominal level.

5 Conclusion

In this work, we employed two different response times estimation variants, based on different regeneration sequences. The obtained results show substantial coverage improvements. This is due to impressive variance reductions, in cases where the marked job regeneration sequences were used. This finding implies the simultaneous use of more than one regeneration sequence for multivariate studies that include response time characteristics. Application of the labeled jobs regenerative variant will improve the experiments' statistical efficiency.

References:

[1] D. L. Iglehart, The regenerative method for simulation analysis, In *Current Trends in Programming Methodology*, Vol. III, Software Modeling, K. M. Chandy and P. T. Yeh, Eds., Prentice Hall, 1978, pp. 52-71

[2] P. Katsaros and C. Lazos, Shared memory parallel regenerative queuing network simulation, In *Proceedings of the 15th European Simulation Multiconference*, The Society for Computer Simulation, Prague, 2001, pp. 736-740

[3] S. S. Lavenberg and C. H. Sauer, Sequential stopping rules for the regenerative method of simulation, *IBM Journal of Research and Development*, Vol. 21, 1977, pp. 545-558

[4] A. M. Law and W. D. Kelton, Confidence intervals for steady-state simulations, II: A survey of sequential procedures, *Management Science*, Vol. 28, No. 5, 1982, pp. 550-562

[5] M. Matsumoto and T. Nishimura, Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator, *ACM Transactions on Modeling and Computer Simulation*, Vol. 8, 1998, pp. 3-30

[6] C. H. Sauer and K. M. Chandy, *Computer systems performance modeling*, Prentice Hall, 1981

[7] G. Shedler, *Regenerative stochastic simulation*, Academic Press, 1993