Use of the Diffusion Approximation to Estimate Run Length in Simulation Experiments

T. Moeller and H. Kobayashi, New York

ABSTRACT

This paper presents an application of the diffusion process approximation to the statistical analysis involved in simulation experiments of queueing systems. The autocovariance function of the queue size process for both a GI/G/1 queue and a cyclic queueing system are obtained. Run length is predicted and the variance (hence confidence intervals) of estimates of performance such as average queue size, queue size distribution, and server utilization are computed using the approximate autocovariance function. These techniques are used in a simulation experiment of a cyclic queueing model of a multi-programmed computer system.

1. INTRODUCTION

The analysis of outputs of simulations of queueing processes is often made difficult because of the high degree of serial correlation in the output time series. If the autocorrelation function is positive over some range of time as is the case'in many queueing processes, a variance estimate made on the assumption of independent and identically distributed outputs may be a serious underestimation of its true value.

In order to overcome such a difficulty, a number of techniques have been developed. The method of blocking of the output samples is the most widely used technique to achieve nearly independent sample outputs. Also commonly used is the method of independent replications of the experiment.

FISHMAN [1971] has proposed a method in which regression analysis is performed during the simulation to dynamically estimate the autocorrelation in the output. This estimate is then used to determine when to halt the simulation. CRANE and IGLEHART [1974] recently discuss the concept of regeneration cycles to obtain groupings of the output which are independent and identically distributed, whereby the variance of the sample means can be calculated in a straightforward manner. Other methods of analysis of variance which take the autocorrelation into account are given in FISHMAN and KIVIAT [1967].

An exact expression for the autocorrelation function of a queueing process, however, is known only for the M/M/1 case, MORSE [1955], who obtained the solution as an infinite series of modified Bessel functions.

For simultations of queueing systems we propose here the use of approximate analytical solutions. The diffusion approximation of the queue size process of a GI/G/1 system and a two stage cyclic queueing system is used to estimate the autocovariance functions of the processes. They, in turn, are used to calculate the run length required to achieve a prespecified value for the variance of the sample mean of the quantity one wishes to estimate.

2. GI/G/1 QUEUEING SYSTEM

Consider a GI/G/1 system with FCFS (first-come, first-served) queue discipline. Let us assume that the interarrival times and service times are both independently and identically distributed with means and variances given by (μ_a,μ_s) and (σ_a^2,σ_s^2), respectively.

Let Q(t) represent the number of customers in the system (i.e., those in service or in queue) at time t. A realization of Q(t) is a random step function with vertical jumps of magnitude one at instants of customer arrivals departures from the system.

If the traffic density $\rho=\mu_a/\mu_s$ is close to unity, then the server is rarely idle, i.e., Q(t) is seldom near the barrier Q=0 and it is well justified to replace Q(t) with an approximate continuous process X(t), t>0. In this case, X(t) is a diffusion process with a reflecting barrier at X=0. For a discussion of the validity of this approximation, see COX and MILLER [1965], KOBAYASHI [1974], and NEWELL [1971].

For a stochastic process X(t) which approximates the queue size process of a GI/G/1 system, the conditional probability density function $p(X_0,X:t)$ of X(t) given that $X(0) = X_0$, satisfies the Fokker-Planck equation:

$$\frac{\partial}{\partial t} p(X_0,X;t) = \frac{\alpha}{2} \frac{\partial^2}{\partial X^2} p(X_0,X;t) - \beta \frac{\partial}{\partial X} p(X_0,X;t) \qquad (1)$$

with the boundary conditions

$$\frac{\alpha}{2} \frac{\partial}{\partial X} p(X_0,X;t) - \beta\, p(X_0,X;t) = 0 \text{ at } X = 0 \qquad (2)$$

where

$$\alpha = \frac{C_a}{\mu_a} + \frac{C_s}{\mu_s}\,,\quad \beta = \frac{1}{\mu_a} - \frac{1}{\mu_s}$$

and

$$C_a = \frac{\sigma_a^2}{\mu_a^2}\,,\quad C_s = \frac{\sigma_s^2}{\mu_s^2}\,. \qquad (3)$$

The solution to Eqn. (1) is given by

$$p(X_0,X;t) = \frac{\partial}{\partial X} \{\Phi(\frac{X-X_0-\beta t}{\alpha t}) - e^{2\frac{\beta}{\alpha}X} \Phi(\frac{X+X_0+\beta t}{\alpha t})\} \qquad (4)$$

where $\Phi(\cdot)$ is the integral of the unit normal distribution, i.e.,

$$\Phi(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} e^{-t^2/2}\, dt. \qquad (5)$$

Using the property that the process X(t) is, in equilibrium, covariance stationary, the autocovariance function of a queue size process Q(t) can be approximated by

$$R(t) = \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} [k - E\{X\}][m - E\{X\}]\ \hat{p}(k)\hat{p}(m,k;t) \qquad (6)$$

where $\hat{p}(m,k,t)$ is the discrete conditional probability density function corresponding to $p(m,X;t)$ and is given by

$$\hat{p}(m,k;t) = \int_{k}^{k+1} p(m,\xi;t)\,d\xi,\quad k=0,1,\ldots \qquad (7)$$

and

$$\hat{p}(k) = \lim_{t\to\infty} \hat{p}(m,k;t). \qquad (8)$$

3. CYCLIC QUEUEING SYSTEM

Let $Q_1(t)$ represent the queue size of the first server of a cyclic queueing system, shown in Fig. 1. The number of jobs N remains constant and the jobs circulate in a closed loop of two servers. Each server follows the FCFS (first-come, first-served) discipline and when service is completed at one server, the job instantaneously moves to the queue of the other server. If the service process at each server is governed by a general probability distribution with means (μ_1,μ_2) and variances (σ_1^2,σ_2^2), and then the $Q_1(t)$ process is a discrete-valued random process with reflecting barriers at $Q_1=0$ and $Q_1=N$.

If the number of jobs N in the system is sufficiently large and the ratio μ_1/μ_2 is close to unity, then $Q_1(t)$ seldom reaches the barrier and it is appropriate to approximate $Q_1(t)$ by a diffusion process q(t). We then derive a normalized diffusion process $y(\tau)$ according to the transformations

$$\tau = t \Big/ \sqrt{\frac{\mu_1(C_1 + C_2\rho)}{(1-\rho)^2}} \tag{9}$$

and

$$y = q \Big/ \sqrt{\frac{C_1 + C_2\rho}{1 - \rho}} \ , \tag{10}$$

and we obtain the corresponding Fokker-Planck equation in KOBAYASHI [1974]:

$$\frac{\partial}{\partial\tau} p(y_0,y;\tau) = \frac{1}{2} \frac{\partial^2}{\partial y^2} p(y_0,y;\tau) - \frac{\partial}{\partial y} p(y_0,y;\tau) \tag{11}$$

with boundary conditions

$$\frac{1}{2} \frac{\partial}{\partial y} p(y_0,t;\tau) - p(y_0,y;\tau) = 0 \text{ at } y = 0 \text{ and } y = b \tag{12}$$

where those parameters which appeared in Eqns. (9) - (12) are defined by

$$b = \left|\frac{N+1}{\frac{C_1 + C_2\rho}{1-\rho}}\right| , \quad C_1 = \frac{\sigma_1^2}{\mu_1} , \quad C_2 = \frac{\sigma_2^2}{\mu_2}$$
$$\rho = \frac{\mu_1}{\mu_2} < 1. \tag{13}$$

Note that τ represents a normalized time and y, the normalized queue size.

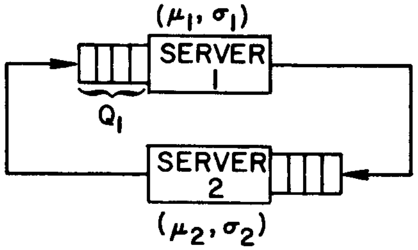

Fig. 1 A two-stage cyclic queueing system

Using the conditional probability density function $p(y_0,y;\tau)$ and its steady state probability density function $p(y) = \lim_{\tau\to\infty} p(y_0,y;\tau)$, the autocovariance function for the normalized queue size process is given by

$$R(\tau) = \int_0^b [y(s) - E\{y\}]p(y(s))dy(s)$$
$$\int_0^b [y(s+\tau) - E\{y\}]\cdot p(y(s+\tau), y(s);\tau)\, dy(s+\tau) \tag{14}$$

A solution to the boundary value problem given by Eqns. (9) and (10) is given in SWEET and HARDIN [1970]

$$p(y_0,y;\tau) = \frac{2e^{2y}}{e^{2b}-1} + e^{y-y_0} e^{-\frac{\tau}{2}} \sum_{n=1}^{\infty} \phi_n(y_0)\phi_n(y)e^{-\frac{\lambda_n^2\tau}{2}} \tag{15}$$

with

$$\phi_n(y) = \sqrt{\frac{2\ \lambda_n^2}{b(\lambda_n^2+1)}} \{ \cos \lambda_n y + \frac{1}{\lambda_n} \sin \lambda_n y\}$$
$$\lambda_n = \frac{n\pi}{b} , \quad n = 1,2,... \tag{16}$$

Using an approximate solution ıor $p(y_0,y;\tau)$ defined by Eqn. (15) in the expression for $R(\tau)$ yields an estimate of the autocovariance function

$$R(\tau) = \sum_{n=1}^{\infty} a_n^2 \, e^{-\frac{(\lambda_n^2+1)\tau}{2}} \qquad (17)$$

where

$$a_n = \left[\frac{2}{e^{2b}-1}\right]^{\frac{1}{2}} \int_0^b \xi \, e^{\xi} \, \phi_n(\xi) \, d\xi \; . \qquad (18)$$

3.1 Mean Value of Queue Size

The sample mean of queue size is given for an observation of length T by

$$\bar{y} = \frac{1}{T} \int_0^T y(t)dt \; . \qquad (19)$$

The variance of $\bar{y}$ may be used to estimate how close the sample mean is to its population mean E{y}. The sample covariance is given in terms of the autocovariance $R(\tau)$ as

$$\mathrm{var}(\bar{y}) = \frac{1}{T} \int_{-T}^{T} \left(1 - \frac{|\tau|}{T}\right) R(\tau) \; d\tau \; . \qquad (20)$$

Since $R(\tau)$ given by Eqn. (17) approaches zero sufficiently fast as $\tau \to \infty$, we obtain the following asymptotic result

$$\lim_{T\to\infty} \int_{-T}^{T} \left(1 - \frac{|\tau|}{T}\right) R(\tau) \; d\tau = 2 \int_0^{\infty} R(\tau)d\tau \; . \qquad (21)$$

For the case where $R(\tau)$ is approximated by Eqn. (17) the assumption for Eqn. (21) holds and thus

$$\mathrm{var}(\bar{y}) = \frac{1}{T} \sum_{n=1}^{\infty} \frac{4 \, a_n^2}{\lambda_n^2+1} \qquad (22)$$

Thus, an estimate of simulation run length T* required to obtain a specified level of variance V* in the sample mean is

$$T^* = \frac{1}{V^*} \sum_{n=1}^{\infty} \frac{4 \, a_n^2}{\lambda_n^2+1} \qquad (23)$$

3.2 Queue Size Distribution Function

Consider an estimation of the distribution function for the queue size process y(t) of the cyclic queueing system. The first-order distribution function of y(t) is given by

$$F(y) = \Pr\{y(t) \leq y\} \; , \; 0 \leq y \leq b \, . \qquad (24)$$

For the process y(t) define a 1-0 valued process $Z_y(t)$ by

$$Z_y(\tau) = \begin{cases} 1 \text{ if } y(\tau) \leq y \\ 0 \text{ if } y(\tau) > y. \end{cases} \qquad (25)$$

The expected value of Z_y is thus the value of the distribution function $F(\cdot)$ at point y:

$$E\{Z_y(t)\} = \Pr\{y(t) \leq y\} = F(y) \, . \qquad (26)$$

For a simulation of run length T, define the sample estimate of the distribution function F(y) by

$$\bar{F}(y) = \frac{1}{T} \int_0^T Z_y(t) \; dt \, . \qquad (27)$$

The variance of this sample value is given by

$$\mathrm{var}\{\bar{F}(y)\} = \frac{1}{T} \int_{-T}^{T} \left(1 - \frac{|\tau|}{T}\right)[G(y,\tau) - F^2(y)]d\tau \qquad (28)$$

where

$$\begin{aligned} G(y,\tau) &= E\{Z_y(t+\tau)Z_y(t)\} \\ &= \Pr\{X(t+\tau) \leq y, \;\; X(t) \leq y\} \qquad (29) \\ &= \int_0^y p(\xi) \int_0^y p(\xi,\eta;\tau) \; d\eta \; d\xi \; . \end{aligned}$$

Since $R(\tau) \to 0$ sufficiently fast as $\tau \to \infty$, we may write for sufficiently large T

$$\mathrm{var}\{\bar{F}(y)\} = \frac{2}{T} \int_0^{\infty} [G(y,\tau) - F^2(y)]d\tau \; . \qquad (30)$$

Using the diffusion process approximation for $p(\xi,\eta;\tau)$ we have an estimate for the variance of $\bar{F}(y)$ given by

$$\mathrm{var}\{\bar{F}(y)\} = \frac{1}{T} \sum_{n=1}^{\infty} \frac{4 \, b_n^2(y)}{\lambda_n^2 + 1} \qquad (31)$$

where

$$b_n(y) = \left[\frac{2}{e^{2b}-1}\right]^{\frac{1}{2}} \int_0^y e^{\xi} \phi_n(\xi) \; d\xi \; . \qquad (32)$$

Using Eqn. (31) an estimate of run length T_y^* may be obtained for the desired level of confidence V_y^* in the variance of the sample distribution point $\bar{F}(y)$:

$$T_y^* = \frac{1}{V_y^*} \sum_{n=1}^{\infty} \frac{4\, b_n^{\,2}(y)}{\lambda_n^{\,2} + 1} \, . \qquad (33)$$

3.3 Simulation Experiment

As an example of the run length and confidence interval prediction techniques we take a cyclic queue model (Fig. 1) of a multiprogrammed computer system. Programs (jobs), in the system wait for service at the CPU (server 1), then after an I/O request they wait for service at the I/O device (server 2). It is assumed that there is a constant number of jobs N in the system and that the CPU has an exponential service time distribution and the I/O device has a five stage erlang service time distribution. For this example, $N=10$, $\rho=\mu_1/\mu_2=.8$ and $\mu_1=4.0$, $\mu_2=5.0$, $\sigma_1=4.0$, $\sigma_2=\sqrt{5}$. Since $\bar{q} = ((C_1 + C_2\rho)/(1-\rho))\bar{y}$ (where $\bar{y}$ is given by Eqn. (19)) is essentially a linear sum of a random variables $y(t)$, $0 \leq t \leq T$, the random variable $\bar{q}$ is asymptotically (i.e., as $T\to\infty$) normally distributed. Thus a confidence interval for $\bar{q}$ for sufficiently large T can be derived using the relation

$$\Pr\left(|\bar{q} - E\{q\}| \leq L \cdot \sqrt{(\mathrm{var}\ \bar{q})} \right) \geq 1 - C \qquad (34)$$

with L such that $\Phi(L) = \frac{C}{2}$, $0 \leq C \leq 1$. Thus the confidence interval is given by $2\,L\sqrt{(\mathrm{var}\ \bar{y})}$ and the confidence level is $100(1-C)\%$.

The simulation was programmed using the SIMPL/I language [IBM, 1972] and run on a IBM 370/158 computer. The experiment is initialized by placing five jobs in each queue and beginning service to the first job in each queue at time zero. The stopping condition was determined from the run length estimate given by Eqn. (23) together with a variance of sample mean such that the confidence interval is .2 and the confidence level is 90%.

The sample mean queue size for the first server and the sample queue size distribution $\{\bar{F}(q)\}_{q=1}^{10}$ are generated from the simulation, and confidence intervals are calculated for $\{\bar{F}(q)\}_{q=1}^{10}$ based on the run length $T^* = 62{,}180$ (in the same units as μ_1) and using Eqn. (31) and Eqn. (34). Table 1 summarizes the results of the simulation.

Of special interest is the fact that the sample value of utilization for the CPU (server 1) is given by $1 - \bar{F}(0)$. The confidence interval for this sample utilization is therefore given by the confidence interval for $\bar{F}(0)$.

$\bar{q} = 1.92$, 90% confidence interval of .2

With $\bar{q}$ the sample mean queue size

q	$\bar{F}(q)$	δ_i	$[\bar{F}-\delta_i,\ \bar{F}+\delta_i]$
0	.245	.002	[.243, .247]
1	.530	.005	[.525, .535]
2	.712	.008	[.704, .720]
3	.826	.011	[.815, .837]
4	.895	.014	[.881, .909]
5	.938	.016	[.922, .959]
6	.965	.017	[.948, .982]
7	.981	.016	[.965, .997]
8	.992	.014	[.972, 1.000]
9	.997	.008	[.887, 1.000]
10	1.000	0	---

Table 1 Estimates of Queue Size Distribution Function For Cyclic Queue System

($N=10$, $\rho=.8$, Run Length = 62,180)

. CONCLUSIONS

The techniques of run length prediction or confidence interval prediction may be applied to other systems for which an expression for the autocovariance function is known. Here a cyclic queueing system and a GI/G/1 were studied. The estimates for run length and confidence intervals take into account the serial correlation of the output time series. An extension of the method to a simulation of a more general queueing network is currently being investigated.

REFERENCES

[1] Cox, D., Miller, Theory of Stochastic Processes, John Wiley and Sons, New York, 1965.

[2] Crane, M., and Iglehart, D., "Simulating Stable Stochastic System I: General Multi-server Queues", JACM, Vol. 21, No. 1, pp. 103-113, 1974.

[3] Fishman, G., "Estimating Sample Size in Computing Simulation Experiments", Management Sci., Vol. 18, No. 1, pp. 21-38, 1971.

[4] Fishman, G, and Kiviat, P., "The Analysis of Simulation Generated Time Series", Management Sci., Vol. 13, No. 7, pp. 525-557, 1967.

[5] IBM, Simpl/I Program Reference Manual, SH 19-5060-0, White Plains, New York, 1972.

[6] Kobayashi, H., "Application of the Diffusion Approximation to Queueing Networks, Part II", to appear in JACM, July 1974.

[7] Morse, P, "Stochastic Properties of Waiting Lines", J. Op. Res. Soc. Am., Vol. 3, No. 3, pp. 255-261, 1955.

[8] Newell, G, Applications of Queueing Theory, Chapman and Hall Ltd., London, 1971.

[9] Sweet, A. L., and Hardin, J. C., "Solutions for Some Diffusion Processes with Two Barriers", J. Appl. Prob., Vol. 7, pp. 423-431, 1970.