SIMULATION RUN LENGTH PLANNING

Ward Whitt
AT&T Bell Laboratories
Murray Hill, NJ 07974, U.S.A.

ABSTRACT

To design a stochastic simulation experiment, it is helpful to have an estimate of the simulation run lengths required to achieve desired statistical precision. Preliminary estimates of required run lengths can be obtained by approximating the stochastic model of interest by a more elementary Markov model that can be analyzed analytically. When steady-state quantities are to be estimated by sample means, we often can estimate required run lengths by calculating the asymptotic variance and the asymptotic bias of the sample mean in the Markov model.

1. INTRODUCTION

Simulation experiments are like exploring trips. We usually have initial goals, but the interesting discoveries often come from the unexpected. We typically do not know in advance precisely how we will proceed and we cannot predict all the benefits. In reality, most simulation experiments are sequences of experiments, with new goals being based on successive discoveries; see Albin (1984). Thus, there obviously is a limit to what can be planned; nevertheless simulation planning is our concern. One reason is that we have to expect something before we can recognize the unexpected.

The purpose of this paper is not so much to suggest that we look before we leap, but to suggest a few things that we might look at. The context is a simulation of a stochastic model to estimate steady-state quantities of interest. Our idea is to develop some expectations and design the initial experiment by doing some preliminary mathematical analysis. We focus on simulation run lengths.

Of course, since we are going to simulate, the stochastic model of interest presumably is relatively complicated, so that it is not easy to calculate the quantities of interest analytically. Thus, we suggest approximating the stochastic model of interest by a more elementary stochastic model that can be analyzed analytically. For the approximating model, in addition to the steady-state quantity of interest, we calculate the asymptotic variance and the asymptotic bias of the sample mean used to estimate the steady-state quantity of interest (assuming that a sample mean will be used). Then we apply these quantities to estimate the simulation run lengths required in the original model to obtain desired statistical precision. The estimated simulation run lengths can then be used, before any data have been collected, to design the experiment, i.e., to determine what cases to consider, what statistical precision to aim for, what experimental budget is appropriate, or even whether to conduct the experiment at all.

There are two steps: First, we must find a suitable approximation for the given stochastic model and, second, we must calculate the asymptotic quantities of interest for the approximating model. Of course, we do not want the preliminary analysis to be harder than doing the experiment itself. The preliminary analysis better be easy or we wouldn't bother with it. Fortunately, there is substantial literature supporting these two steps.

In Whitt (1989a) we show how this program can be carried out for a large class of stochastic models. The models considered are those for which the stochastic process of interest can be approximated by reflecting Brownian motion (RBM). Through much previous work on heavy-traffic limit theorems and diffusion approximations, it is known that many queueing processes can be approximated by RBM, at least roughly, so that the class of models covered is relatively large. The class includes the standard GI/G/1 queue, as was shown previously in related work by Blomqvist (1969), Moeller and Kobayashi (1974) and Woodside, Pagurek and Newell (1980), but also applies to many other models. Similar ideas are expressed in Chapter 5 of Newell (1982); e.g., p. 151. In Whitt (1989a) we also show how to calculate the asymptotic quantities of interest for RBM to obtain very simple approximate formulas. Moreover, we show that the scaling of time in the heavy-traffic limit theorem plays an essential role in determining the form of the final formulas.

Of course, RBM is by no means the only stochastic process that can be used as an approximation. For example, the Ornstein-Uhlenbeck diffusion process is a natural candidate for infinite-server queues, which also yields very simple formulas. In Whitt (1989b) we apply recurrent potential theory for Markov processes to obtain asymptotic formulas for a large class of Markov processes, including general birth and death processes and diffusion processes. In Heidelberger and Whitt (1989) we compare the asymptotic formulas, and thus the simulation efficiency, for several different queueing models. There we show that it usually is easier to obtain reliable estimates for an infinite-server queue than for a single-server queue. It usually is even easier for small closed queueing networks. Roughly speaking, simulation efficiency increases (estimation becomes easier) as the relaxation time (the time required to reach steady state; see Keilson (1979)) decreases.

Our purpose here is to provide a brief overview. In Section 2 we review the standard statistical theory leading up to the large-sample formula for the required simulation run length with the relative width criterion; i.e., the run length such that the width of the confidence interval divided by the quantity being estimated (which is presumed to be positive) takes some prescribed value. In Section 3 we review some of the RBM-type examples from Whitt (1989a), including the M/M/1 queue, RBM, the GI/G/m queue and a packet queue model from Fendick, Saksena and Whitt (1989). The packet queue model is relatively complicated, so that an exact analysis is evidently not possible with current methodology. However, simple formulas to determine appropriate simulation run lengths can be obtained from an RBM approximation. These examples show that it can be very misleading to do an exploratory simulation with one set of parameter values to determine appropriate simulation run lengths, because the appropriate run lengths are very different for different traffic intensities. Finally, in Section 4 we review the asymptotic formulas for Markov processes, including the explicit formulas for birth and death processes and diffusion processes in Whitt (1989b).

2. STATISTICAL BACKGROUND

Let $X(t)$ be the stochastic process of interest, which we assume converges in distribution to a proper limit $X(\infty)$ as $t \rightarrow \infty$. To represent various steady-state quantities of interest, consider $Y(t) = f(X(t))$ where f is a real-valued function on the state space of $X(t)$. For example, if f is the indicator function of the interval $(-\infty, x]$, then $Y(t) = P(X(t) \leq x)$; if $f(x) = x^2$, then $Y(t) = X(t)^2$.

Suppose that the *steady-state mean* $\bar{f} \equiv E\,Y(\infty) \equiv Ef\,(X(\infty))$ is to be estimated by the *sample mean* $\bar{Y}(t) = t^{-1} \int_0^t Y(s)\,ds$. The standard statistical analysis is based on the *central limit assumption* for $\bar{Y}(t)$ as $t \rightarrow \infty$; i.e., the assumption that $t^{1/2}\,(\bar{Y}(t) - \bar{f}\,)$ converges in distribution to a normal distribution with mean 0 and variance $\bar{\sigma}^2$ as $t \rightarrow \infty$. (Necessarily $\bar{Y}(t)$ converges in probability to $\bar{f}$ as well.) Typically (under extra uniform integrability), t Var $(\bar{Y}(t))$ converges to $\bar{\sigma}^2$ as $t \rightarrow \infty$ too; hence we call $\bar{\sigma}^2$ the *asymptotic variance of the sample mean.* Throughout this paper we assume that these limits are well defined. (For some relevant existence theory, see Glynn (1989).)

Hence, for sufficiently large t (which depends on the model and the function f), it is almost always appropriate to regard $\bar{Y}(t)$ as being approximately normally distributed with mean $\bar{f}$ and variance $\bar{\sigma}^2/t$. However, we typically cannot directly calculate $\bar{f}$ and $\bar{\sigma}^2$. Moreover, there is the question of how large t needs to be before the normal approximation is justified, which we will not discuss here; see Asmussen (1989, 1980). We presume that the value of t required to obtain desired statistical precision under the normal approximation is large enough. We do assume that t will indeed be relatively large, so that the large sample theory applies.

Based on the normal approximation for $\bar{Y}(t)$, a $(1-\beta)(100)\,\%$ *confidence interval* for $\bar{f}$ is

$$[\bar{Y}(t) - z_{\beta/2}\,(\bar{\sigma}^2/t)^{1/2}\;,\;\; \bar{Y}(t) + z_{\beta/2}\,(\bar{\sigma}^2/t)^{1/2}] \quad (1)$$

where $\Phi(z_{\beta/2}) - \Phi(-z_{\beta/2}) = 1-\beta$ with Φ being the standard (mean 0, variance 1) normal cumulative distribution function. The *relative width* of the confidence interval (1) is then

$$w_r(\beta) = \frac{2\,\bar{\sigma}\,z_{\beta/2}}{t^{1/2}\,\bar{f}}\;. \quad (2)$$

For (2) to be meaningful, we assume that $\bar{f} > 0$. From (2) we see that the *required simulation run length* for specified relative width ε and level of precision β is

$$t_r(\varepsilon, \beta) = \frac{4\,\bar{\sigma}^2\,z_{\beta/2}^2}{\varepsilon^2\,\bar{f}^2}\;. \quad (3)$$

Hence, with the relative width criterion, the required run length is proportional to $\bar{\sigma}^2/\bar{f}^2$, which we call the (relative width) *run-length ratio.*

To properly interpret the run-length ratio $\bar\sigma^2/\bar f^2$, recall that $\bar\sigma^2$ is typically much greater than the steady-state variance $\sigma_\infty^2 = \mathrm{Var}\,(Y(\infty))$ due to positive correlations in the stochastic process $Y(t)$. Indeed, we often find it helpful to represent the run-length ratio as the product of two factors, by setting

$$\frac{\bar\sigma^2}{\bar f^2} = \left[\frac{\bar\sigma^2}{\sigma_\infty^2}\right]\left[\frac{\sigma_\infty^2}{\bar f^2}\right] . \tag{4}$$

We call the first factor $(\bar\sigma^2/\sigma_\infty^2)$ the *correlation factor*, because it describes the effect of the correlations over time; we call the second factor $\sigma_\infty^2/\bar f^2$ (which is the squared coefficient of variation of $Y(\infty)$) the *steady-state variability factor* because it describes the effect of the variability of the steady-state distribution.

So far we have not mentioned bias, but since we presumably cannot start simulating in steady-state, $E\,\bar Y(t) \neq \bar f$. However, typically $t(E\,\bar Y(t) - \bar f\,)$ converges as $t \to \infty$ to a finite limit $\bar\beta$, which we call the *asymptotic bias of the sample mean*. Thus, for sufficiently large t, the *bias* $(E\,\bar Y(t) - \bar f\,)$ is approximately $\bar\beta/t$ and the relative bias $(E\,\bar Y(t)/\bar f\,) - 1$ is approximately $\bar\beta/\bar f\,t$. Since the relative bias is of order t^{-1}, the relative bias is asymptotically negligible compared to the relative width of the confidence interval in (2) as $t \to \infty$. However, to determine whether the bias can reasonably be ignored, it is helpful to approximate $\bar\beta$ as well as $\bar f$ and $\bar\sigma^2$.

3. EXAMPLES

In this section we present some examples, drawing on Whitt (1989a) and references cited there.

3.1 The M/M/1 queue

For the queue length process (including the customer in service) in the M/M/1 queue with service rate 1 and arrival rate (and traffic intensity) ρ, if we consider $f\,(k) = k$, then

$$\bar f = \frac{\rho}{1-\rho}\ , \quad \sigma_\infty^2 = \frac{\rho}{(1-\rho)^2}\ , \quad \bar\sigma^2 = \frac{2\rho(1+\rho)}{(1-\rho)^4} \tag{5}$$

see Section 3.2 of Whitt (1989a), so that the run-length ratio is

$$\frac{\bar\sigma^2}{\bar f^2} = \frac{2(1+\rho)}{\rho(1-\rho)^2}\ , \tag{6}$$

which approaches $+\infty$ as $\rho \to 0$ and as $\rho \to 1$. Note that the steady-state variability factor $\sigma_\infty^2/\bar f^2 = 1/\rho$ produces all of the light-traffic explosion but none of the heavy-traffic explosion, whereas the correlation factor $\bar\sigma^2/\sigma_\infty^2 = 2(1+\rho)/(1-\rho)^2$ produces all of the heavy-traffic explosion but none of the light-traffic explosion.

Abate and Whitt (1987) developed approximations for the bias of the sample mean starting at zero for any t and determined the asymptotic bias exactly, which is $\bar\beta = \rho/(1-\rho)^3$. (Note that the asymptotic bias is just the mean $\rho/(1-\rho)$ times the mean of the first-moment *cdf*; see Section 3 of Abate and Whitt.) Hence, for large t, the relative bias $\bar\beta/\bar f t$ is approximately $1/(1-\rho)^2 t$, which is asymptotically negligible compared to $1/\sqrt{t}$ and the relative width of a confidence interval.

3.2 Time Scaling

Suppose that the process of interest $X(t)$ is equal to $yZ(zt)$ for positive constants y, z and some other process $Z(t)$ with asymptotic parameters $\bar f_Z$, $\sigma_{\infty\,Z}^2$ and $\bar\sigma_Z^2$. It is easy to see that the corresponding asymptotic parameters for $X(t)$ are $\bar f = y\bar f_Z$, $\sigma_\infty^2 = y^2\sigma_{\infty\,Z}^2$ and $\bar\sigma^2 = y^2\bar\sigma_Z^2/z$, see Section 4.2 of Whitt (1989a), so that

$$\bar\sigma^2/\bar f^2 = (\bar\sigma_Z^2/\bar f_Z)/z\ , \tag{7}$$

which shows the fundamental role played by the time scaling z.

3.3 RBM Approximations

Let $R(t)$ be RBM with drift coefficient -1 and diffusion coefficient 1, which has an exponential stationary distribution with mean $1/2$. Then $\bar f = 1/2$, $\sigma_\infty^2 = 1/4$, $\sigma_\infty^2/\bar f^2 = 1$, $\bar\sigma^2 = 1/2$ and $\bar\sigma^2/\bar f^2 = \bar\sigma^2/\sigma_\infty^2 = 2$; see Section 4.1 of Whitt (1989a). Here we see that the correlation factor and the steady-state variability factor are of approximately the same order. The interesting phenomena occur in the scaling associated with an approximation. The standard heavy-traffic approximation for a queueing process $X(t)$ associated with an m-server queue with arrival rate and traffic intensity ρ is

$$((1-\rho)/a)X(at/(1-\rho)^2) \approx R(t)$$

or, equivalently,

$$X(t) \approx aR((1-\rho)^2 t/a)/(1-\rho)\ , \tag{8}$$

where a is a constant depending on the variability of the arrival and service processes. Hence, $\bar f \approx a/2(1-\rho)$, $\sigma_\infty^2 \approx \bar f^2$, $\bar\sigma^2 \approx a^3/2(1-\rho)^4$ and the run-length ratio is $(\bar\sigma^2/\bar f^2) \approx 2a/(1-\rho)^2$. We see that the run-length ratio is directly proportional to a and inversely proportional to $(1-\rho)^2$. If $X(t)$ is the queue length process in the standard GI/G/m model with m servers, then $a = c_A^2 + c_S^2$ where c_A^2

and c_S^2 are the squared coefficients of variation (variance divided by the square of the mean) of an interarrival times and a service time, respectively; see Section 5.1 of Whitt (1989a).

3.4 A Packet Queue Model

To illustrate the intended applications, we now present an example from Fendick, Saksena and Whitt (1989) and Sections 5.3 and 6.2 of Whitt (1989a). The model is for a packet switch with k classes of traffic and variable packet lengths. We assume that the switch acts as a first-come first-served single-server queue with unlimited waiting space. The service times are proportional to packet length. For class i, the service times are assumed to be i.i.d. with a general distribution having mean τ_i and squared coefficient of variation c_{si}^2. For each class, traffic consists of messages divided into packets. For the class i packet stream, we assume that packets arrive in batches (the messages), with successive batch sizes being i.i.d. with a general distribution having mean m_i and squared coefficient of variation c_{bi}^2. For each class there are spaces between the arrival epochs of packets within the same batch. For class i, the spaces are i.i.d. with a general distribution having mean ξ_i and squared coefficient of variation c_{xi}^2. Following the arrival of all packets in a batch there is an idle period before the arrival of the first packet of the next batch, with a general distribution having mean ω_i and squared coefficient of variation c_{Ii}^2. We assume that all the service times, batch sizes, spacings and idle periods are mutually independent.

Let λp_i be the arrival rate of messages for class i, where $p_1 + \ldots + p_k = 1$. From above $\lambda p_i = 1/(m_i \xi_i + \omega_i)$. The associated packet arrival rate for class i is $\bar{\lambda} q_i = \lambda p_i m_i = m_i/(m_i \xi_i + \omega_i)$, where $q_i = p_i m_i / \sum_{j=1}^{k} p_j m_j$ is the proportion of all arrivals that are of class i and $\bar{\lambda}$ is the total packet arrival rate. Let $r_i = \tau_i/\tau$ where τ is the average service time for all classes, i.e., $\tau = \sum_{i=1}^{k} q_i \tau_i / \sum_{i=1}^{k} q_i$. Let β_i be the proportion of busy time in each busy cycle for class i, defined by $\beta_i = m_i \xi_i/(m_i \xi_i + \omega_i)$. For convenience, choose measuring units so that $\tau = 1$ and the traffic intensity is $\rho = \bar{\lambda}$.

It should be apparent that this model is difficult to analyze exactly, but we have proved a heavy-traffic limit theorem establishing a diffusion approximation for the workload or virtual waiting time process, so that (8) is valid with $a = c_A^2 + c_S^2 - 2 c_{AS}^2$, where

$$c_A^2 = \sum_{i=1}^{k} q_i c_{Ai}^2 ,$$

$$c_S^2 = \sum_{i=1}^{k} q_i [r_i^2 c_{si}^2 + (r_i - 1)^2 c_{Ai}^2] ,$$

$$c_{AS}^2 = \sum_{i=1}^{k} q_i (1 - r_i) c_{Ai}^2$$

$$c_{Ai}^2 = m_i (1 - \beta_i)^2 (c_{bi}^2 + c_{Ii}^2) + \beta_i^2 c_{xi}^2 . \qquad (9)$$

As indicated in Fendick et al., quite large values of the parameters c_A^2, c_S^2 and c_{AS}^2 can occur; typical values are $c_A^2 = 20$, $c_S^2 = 35$ and $c_{AS}^2 = -7$ yielding $b = 60$.

Simulation experience indicates that the RBM approximation is effective for simulation run length planning, even though the approximation errors in light traffic from (8) and (9) are enormous; see Table 4 of Whitt (1989a).

4. THE ASYMPTOTIC VARIANCE OF THE SAMPLE MEAN

From Section 2 it is clear that simulation run length planning can easily be performed if we can approximately determine the asymptotic variance and the asymptotic bias of the sample mean ($\bar{\sigma}^2$ and $\bar{\beta}$) as well as the steady-state mean ($\bar{f}$) itself. In this section we present formulas for these quantities for diffusion processes and other basic Markov processes, drawing on Whitt (1989b). Of course, there is a large body of related work; e.g., see Hordijk, Iglehart and Schassberger (1976), Hazen and Pritsker (1980), Glynn (1984, 1989), Grassman (1987) and references cited there.

4.1 Stationary Processes and the Spectral Density

A starting point for calculating $\bar{\sigma}^2$ is the formula

$$\bar{\sigma}^2 = \lim_{t \to \infty} \int_{-t}^{t} \left[1 - \frac{|s|}{t} \right] R(s)\, ds = 2 \int_{0}^{\infty} R(t)\, dt \ ,(10)$$

where $R(t) = E\, Y^*(0)\, Y^*(t) - (E\, Y^*(0))^2$ is the (auto) *covariance function* of the stationary version $Y^*(t)$ of the stochastic process $Y(t)$, which we assume is well defined; i.e., we assume that $E[Y^*(t)^2] < \infty$ so that $R(t)$ is finite for each t and we assume that $\int_0^\infty |R(t)| dt < \infty$; see Chapter 5 of Fishman (1978). However, (10) is not too useful by itself because typically it is relatively difficult to calculate $R(t)$, except for certain special cases, e.g., for reversible Markov chains $R(t)$ can be calculated relatively easily via the spectral representation because the eigenvalues are all real; see Chapter 7 of Keilson (1979). However, we have in mind much easier calculations, as in Section 3.

One way to actually calculate $\bar{\sigma}^2$ starting from (5) is to calculate the *spectral density* (the Fourier transform of $R(t)$) at 0, paralleling the estimation procedure in Heidelberger and Welch (1981) and references cited there, because $\bar{\sigma}^2 = 2\pi\, s(0)$, where $s(\omega)$ is the spectral density, defined by

$$s(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega t}\, R(t)\, dt\,. \qquad (11)$$

4.2 Regenerative Processes

Another starting point for calculating $\bar{\sigma}^2$ is the regenerative process formula

$$\bar{\sigma}^2 = \frac{\mathrm{Var}\left[\int_{T_0}^{T_1} (Y(t) - \bar{f}\,)\, dt\right]}{E\,(T_1 - T_0)}\;, \qquad (12)$$

assuming that $\{T_n : \; n \geq 0\}$ is a sequence of regeneration times for $X(t)$ with $E(T_1 - T_0)^2 < \infty$ and $\left[E \int_{T_0}^{T_1} |Y(t)|\, dt\right]^2 < \infty$; e.g., see Glynn and Whitt (1987). (The argument is essentially the same as in I.14-16 of Chung (1967).) Then we can also calculate $\bar{f}$ by

$$\bar{f} = \frac{E\int_{T_0}^{T_1} Y(t)\, dt}{E\,(T_1 - T_0)}\;. \qquad (13)$$

Obviously, neither (12) nor (13) is a very tractable expression by itself, but in Markov chains (12) and (13) can be the basis for calculation, as shown by Hordijk, Iglehart and Schassberger (1976). Then we can calculate both $\bar{f}$ and $\bar{\sigma}$ by performing successive approximations of the form $x_{k+1} = x_0 + Px_k$, where P is the transition matrix of a transient Markov chain. Ways to improve and bound the rate of convergence of such successive approximations are described by van der Wal and Schweitzer (1987).

4.3 Continuous-Time Markov Chains

We obtain more tractable expressions when we impose more structure. Now suppose that $X(t)$ is an irreducible continuous-time Markov chain (CTMC) with state space $\{0, 1, \ldots, n\}$ and generator matrix Q. (Similar results hold for infinite state spaces under regularity conditions, e.g., assuming (12) and (13) are valid. Similar results also hold for discrete-time Markov chains.) Let Π be the matrix with all rows equal to the stationary probability vector π. It is well known that π is the unique solution to $xQ = \theta$ where $\theta = (0, \ldots, 0)$ and $\sum_{i=0}^{n} x_i = 1$. From Hazen and Pritsker (1980), Glynn (1984), Grassman (1987), Whitt (1989b) and others (the basic theory was established much earlier),

$$\bar{\sigma}^2 = 2 \sum_{i=0}^{n} \sum_{j=0}^{n} f_i\, \pi_i\, Z_{ij}\, f_j \qquad (14)$$

where $f_i = f\,(i)$ and Z is the *fundamental matrix*, which can be defined by

$$Z = (\Pi - Q)^{-1} - \Pi\,. \qquad (15)$$

Alternatively, $\bar{\sigma}^2 = 2 \sum_{j=0}^{n} x_j\, f_j$ where x is the unique solution to $xQ = -y$ with $y_i = (\, f_i - \bar{f}\,)\, \pi_i$ and $\sum_{i=0}^{n} x_i = 0$. Hence, if the CTMC is a birth and death process or, more generally, is *skip free* ($Q_{ij} = 0$ for all $j \geq i+2$ or for all $j \leq i-2$), then we can calculate x and thus $\bar{\sigma}^2$ *recursively* instead of inverting the matrix $(\Pi - Q)$ in (15). (We initially let $x_0 = 1$ and normalize afterwards so that $\Sigma x_i = 1$.) For birth and death processes, this observation was made by Grassman (1987). If the CTMC is not skip free, then we can calculate iteratively; see Section 5 of Whitt (1989b).

Paralleling (14), the asymptotic bias starting with initial probability vector α, say $\bar{\beta}_\alpha$, is

$$\bar{\beta}_\alpha = \sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_i\, Z_{ij}\, f_j \qquad (16)$$

for Z in (15). We can also calculate $\bar{\beta}_\alpha$ by $\bar{\beta}_\alpha = \sum_{j=0}^{n} x_j\, f_j$ where x is the unique solution to $xQ = -\alpha + \pi$ with $\sum_{i=0}^{n} x_i = 0$.

4.4 Birth and Death Processes

We now continue towards more concrete formulas by imposing even more structure. Let $X(t)$ be a positive recurrent birth and death process on the set $\{0, 1, \ldots, n\}$ with birth rates λ_i, death rates μ_i and stationary probabilities $\pi_i = \pi_0\, (\lambda_0\, \lambda_1 \ldots \lambda_{i-1}) / (\mu_1 \ldots \mu_i)$. (The process is reflecting at 0 and n; i.e., $\lambda_n = \mu_0 = 0$.) Then

$$\bar{\sigma}^2 = 2 \sum_{j=0}^{n-1} \frac{1}{\lambda_j \pi_j} \left[\sum_{i=0}^{j} (f_i - \bar{f}) \pi_i \right]^2 \qquad (17)$$

and

$$\bar{\beta}_\alpha = \sum_{j=0}^{n-1} \frac{1}{\lambda_j \pi_j} \sum_{i=0}^{j} (f_i - \bar{f}) \pi_i \sum_{i=0}^{j} (\alpha_i - \pi_i) \text{ .(18)}$$

((17) was communicated by Burman (1980).) For the M/M/1 queue in Section 3 we easily obtain (6) from (17). For computation in the general case, it is often convenient to move π_j in (17) and (18) into the inner sum; see (17) of Whitt (1989b).

4.5 Diffusion Processes

Finally, let $X(t)$ be a positive recurrent diffusion process on the interval $[a, b]$ with drift coefficient $\mu(x)$, diffusion coefficient $\sigma^2(x)$ and reflecting boundaries at a and b. Since a diffusion process is the continuous-state analog of a birth and death process, we obtain continuous-state analogs of (17) and (18). First, the *stationary density* is $\pi(y) = m(y)/M(b)$ where $m(y) = 2/\sigma^2(y)\, s(y)$ is the *speed density*, $M(y) = \int_a^y m(x)\,dx$ and $s(y) = \exp\left[-\int_a^y [2\mu(x)/\sigma^2(x)]\,dx \right]$ is the *scale density*, assuming all integrals are finite. Then

$$\bar{\sigma}^2 = 2 \int_a^b \frac{2}{\sigma^2(y)\, \pi(y)} \left[\int_a^y (f(x) - \bar{f})\, \pi(x)\, dx \right]^2 dy \qquad (19)$$

and

$$\bar{\beta}_\alpha = a \int_a^b \frac{2}{\sigma^2(y)\, \pi(y)} \times$$

$$\left[\int_a^y (f(x) - \bar{f})\, \pi(x)\, dx \int_a^y (\alpha(x) - \pi(x))\, dx \right] dy. \qquad (20)$$

(Formula (19) appears on p. 94 of Mandl (1968).) For RBM in Section 3 we easily obtain $\bar{\sigma}^2 = 1/2$ from (19).

ACKNOWLEDGMENT

I thank Philip Heidelberger for encouraging me to continue beyond Whitt (1989a).

REFERENCES

Abate, J. and Whitt, W. (1987). Transient behavior of the M/M/1 queue: starting at the origin. *Queueing Systems* 2, 41-65.

Albin, S. L. (1984). Simulation to develop and test queue approximations: a case study. *Simulation* 43, 279-285.

Asmussen, S. (1989). Validating the heavy traffic performance of regenerative simulation. *Stochastic Models*, to appear.

Asmussen, S. (1990). Regenerative simulation in heavy traffic. *Mathematics of Operations Research*, to appear.

Blomqvist, N. (1969). Estimation of waiting-time parameters in the GI/G/1 queueing system, part II: heavy-traffic approximations. *Skandinavisk Aktuarietidskrift*, Uppsala, Sweden, 125-136.

Burman, D. Y. (1980). A functional central limit theorem for birth and death processes. Talk at the ORSA/TIMS Conference in Washington, D.C. and personal communication.

Chung, K. L. (1967). *Markov Chains*, Second ed., Springer-Verlag, NY.

Fendick, K. W., Saksena, V. R. and Whitt, W. (1989). Dependence in packet queues. *IEEE Transactions on Communications*, to appear.

Fishman, G. S. (1978). *Principles of Discrete Event Simulation*, Wiley, New York.

Glynn, P. W. (1984). Some asymptotic formulas for Markov chains with applications to simulation. *Journal of Statistical Computation and Simulation* 19, 97-112.

Glynn, P. W. (1989). Poisson's equation for Markov chains. In preparation.

Glynn, P. W. and Whitt, W. (1987). Sufficient conditions for functional-limit-theorem versions of $L = \lambda W$. *Queueing Systems* 1, 279-287.

Grassman, W. K. (1987). The asymptotic variance of a time average in a birth-death process. *Annals of Operations Research* 8, 165-175.

Hazen, G. B. and Pritsker, A. A. B. (1980). Formulas for the variance of the sample mean in finite state Markov processes. *Journal of Statistical Computation and Simulation* 12, 25-40.

Heidelberger, P. and Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM* 24, 233-245.

Heidelberger, P. and Whitt, W. (1989). Simulation efficiency in some queueing models. In preparation.

Hordijk, A., Iglehart, D. L. and Schassberger, R. (1976). Discrete time methods for simulating continuous time Markov chains. *Advances in Applied Probability* 8, 772-788.

Keilson, J. (1979). *Markov Chain Models — Rarity and Exponentiality*, Springer-Verlag, New York.

Mandl, P. (1968). *Analytical Treatment of One-Dimensional Markov Processes*, Springer-Verlag, NY.

Moeller, T. and Kobayashi, H. (1974). Use of diffusion approximation to estimate run length in simulation experiments. COMPSTAT 1974, *Proceedings of Computational Statistics*, (G. Bruckmann, F. Fershel and L. Schmetterer, eds.) Physica-Verlag, Vienna, 363-372.

van der Wal, J. and Schweitzer, P. J. (1987). Iterative bounds on the equilibrium distribution of a finite Markov chain. *Probability in the Engineering and Information Sciences* 1, 117-131.

Whitt, W. (1989a). Planning queueing simulations. *Management Science* 35, to appear.

Whitt, W. (1989b). Asymptotic formulas for Markov chains with application to simulation. Submitted for publication.

Woodside, C. M., Pagurek, B. and Newell, G. F. (1980). A diffusion approximation for correlation in queues. *Journal of Applied Probability* 17, 1033-1047.

AUTHOR'S BIOGRAPHY

WARD WHITT is a member of technical staff in the Mathematical Sciences Research Center of AT&T Bell Laboratories. He received an A.B. in mathematics from Dartmouth College in 1964, and a Ph.D. in operations research from Cornell University in 1969. Before joining AT&T Bell Laboratories in 1977, he was on the faculty of the Department of Operations Research at Stanford University (1968-1969) and the Department of Administrative Sciences at Yale University (1969-1977). His research interests are primarily in probability theory and its applications, especially to queues. He has been involved with several simulation studies to develop and evaluate approximations for queueing models. He is a member of ORSA and IMS.