On Approximate Computer System Models

EROL GELENBE

Institut de Recherche d'Informatique et d'Automatique, Rocquencourt, France

ABSTRACT A new treatment of the boundary conditions of diffusion approximations for interconnected queueing systems is presented The results have applications to the study of the performance of multiple-resource computer systems In this approximation method additional equations to represent the behavior of the queues when they are empty are introduced. This reduces the dependence of the model on heavy traffic assumptions and yields certain results which would be expected from queueing or renewal theory The accuracy of the approach is evaluated by comparison with certain known exact or numerical results.

KEY WORDS AND PHRASES. computer system performance, queueing theory, operating systems, multiprogramming, approximate models of performance

CR CATEGORIES 4.32, 4 35, 4.6, 5 5, 6.29, 8.1

1. *Introduction*

Recently, considerable interest has been shown for obtaining approximate solutions to queueing models using diffusion processes [5, 12]. The interest of these approximations lies essentially in the fact that explicit though approximate results are obtainable for relatively complex situations where the only possible alternative lies in numerical methods or simulation experiments. It is not surprising that this approach has been applied to the mathematical modeling of multiple-resource computer systems [7, 8, 11], for which analytical results are otherwise difficult to obtain.

The basic argument used in these approximations (see, for instance, [7, p. 570]) is that the number of users $N(t)$ in queue at time t will tend to become normally distributed for large t, with mean $b(t).t$ and variance $\alpha(t).t$ if "boundary conditions" (such as the case where the queue is empty) are neglected, thus one argues that the process $N(t)$ can be approximated by a diffusion process. Reflecting boundaries (for instance, at the origin for a single-server queue) are imposed in order to keep the process in the desired region (for instance, on the positive real line) but also to insure that no probability mass collects at the boundaries. These models are then only useful in heavy traffic conditions (for a single-server queue when the traffic intensity is close to or greater than one).

The interest for such models as vehicles for representing the behavior of multiple interconnected resource computer systems stems from the need to predict performance with a reasonable degree of accuracy for design and evaluation. Explicit results yielding performance measures from mathematical models are sought since they reduce the need for costly simulation experiments and yield a clearer insight into the basic assumptions and into the interaction of model parameters.

The advantage of diffusion models over other techniques is that they can yield explicit solutions with a high degree of accuracy for nonexponential service time assumptions

Author's present address· Chaire d'Informatique de l'Université de Liège, Av des Tilleuls 59, Liège, Belgium.

which often can only be treated numerically (if at all) by conventional queueing theory techniques or at high cost and lower precision using simulation runs.

In this paper we introduce a new diffusion approximation for queueing systems and we apply it to a multiprogramming system consisting of a central processing unit (CPU) and a secondary memory (SM) or input-output device, assuming that a fixed number of programs circulate in the system in alternate cycles of computation and input-output.

The novelty of our approximation technique with respect to approaches derived from the work of Newell [12] or Gaver [5] is in the treatment of boundary conditions for the model equations. By assuming that the case when a queue is emptied is represented by letting the diffusion process approach a reflecting boundary (as is done in [5, 12]), the actual behavior of the queueing process, in which the number in queue remains zero for a nonzero amount of time, is not adequately represented. Known queueing theory results (such as the stationary probability of an empty queue) have then to be introduced as initial conditions, and fairly arbitrary modifications [11] are made to the stationary solutions in order to conform to reality. In our approach, however, equations are introduced describing the dynamics of the system when a queue is empty. The fact that the number in queue remains zero for a random but nonzero period of time, and that after an arrival occurs it increases instantaneously to a new level, is properly taken into account. This leads to a self-contained model yielding some known queueing theory results exactly, and more accurate approximations for light traffic conditions when the queues are more frequently empty.

In Section 2 we present the equations of the approximate model that we propose without particular reference to a queueing system. Our objective is to interpret these equations in terms of the flow of probability masses. In Section 3 the modeling method is applied to the M/G/1 queue in order to illustrate its use. Section 4 is devoted to an approximate model of a multiprogramming system; the stationary queue length distributions and the CPU utilization are computed. Numerical examples are then presented in order to evaluate the accuracy of our approach.

2. *The Instantaneous Return Process*

Let us briefly and informally present the basic equations for the model, which we call the instantaneous return process, and for which an informal presentation can be found in [1] and a rigorous treatment in [2, 3].

Consider the stochastic process $\{X(t), t \geq 0\}$ on the closed interval $[0, M]$ of the positive real line. On the open interval $]0,M[$ the process behaves as a diffusion (Wiener) process. However, when the process reaches one of the boundaries (0 or M) it remains there for an exponentially distributed time, after which it jumps instantaneously back into $]0,M[$, distributing itself with some probability density function over the open interval, reinitializing the diffusion process. Since the holding times on the boundaries are exponentially distributed, the instantaneous return process $\{X(t), t \geq 0\}$ retains the Markov property. Though it is easy to imagine a semi-Markov version of this process with arbitrarily distributed holding times at the boundaries, we do not know of mathematical results covering this case.

Let $m_1(t)$, $m_2(t)$ be the probability masses concentrated at the lower and upper boundary and $1/\lambda_1$, $1/\lambda_2$ the expected holding times at 0 and M, respectively.

The probability density function of the point from which the diffusion process starts once again immediately after a jump is $f_1(x)$ if the jump originated at 0 and $f_2(x)$ if it originated at M. Denote also by $A_{x,t}$ the forward operator,

$$A_{x,t}f = -(\partial/\partial t)f(x, t) - (\partial/\partial t)[b(x, t)f(x, t)] + \tfrac{1}{2}(\partial^2/\partial x^2)[\alpha(x, t)f(x, t)] \qquad (1)$$

and by $B_{x,t}$ the operator

$$B_{x,t}f = -b(x, t)f(x, t) + \tfrac{1}{2}(\partial/\partial_x)[\alpha(x, t)f(x, t)], \qquad (2)$$

where $f = f(x, t)$ is the probability density function of the process at time t, and

$$b(x, t) = \lim_{\Delta t \to 0} \frac{E\{X(t + \Delta t) - X(t) \,|\, X(t) = x\}}{\Delta t} \tag{3}$$

$$\alpha(x, t) = \lim_{\Delta t \to 0} \left[\frac{E\{[X(t + \Delta t) - X(t)]^2 \,|\, X(t) = x\}}{\Delta t} - \frac{[E\{X(t + \Delta t) - X(t) \,|\, X(t) = x\}]^2}{\Delta t} \right] \tag{4}$$

are the instantaneous rate of change of the mean and variance of $X(t)$ given that $X(t) = x$.

The equations for the instantaneous return process are [1–3]:

$$A_{x,t}f + \lambda_1 m_1(t) f_1(x) + \lambda_2 m_2(t) f_2(x) = 0, \tag{5}$$

$$(d/dt) m_1(t) = B_{0,t}f - \lambda_1 m_1(t), \tag{6}$$

$$(d/dt) m_2(t) = -B_{M,t}f - \lambda_2 m_2(t). \tag{7}$$

These equations have a simple interpretation. The term

$$B_{0,t}f = \lim_{x \to 0} [-b(x, t) f(x, t) + \tfrac{1}{2}(\partial/\partial x)(\alpha(x, t) f(x, t))]$$

represents the rate of flow of the probability mass from the region]0, M[to the lower boundary, while $-B_{M,t}f$ is the flow to the upper boundary. $\lambda_1 m_1(t)$ is the rate of flow of mass from the lower boundary into]0, M[and $\lambda_2 m_2(t)$ is the corresponding quantity from the upper boundary. The conservation of mass should be satisfied; that is, we must have

$$(\partial/\partial t)[\textstyle\int_0^M f(x, t)dx + m_1(t) + m_2(t)] = 0. \tag{8}$$

From (5) and (1) and since $\int_0^M f_1(x)dx = 1$, $\int_0^M f_2(x)dx = 1$,

$$\begin{aligned}(\partial/\partial t) \textstyle\int_0^M f(x, t)dx &= [-b(x, t)f(x, t) + \tfrac{1}{2}(\partial/\partial x)[\alpha(x, t)f(x, t)]]_0^M \\ &\quad + \lambda_1 m_1(t) + \lambda_2 m_2(t) \\ &= B_{M,t}f - B_{0,t}f + \lambda_1 m_1(t) + \lambda_2 m_2(t). \end{aligned} \tag{9}$$

Using (6), (7), and (9) we have

$$\begin{aligned}(\partial/\partial t) \textstyle\int_0^M f(x, t)dx + (dm_1(t)/dt) + (dm_2(t)/dt) &= B_{M,t}f - B_{0,t}f + \lambda_1 m_1(t) \\ &\quad + \lambda_2 m_2(t) + B_{0,t}f - \lambda_1 m_1(t) \\ &\quad - B_{M,t}f - \lambda_2 m_2(t) \\ &= 0, \end{aligned}$$

which verifies (8). The sum of the probability mass is one:

$$\textstyle\int_0^M f(x, t)dx + m_1(t) + m_2(t) = 1. \tag{10}$$

Equation (5) also can be interpreted in terms of the flow of probability masses. Let Ω be a subinterval of]0, M[. Then we write

$$\textstyle\int_\Omega A_{x,t}f dx + \lambda_1 m_1(t) \int_\Omega f_1(x)dx + \lambda_2 m_2(t) \int_\Omega f_2(x)dx = 0$$

or from (1),

$$\begin{aligned}(\partial/\partial t) \textstyle\int_\Omega f(x, t)dx = &\textstyle\int_\Omega [-(\partial/\partial x)(b(x, t)f(x, t)) + \tfrac{1}{2}(\partial^2/\partial x^2)(\alpha(x, t)f(x, t))]dx \\ &+ \lambda_1 m_1(t) \textstyle\int_\Omega f_1(x)dx + \lambda_2 m_2(t) \int_\Omega f_2(t)dx, \end{aligned} \tag{11}$$

which states that the rate of change of the probability mass in Ω is equal to the flow of mass from the boundaries 0 and M (the last two terms on the right-hand side of (11)) plus the flow out of the region Ω, yielding (5).

Although a more general case may be considered, we shall assume that the boundaries act as absorbing boundaries for the diffusion process *until* the jump occurs which corresponds to the queueing problems of interest. Thus we take for all $t \geq 0$,

$$\lim_{x \to 0} f(x, t) = \lim_{x \to M} f(x, t) = 0 \tag{12}$$

as with absorbing boundaries. Of course, initial conditions for $m_1(t)$, $m_2(t)$, and $f(x, t)$ will also have to be given.

Let us note that λ_1, λ_2 may be functions of time, and that $f_1(x), f_2(x)$ could be taken to be functions of the instant of time at which the jump occurs.

3. *Approximation to the M/G/1 Queue*

We shall apply the instantaneous return model to the M/G/1 queueing system. The process $\{X(t), t \geq 0\}$, taken on the interval $[0, \infty]$ (the nonnegative real line), approximates the number of requests in queue including the one in service.

In the M/G/1 queue the interarrival and service times are independent of each other and of the number in queue. The interarrival process is Poisson of parameter λ, and the service times are independent identically distributed random variables with common distribution function of mean $1/\mu$ and variance V_s. The variance of the time between successive arrivals is of course $V_a = (1/\mu)^2$.

Via an extension of arguments based on the central limit theorem [4], it can be shown that the parameters b and α defined in (3) and (4) are given asymptotically (as $t \to \infty$) by (see also [7, 12]), $b = \lambda - \mu$, $\alpha = \lambda^3 V_a + \mu^3 V_s$, provided the queue never becomes empty. The diffusion approximation is based on assuming that this asymptotic behavior describes the number in queue as soon as a busy period begins, that is, as soon as the number in queue moves from zero to one. Let $f(x, t)dx$ denote the probability that $X(t)$, the approximate representation of the number of customers in queue, lies in $[x, x + dx]$; also let $m(t)$ be the approximated probability that the queue is empty at time t. Since $X(t)$ takes values in the nonnegative real line, there is no upper boundary to be considered.

After an arrival to the empty queue, the number in queue jumps instantaneously to $+1$; therefore we must let $f_1(x)$ be the Dirac delta function. We shall only consider the stationary state defined by $((\partial f/\partial t) = 0, (dm/dt) = 0)$ so that (5) and (6) become

$$-b(\partial f/\partial x) + \tfrac{1}{2}\alpha(\partial^2 f/\partial x^2) = -\lambda m \delta(x - 1), \tag{13}$$

$$\lim_{x \to 0} [-bf + \tfrac{1}{2}\alpha(\partial f/\partial x)] = \lambda m, \tag{14}$$

where $\delta(x - 1)$ is a Dirac density function concentrated at $x = 1$, m is the stationary probability mass at $x = 0$, and α, b are independent of x, t. Since the arrival process is Poisson, we have $\alpha = \lambda + \mu^3 V_s$.

Let us note

$$\gamma = 2b/\alpha = -2(1 - \rho)(\rho + K_s)^{-1}, \tag{15}$$

where $\rho = \lambda/\mu$ and $K_s = \mu^2 V_s$. The solution to (13) with boundary conditions (14) and $\lim_{x \to 0} f(x) = 0$ from (12) is

$$f = \begin{cases} (m\lambda/b)[e^{\gamma x} - 1], & 0 \leq x \leq 1, \\ (m\lambda/b)[1 - e^{-\gamma}]e^{\gamma x}, & x \geq 1. \end{cases} \tag{16}$$

To compute m we use

$$m + \textstyle\int_0^\infty f dx = 1, \tag{17}$$

which yields after some computation, and on condition that $\gamma < 0$ (i.e. $\rho < 1$),

$$m = 1 - \rho \tag{18}$$

for the probability of an empty queue, which one would expect to obtain. Finally let us compute n, the average number of customers in the system, at steady state. Let us discretize the density by setting

$$\pi_0 = m; \qquad \pi_i = \int_{i-1}^{i} f dx, \quad i > 0,$$

where the π_i, $i \geq 0$, denote the stationary probabilities for the number of customers in queue. We can then compute

$$n = \sum_{i=1}^{\infty} i\pi_i = \rho[1 - \gamma^{-1}] = \rho\left[1 + \frac{(\rho + K_s)}{2(1 - \rho)}\right]. \tag{19}$$

This differs from the Pollaczek-Khintchine formula which yields

$$\hat{n} = \rho\left[1 + \frac{\rho(1 + K_s)}{2(1 - \rho)}\right]. \tag{20}$$

Other approaches to diffusion approximations [5, 12] do not yield (18) or (20). Though we are able to obtain the probability of an empty queue correctly with our approach, we have been unable to obtain (20) exactly. Notice that the exact result would have been obtained if γ were

$$\hat{\gamma} = -(2/\rho)(1 - \rho)(K_a + K_s)^{-1}$$

instead of (15). The error in the average queue length is given by $n - \hat{n} = \rho K_s/2$.

4. *Approximation of a Closed Two-Server System*

Consider the closed two-server system shown in Figure 1. The system contains a fixed number M of customers, which we shall call programs, and we shall call the two servers the central processing unit (CPU) and the secondary memory (SM), respectively. This model has been analyzed by Gaver and Shedler [7] using a diffusion approximation with reflecting boundaries. In order to choose appropriate boundary conditions to the diffusion equation, they suggest one of three approaches: the use of renewal theory to obtain an exact fit for (1) the case $M = 1$, or (2) the case $M = 2$, or (3) the use of the basic queueing theory result $\pi_0 = 1 - \rho$ to obtain an exact fit for $M = \infty$. The results obtained in [7] yield very good fits to an exact solution; conceptually, however, one is not fully satisfied since the boundary conditions have to be introduced from queueing or renewal theory.

In this section we analyze the same model as in [7] (shown in Figure 1), making use of the instantaneous return process. Except for the specification of the constants α, b of the diffusion equation, no results from queueing or renewal theory need be used to solve for the probability distribution function $F(x)$ of the number of programs in the CPU queue at steady state. We shall also obtain the probability of an empty CPU queue which will satisfy the queueing theory result for $M = \infty$ mentioned above. It will also be shown that our result satisfies another familiar queueing theory result: for each server the stationary arrival rate is equal to the stationary departure rate for any value of M.

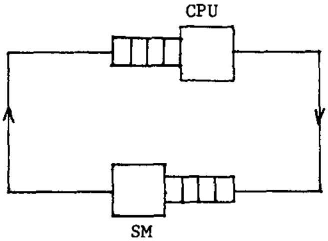

FIG. 1

As previously, let $f(x)$ be the stationary probability density function for the diffusion process in $]0, M[$, where x is the value approximating the number of programs in the CPU queue; $M - x$ will correspond to the number in the SM queue. m_1, the stationary probability mass at the origin, represents the probability that the CPU queue is empty at steady state, and m_2 is the steady-state probability that it contains M programs. The instantaneous return model allows us to approximate the system under the following conditions.

(a) When the number of programs in the CPU queue is neither 0 nor M, the distribution of service time at both CPU and SM is arbitrary with mean $1/\mu$ and $1/\lambda$, respectively, and variance V_s and V_a, respectively.

(b) When the number of programs in the CPU queue is 0, the service time at the SM is exponentially distributed with parameter λ; if it is M, the CPU service time is exponentially distributed with parameter μ.

We assume that μ, λ, V_s, V_a are *not* functions of x. They may, however, be functions of M. The "jump" of the process from the lower boundary represents an arrival from the SM when the CPU queue is empty, the jump from the upper boundary is a departure from the CPU queue when all programs are there.

At steady state we set the following equations (from (5), (6), (7)) for the process:

$$\tfrac{1}{2}\alpha f'' - bf' = -\lambda m_1 \delta(x - 1) - \mu m_2 \delta(x - M + 1), \tag{21}$$

$$\lim_{x\to 0} \tfrac{1}{2}\alpha f' - bf = \lambda m_1, \tag{22}$$

$$\lim_{x\to M} \tfrac{1}{2}\alpha f' - bf = \mu m_2, \tag{23}$$

where $\delta(.)$ is the Dirac density function, and

$$b = \lambda - \mu, \qquad \alpha = \lambda^3 V_a + \mu^3 V_s. \tag{24}$$

The boundary conditions on f are $\lim_{x\to 0} f = \lim_{x\to M} f = 0$. We integrate (21) once to obtain

$$\tfrac{1}{2}\alpha f' - bf = -\lambda m_1 \phi(x - 1) - \mu m_2 \phi(x - M + 1) + c, \tag{25}$$

where c is a constant, and

$$\phi(x) = \begin{cases} 0, & \text{for } x < 0, \\ 1, & \text{for } x \geq 0. \end{cases}$$

Using (22) we see that

$$c = \lambda m_1 \tag{26}$$

and also that (26) satisfies (23). We then solve (25) in the three regions $0 \leq x \leq 1$, $1 \leq x \leq M - 1$, and $M - 1 \leq x \leq M$. It can be shown that the solution to (25) must be continuous so that we may use the continuity of f at $x = 1$ and $x = M - 1$. Solving for f and using (12) we obtain

$$f = \begin{cases} -(\lambda m_1/b)[1 - e^{\gamma x}], & 0 \leq x \leq 1, \\ -(\lambda m_1/b)[e^{-\gamma} - 1]e^{\gamma x}, & 1 \leq x \leq M - 1, \\ -(\mu m_2/b)[e^{\gamma(x-M)} - 1], & M - 1 \leq x \leq M, \end{cases} \tag{27}$$

$$m_2 = (\lambda m_1/\mu) e^{\gamma(M-1)}, \tag{28}$$

where, as previously, $\gamma = (2b/\alpha)$. Also, using

$$\int_0^M f dx + m_1 + m_2 = 1, \tag{29}$$

we obtain

$$\int_0^M f dx = -(\lambda m_1/b)[1 - e^{\gamma(M-1)}]$$

so that

$$m_1 + (\lambda m_1/\mu)e^{\gamma(M-1)} - (\lambda m_1/b)[1 - e^{\gamma(M-1)}] = 1.$$

Then, setting $\rho = \lambda/\mu$,

$$m_1 = (1 + \rho e^{\gamma(M-1)} + \rho(1 - \rho)^{-1}[1 - e^{\gamma(M-1)}])^{-1},$$

yielding the following proposition.

PROPOSITION 1. *The approximation by the instantaneous return process for the closed two-server system yields the stationary probability of an empty CPU queue as*

$$m_1 = (1 - \rho)(1 - \rho^2 e^{\gamma(M-1)})^{-1}. \tag{30}$$

If $\rho < 1$ and is independent of M, we have the familiar queueing theory result for $M \to \infty$: $m_1 = 1 - \rho$. Furthermore, when $\rho = 1$, by taking appropriate limits, we have

$$m_1 = (\tfrac{1}{2})(1 + (M - 1)(K_a + K_s)^{-1})^{-1}, \tag{31}$$

where (31) is valid if the squared coefficients of variation $K_a = \lambda^2 V_a^2$, $K_s = \mu^2 V_s^2$ do not vary with ρ. Note also that if $M = 1$, we obtain $m_1 = \mu(\lambda + \mu)^{-1} = (1/\lambda)(1/\lambda + 1/\mu)^{-1}$, which is the result we would expect from renewal or queueing theory.

Another result is that (28) and (30) imply $(1 - m_2)\lambda = (1 - m_1)\mu$, which states that the arrival rate to the CPU queue is equal to the departure rate, at steady state.

The results which are summarized in Proposition 1, and in particular (30), are useful in predicting the performance of the multiprogramming system model of Figure 1 with arbitrary distribution functions of processing time at the CPU and SM. Defining the *stationary CPU utilization* η as the stationary probability of having a nonempty CPU queue, we have $\eta = 1 - m_1$. This measure is of practical significance since it can be used as an indicator of the stationary system throughput (number of programs processed per unit time).

In order to evaluate the accuracy of our approximation method we have compared it numerically in Figure 2 with results obtained by the diffusion approximation of Gaver and Shedler [7] and with the exact semi-Markov analysis of Shedler [14]. The quantity being tabulated is the stationary CPU utilization, and for each case the SM service time is exponential of mean one ($\lambda = 1$). The CPU service time is Erlang (1, 2, 3, 4, 5, ∞), and $1/\mu$ is varied between 0.25 and 0.9. The number of programs sharing the multiprogramming system varies between one and ten. Columns marked S-M refer to results of exact semi-Markov analysis while Diff. refers to the approach in [14]; I-R refers to our approach. The relative error in η obtained from the instantaneous return approximation, defined as the absolute difference between that value of η and the exact value (from semi-Markov analysis) divided by the latter, is at most roughly 3 percent. Maximum error seems to be attained for $M = 2$ and a constant (Erlang-∞) service time at the CPU. This relative error is comparable to the best error margin one might be able to achieve via careful and lengthy simulation experiments or using measurements on a real system. Our results seem generally somewhat more accurate than those in [14], but it is difficult to make meaningful comparative comments when the level of error in both is so low. For $M = 1$ the factor η is obtained exactly by our analysis and the same is true for $M = \infty$. For larger values of M, say M above 6, the utilization factor tends rapidly towards the asymptotic value.

In [13] detailed comparisons are given of significant measures obtained by diffusion approximations with other modeling techniques, and in particular with simulation results, although unfortunately confidence intervals for the latter are not provided.

5. *Conclusions*

Diffusion approximations appear to be attractive means of approximating queueing networks as models of systems of interconnected resources since they allow a more detailed

M	E[S]	Erlang - 1 (exponential)			Erlang - 2			Erlang - 3			Erlang - 4			Erlang - 5			Erlang - ∞ (constant)		
		S-M	I-R	Diff	S-M	I-R	Diff.	S-M	I-R	Diff.	S-M	I-R	Diff	S-M	I-R	Diff.	S-M	I-R	Diff.
1	0.25		0.200			0.200			0.200			0.200			0.200			0.200	
2		0.236	0.236	0.233	0.243	0.244	0.237	0.245	0.246	0.238	0.246	0.248	0.239	0.247	0.248	0.239	0.249	0.250	0.241
3		0.247	0.246	0.245	0.249	0.249	0.247	0.250	0.250	0.247	0.250	0.250	0.247	0.250	0.250	0.247	0.250	0.250	0.248
4		0.249	0.249	0.249	0.250	0.250	0.249	0.250	0.250	0.249	0.250	0.250	0.249	0.250	0.250	0.249	0.250	0.250	0.250
5		0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250
6		0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250
7			0.250			0.250			0.250			0.250			0.250			0.250	
8			0.250			0.250			0.250			0.250			0.250			0.250	
9			0.250			0.250			0.250			0.250			0.250			0.250	
10			0.250			0.250			0.250			0.250			0.250			0.250	
1	0.50		0.333			0.333			0.333			0.333			0.333			0.333	
2		0.429	0.426	0.424	0.444	0.449	0.444	0.451	0.459	0.451	0.455	0.465	0.454	0.458	0.466	0.456	0.468	0.482	0.464
3		0.467	0.465	0.464	0.480	0.482	0.476	0.485	0.488	0.480	0.487	0.491	0.482	0.489	0.493	0.483	0.494	0.498	0.481
4		0.484	0.482	0.482	0.493	0.494	0.490	0.495	0.497	0.492	0.496	0.498	0.493	0.497	0.498	0.493	0.499	0.500	0.495
5		0.492	0.491	0.491	0.497	0.498	0.495	0.498	0.499	0.497	0.499	0.499	0.497	0.499	0.500	0.497	0.500	0.500	0.498
6		0.496	0.495	0.495	0.499	0.499	0.498	0.500	0.500	0.499	0.500	0.500	0.499	0.500	0.500	0.499	0.500	0.500	0.499
7			0.498			0.500			0.500			0.500			0.500			0.500	
8			0.499			0.500			0.500			0.500			0.500			0.500	
9			0.499			0.500			0.500			0.500			0.500			0.500	
10			0.500			0.500			0.500			0.500			0.500			0.500	
1	0.75																		
2		0.568	0.567	0.566	0.591	0.599	0.608	0.601	0.613	0.623	0.606	0.620	0.631	0.610	0.626	0.636	0.626	0.648	0.655
3		0.634	0.634	0.633	0.660	0.665	0.666	0.670	0.678	0.677	0.676	0.685	0.683	0.679	0.689	0.686	0.694	0.706	0.700
4		0.672	0.672	0.672	0.695	0.699	0.697	0.704	0.709	0.706	0.709	0.714	0.710	0.711	0.717	0.712	0.723	0.729	0.722
5		0.696	0.695	0.695	0.716	0.718	0.716	0.722	0.726	0.722	0.726	0.729	0.725	0.728	0.732	0.727	0.736	0.740	0.734
6		0.712	0.711	0.711	0.728	0.729	0.727	0.733	0.735	0.732	0.736	0.738	0.734	0.737	0.739	0.736	0.743	0.745	0.741
7			0.722			0.737			0.741			0.743			0.744			0.747	
8			0.729			0.741			0.744			0.746			0.746			0.749	
9			0.735			0.744			0.746			0.747			0.748			0.749	
10			0.739			0.746			0.748			0.748			0.749			0.750	
1	0.90																		
2		0.590	0.590	0.589	0.614	0.622	0.635	0.624	0.637	0.652	0.630	0.645	0.661	0.633	0.650	0.666	0.651	0.673	0.688
3		0.661	0.661	0.661	0.689	0.694	0.697	0.700	0.708	0.710	0.701	0.715	0.717	0.709	0.719	0.721	0.726	0.738	0.737
4		0.702	0.702	0.702	0.728	0.731	0.732	0.738	0.743	0.742	0.743	0.749	0.747	0.746	0.752	0.750	0.759	0.767	0.762
5		0.729	0.729	0.729	0.751	0.754	0.753	0.759	0.763	0.761	0.763	0.768	0.764	0.766	0.770	0.767	0.776	0.781	0.776
6		0.747	0.747	0.747	0.766	0.768	0.766	0.773	0.775	0.773	0.776	0.779	0.776	0.778	0.781	0.778	0.786	0.789	0.784
7			0.759			0.777			0.783			0.786			0.788			0.793	
8			0.769			0.784			0.789			0.791			0.792			0.796	
9			0.776			0.788			0.792			0.794			0.795			0.798	
10			0.781			0.792			0.794			0.796			0.796			0.799	
1	0.90																		
2		0.631	0.631	0.631	0.656	0.664	0.685	0.667	0.679	0.705	0.675	0.687	0.715	0.677	0.692	0.721	0.694	0.715	0.748
3		0.709	0.709	0.709	0.739	0.744	0.753	0.751	0.759	0.769	0.757	0.766	0.777	0.762	0.771	0.782	0.780	0.792	0.803
4		0.756	0.756	0.756	0.785	0.788	0.792	0.796	0.801	0.806	0.802	0.807	0.812	0.806	0.811	0.816	0.821	0.829	0.832
5		0.787	0.786	0.787	0.813	0.816	0.818	0.823	0.827	0.829	0.828	0.832	0.834	0.831	0.836	0.837	0.845	0.850	0.851
6		0.808	0.808	0.808	0.832	0.834	0.835	0.841	0.844	0.844	0.846	0.849	0.849	0.848	0.852	0.852	0.860	0.864	0.863
7			0.824			0.848			0.856			0.860			0.863			0.873	
8			0.837			0.857			0.865			0.868			0.871			0.879	
9			0.846			0.865			0.872			0.875			0.877			0.884	
10			0.854			0.871			0.877			0.880			0.881			0.888	

FIG. 2

characterization of service and interarrival time statistics with a greater economy of representation.

The basic problems raised by these approximations are (1) the choice of the diffusion parameters $\beta(x, t)$ and $\alpha(x, t)$, (2) the choice of the proper boundary conditions, and (3) the selection of the discretization of the probability density function $f(x, t)$ in the neighborhood of integer valued points $x = i$ in order to approximate the probability of finding i customers in queue at time t. In this paper we have suggested a solution to the issue raised in (2) using the instantaneous return process. This is distinct from previous approaches and seems to yield good approximations in the cases which have been discussed. Problem (1) can be treated either using asymptotic renewal theory [7, 12] or as in [6] using Wald's identity and the results of Haji and Newell [9]. Problem (3) does not seem to have been adequately treated as yet.

Our approach to (2) has the advantage that we do not have to call upon queueing theory results in order to obtain the integration constants of the diffusion equation; in fact some familiar queueing theory results are directly obtained from the instantaneous return model. Also our results are less dependent on heavy traffic assumptions. The approach, which appears to be a useful tool for the computation of performance measures for multiprogramming computer systems, is being extended to general queueing networks.

ACKNOWLEDGMENTS. The author wishes to thank Mr. Quadrat and Mr. Viot of IRIA-Laboria, as well as the anonymous referees, for their comments.

REFERENCES

(Note. Reference [10] is not cited in the text.)

1. Bharucha-Reid, A. T. *Elements of the Theory of Markov Processes and Their Applications.* McGraw-Hill, New York, 1960.
2. Feller, W. The parabolic differential equations and the associated semigroups of transformations. *Ann. Math. 55* (1952), 468–519.
3 Feller, W Diffusion processes in one dimension. *Trans Amer. Math. Soc 77* (1954), 1–31.
4. Feller, W. *An Introduction to Probability Theory and Its Applications, Vols. I, II.* Wiley, New York, 1966.
5. Gaver, D. P. Diffusion approximations and models for certain congestion problems *J. Appl. Probabil. 5* (1968), 607–623.
6 Gaver, D. P., and Shedler, G. S. Approximate models for processor utilization in multiprogrammed computer systems Res. Rep., Naval Postgraduate School, Monterey, Calif., Sept. 1972
7. Gaver, D. P , and Shedler, G. S. Processor utilization in multiprogramming systems via diffusion approximations *Oper. Res. 21* (1973), 569–576.
8. Gelenbe, E. Modèles de systèmes informatiques. Ph D. Th., Etat ès Sciences Mathématiques, U. de Paris, Paris, France, 1973.
9 Haji, R , and Newell, G. F. A relation between stationary queue and waiting time distributions. *J. Appl. Probabil. 8*, 3 (1971), 617–620
10. Jackson, J. R. Jobshop-like queueing systems. *Manage Sci. 10* (1963), 131–142.
11. Kobayashi, H. Application of the diffusion approximation to queueing networks I: Equilibrium queue distributions. *J ACM 21*, 2 (April 1974), 316–328.
12. Newell, G F. *Applications of Queueing Theory.* Chapman and Hall, London, 1971.
13. Reiser, M., and Kobayashi, H Accuracy of the diffusion approximation for some queueing systems *IBM J. Res Devel 18*, 2 (March 1974), 110–124
14 Shedler, G. S. A cyclic queue model of a paging machine. IBM Res. Rep. RC-2814, IBM Watson Res Center, Yorktown Heights, N Y., March 1970.

RECEIVED MARCH 1974, REVISED JULY 1974