Application of the Diffusion Approximation to Queueing Networks II: Nonequilibrium Distributions and Applications to Computer Modeling

HISASHI KOBAYASHI

IBM Thomas J. Watson Research Center, Yorktown Heights, New York

ABSTRACT: Quite often explicit information about the behavior of a queue over a fairly short period is wanted. This requires solving the nonequilibrium solution of the queue-length distribution, which is usually quite difficult mathematically. The first half of Part II shows how the diffusion process approximation can be used to answer this question. A transient solution is obtained for a cyclic queueing model using the technique of eigenfunction expansion. The second half of Part II applies the earlier results of Part I to modeling and performance problems of a typical multiprogrammed computer system. Such performance measures as utilization, throughput, response time and its distribution, etc., are discussed in some detail.

KEY WORDS AND PHRASES: nonequilibrium theory of queues, transient behavior, diffusion equation, cyclic queueing system, eigenfunction expansion, partial differential equation, computer system performance evaluation, performance measure, utilization, system throughput, response time

CR CATEGORIES: 4.32, 5.13, 5.17, 5.5

1. *Transient Behavior of Queue Distributions*

In Part I [5], we were concerned solely with limiting, or equilibrium, probability distributions of queue sizes in a network of queues. These equilibrium queue distributions provide information about the properties of the queueing network averaged over a long time period. Equally important to practical applications, however, is the rate of convergence to the equilibrium. The workload of a computer system, for example, is rarely stationary over a long period of time, and therefore we need explicit information about the behavior over a fairly short interval. In measuring operational systems or simulation models, some estimate of the transient time is required in determining a priori the length of sufficient observation period and/or appropriate sampling rate, since the uncertain limits of statistical estimates are usually given in terms of independent samples. Cox and Smith [2] discuss some other examples where nonequilibrium theory is needed.

The nonequilibrium theory of queues is usually much more difficult mathematically than the equilibrium theory. Consider, for example, an M/M/1 system, i.e. a single-server queue with Poisson arrivals and exponential service times. The transient solution to this simplest queueing system is given in terms of infinite series of the hyperbolic Bessel functions (see [2, p. 64]) and is far from convenient. Difficulty in obtaining solutions to more complicated cases is quite apparent, and thus we shall have to content ourselves with approximate solution. The diffusion process approximation can be useful in this regard, as we will demonstrate in the rest of this section.

Copyright © 1974, Association for Computing Machinery, Inc. General permission to republish, but not for profit, all or part of this material is granted provided that ACM's copyright notice is given and that reference is made to the publication, to its date of issue, and to the fact that reprinting privileges were granted by permission of the Association for Computing Machinery.

An earlier version of this paper was presented at the Seventh Annual Princeton Conference on Information Sciences and Systems, Princeton University, Princeton, N. J., March 22–23, 1973 [4].

Author's address: IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598.

1.1. A SINGLE-SERVER QUEUE. Recall that $x(t)$ defined in [5] represents the diffusion process which approximates the length of a single-server queue for which the mean value and squared coefficient of variation of interarrival time are μ_a and c_a, respectively, and those of service time, μ_s and c_s. Then the probability density function of $x(t)$ given that $x(0) = x_0$ is given by [5, eq. (2.4)]:

$$p(x_0, x; t) = (\partial/\partial x)\{\Phi((x - x_0 - \beta t)/(\alpha t)^{\frac{1}{2}}) - \exp(2\beta x/\alpha)\Phi(-(x + x_0 + \beta t)/(\alpha t)^{\frac{1}{2}})\}, \quad (1.1)$$

where $\Phi(x) = \int_{-\infty}^{x} (2\pi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}z^2\}\, dz$, $\alpha = c_a/\mu_a + c_s/\mu_s$, and $\beta = 1/\mu_a - 1/\mu_s$. In order to obtain a coordinate-free solution, let us apply the following scaling transformation [9] to queue-length variable x and time t:

$$y = x/|\alpha/\beta|, \quad \tau = t/(\alpha/\beta^2). \quad (1.2)$$

Then for $\beta < 0$ (or equivalently $\rho = \mu_s/\mu_a < 1$), eq. (1.1) becomes

$$p(y_0, y; \tau) = (\partial/\partial y)\{\Phi((y - y_0 + \tau)/\tau^{\frac{1}{2}}) - e^{-2y}\Phi(-(y + y_0 - \tau)/\tau^{\frac{1}{2}})\}, \quad y > 0. \quad (1.3)$$

Similarly, for $\beta > 0$ (i.e. $\rho > 1$) we have

$$p(y_0, y; \tau) = (\partial/\partial y)\{\Phi((y - y_0 - \tau)/\tau^{\frac{1}{2}}) - e^{2y}\Phi(-(y + y_0 + \tau)/\tau^{\frac{1}{2}})\}, \quad y > 0. \quad (1.4)$$

The curves of Figure 1 are plots of (1.3) for values of time parameter τ, where the dashed curves are obtained for the initial value $y_0 = 0$, and the solid curves are for $y_0 = 2$. We can see that the equilibrium state is reached approximately by $\tau = 5$ for both cases. Clearly for the initial value y_0 greater than 2, the transient interval cannot be shorter than 5.

When $\beta > 0$, the queue is oversaturated and no equilibrium solution exists. Not only the mean queue size increases as time elapses, but also the variance increases; thus the density function becomes broader and broader. Figure 2 shows the distribution function (i.e. integration of (1.4)), when the initial queue size is $y_0 = 0$ and $y_0 = 2$. The results tell us how fast the queue will build up in an overloaded system.

Before we proceed further let us examine the scaling factor of transformation (1.2):

$$\alpha/\beta^2 = \mu_s(c_s + c_a\rho)/(1 - \rho)^2. \quad (1.5)$$

We see from this expression that the transient period is inversely proportional to the square of $1 - \rho$, the probability that the server is idle. A similar observation can be made on the scaling factor of the queue size: $|\alpha/\beta| = (c_s + c_a\rho)/|1 - \rho|$.

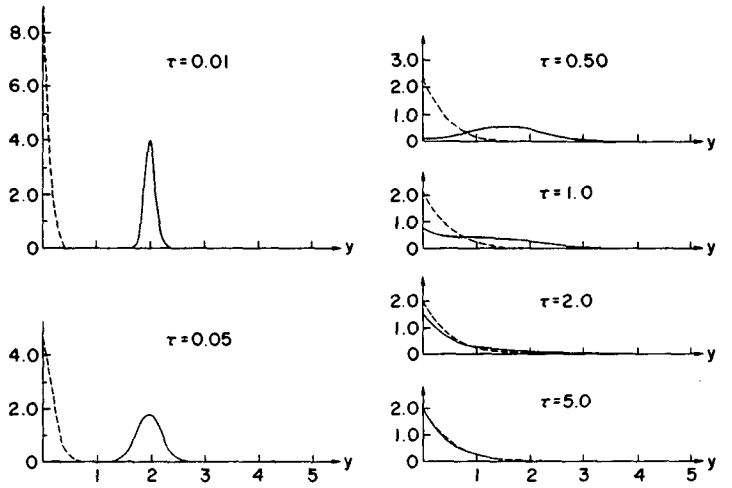


FIG. 1. The transient solution $p(y_0, y; \tau)$ for $\rho < 1$ ($\beta < 0$), with the initial condition $y_0 = 0$ (dashed curves), and $y_0 = 2$ (solid curves).

1.2. A CYCLIC QUEUEING SYSTEM. Let us now consider the cyclic queueing system which was discussed in [5, Ex. 4.1]. Service times at Station i are i.i.d. with the mean μ_i and the squared coefficient of variation c_i, $i = 1, 2$. The system is a closed queueing network; hence N, the total number of jobs in the system, remains constant. Let us denote by $x(t)$ the diffusion process which approximates the queue size $n_1(t)$ (Figure 3). Then the corresponding diffusion equation is

$$(\partial/\partial t)p(x_0, x; t) = \tfrac{1}{2}\alpha^\circ(\partial^2/\partial x^2)p(x_0 x; t) - \beta^\circ(\partial/\partial x)p(x_0, x; t), \tag{1.6}$$

where α° and β° are α_{11}° of [5, eq. (4.3)] and β_1° of [5, eq. (4.4)], respectively: $\alpha^\circ = c_1/\mu_1 + c_2/\mu_2$, $\beta^\circ = 1/\mu_2 - 1/\mu_1$. We now want to solve (1.6) with the boundary condition $0 \le x(t) \le N + 1$, for every $t \ge 0$. Note that the upper boundary is $N + 1$ instead of N, since the queue size n_1 corresponds to unit interval $n_1 \le x \le n_1 + 1$, where $n_1 = 0, 1, 2, \cdots, N$. (See [5, eq. (2.6)].) By applying the scaling transformations

$$y = x/|\,\alpha^\circ/\beta^\circ\,| = x/|\,(c_1 + c_2\rho)/(1 - \rho)\,|,$$
$$\tau = t/(\alpha^\circ/\beta^{\circ 2}) = t/\mu_1(c_1 + c_2\rho)/(1 - \rho)^2, \tag{1.7}$$

where $\rho = \mu_1/\mu_2$, we now have the coordinate-free diffusion equation

$$(\partial/\partial\tau)p(y_0, y; \tau) = \tfrac{1}{2}(\partial^2/\partial y^2)p(y_0, y; \tau) - \delta\cdot(\partial/\partial y)p(y_0, y; \tau) \tag{1.8}$$

with two reflecting barriers at $y = 0$ and $y = b$:

$$\tfrac{1}{2}(\partial/\partial y)p(y_0, y; \tau) - \delta\cdot p(y_0, y, \tau) = 0 \quad \text{at} \quad y = 0 \quad \text{and} \quad y = b, \tag{1.9}$$

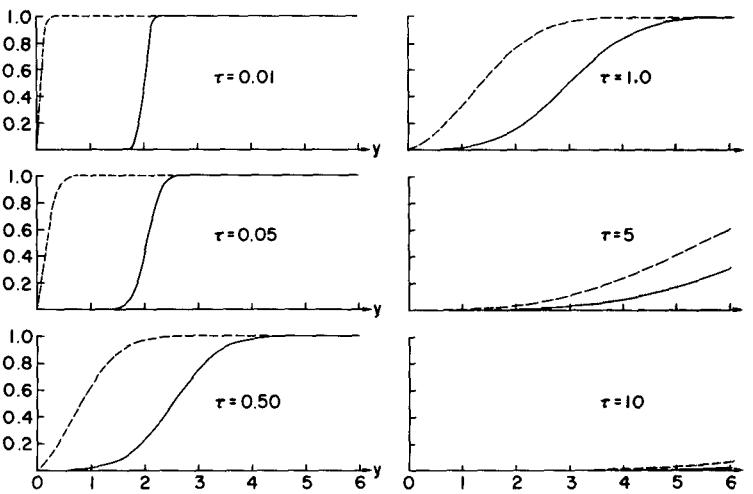
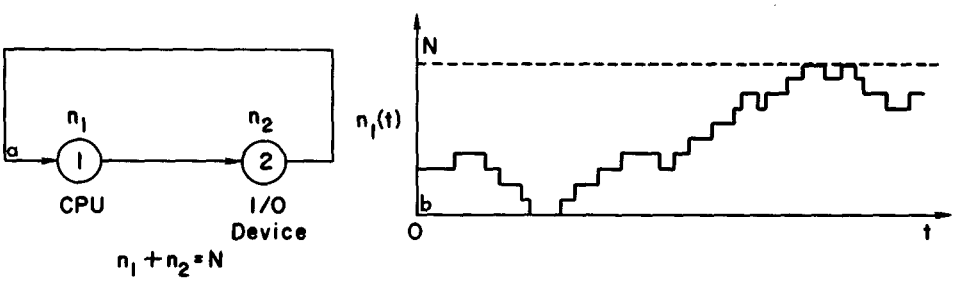


FIG. 2. The transient solution $\int_0^y p(y_0, z; \tau)\,dz$ for $\rho > 1$ ($\beta > 0$) with the initial condition $y_0 = 0$ (dashed curves), and $y_0 = 2$ (solid curves).



(a) (b)

FIG. 3. (a) A cyclic queueing model; (b) a typical behavior of the CPU queue size $n_1(t)$

where

$$\delta = \begin{cases} 1 & \text{if } \rho < 1, \\ 0 & \text{if } \rho = 1, \\ -1 & \text{if } \rho > 1, \end{cases} \tag{1.10}$$

and $b = (N+1)/|(c_1 + c_2\rho)/(1-\rho)|$. By applying the method of "eigenfunction expansion," a technique frequently used in solving partial differential equations, we obtain the following solution to (1.8) (Appendix A):

$$p(y_0, y; \tau) = \begin{cases} 2\delta e^{2\delta y}/(e^{2\delta b} - 1) + \exp[\delta(y - y_0 - \delta\tau/2)] \\ \qquad \cdot \sum_{n=1}^{\infty} \phi_n(y)\phi_n(y_0) \exp(-\lambda_n^2 \tau/2), & 0 \le y \le b, \\ 0, \quad \text{elsewhere}, \end{cases} \tag{1.11}$$

where $\phi_n(y)$'s are eigenfunctions associated with eigenvalues λ_n's:

$$\phi_n(y) = [2\lambda_n^2/b(\lambda_n^2 + 1)]^{\frac{1}{2}} \{\cos \lambda_n y + (\delta/\lambda_n) \sin \lambda_n y\},$$

and $\lambda_n = n\pi/b$, $n = 1, 2, 3, \cdots$. The first term of (1.11) represents the steady-state probability and the second term gives the transient part in terms of eigenfunction expansion. Note that (1.11) satisfies the initial condition $y = y_0$, i.e. $p(y_0, y; \tau) = \delta(y - y_0)$, since the delta function is expressed in terms of the eigenfunctions as shown by (A-7) in Appendix A. The second term of (1.11) is an infinite series, but can be well approximated by finite terms, since the factor $\exp\{-\frac{1}{2}\lambda_n^2 \tau\}$ approaches zero as n increases.

Example 1.1. Let us consider the numerical example discussed in [5, Ex. 4.2]: We choose parameters $\rho = 0.75$, $c_1 = 1$, $c_2 = 0.2$, and $N = 10$, which lead to $\delta = 1$ and $b = 2.39$. The set of curves of Figure 4 are plots of (1.11) for various values of time τ when the initial value of queue length at CPU (Station 1) is $y_0 = 0$ (i.e. CPU is empty), $y_0 = b/2$ (i.e. $n_1 = N/2 = 5$), and $y_0 = b$ (i.e. I/O is empty). As we see from these curves, the equilibrium distribution is reached approximately by $\tau = 5$ again, similar to a single server queue system shown in Figure 1. This agreement is not a coincidence. A cyclic queue system is approximately equivalent to a singer server system with a finite waiting room. The equivalence holds exactly if one of the stations has an exponential service time distribution [1, 6].

2. *Resource Utilization, System Throughput, and Response Times*

Among most frequently used performance measures for computer system performance evaluation are the CPU and device utilizations, system throughput, the response time

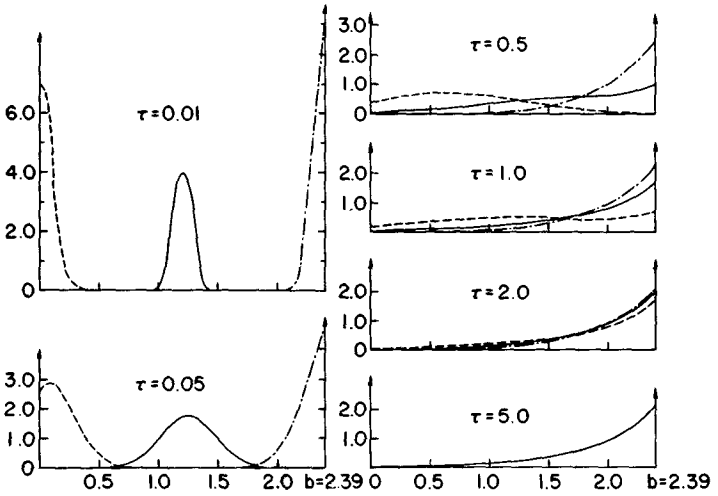


FIG. 4. The transient behavior $p(y_0, y; \tau)$ of a cyclic queueing system with the initial condition $y_0 = 0$, $y_0 = b/2$, and $y_0 = b$.

distribution, etc. In this section we define these terms precisely and show how our results based on the diffusion approximation technique can be applied to performance evaluation and prediction of a computer system.

2.1. Utilization and Throughput in a Queueing Network Model. In an open network of queues the utilization of an mth processor (or server) u_m is, according to [5, eq. (3.12)], given by

$$u_m = \rho_m = e_m \mu_m / \mu_0 \tag{2.1}$$

if the system is stable, i.e. the rightmost expression of (2.1) is less than unity for all m. Here μ_m is the average service time at processor m, μ_0 is the average interarrival time, and e_m is the average number of visits to processor m by a task during its lifetime (i.e. between its arrival and departure). Recall that vector $[1, e_1, e_2, \cdots, e_m]$ is the solution of [5, eq. (3.11)]:

$$e_m = \sum_{l=0}^{M} e_l r(l, m), \qquad 1 \leq m \leq M, \tag{2.2}$$

and e_m corresponds to the average number of visits that a job makes to Station m during its entire life in the system. Here $r(l, m)$ is the transition probability from Station l to Station m as was defined in [5].

If the task arrival rate $1/\mu_0$ is sufficiently high so that the load on some processor exceeds its processing capacity, i.e. if

$$\max_m \{e_m \mu_m\} = e_{m^*} \mu_{m^*} > \mu_0, \tag{2.3}$$

then the processor m^* becomes a bottleneck of the system and its utilization (busy factor) is 100 percent:

$$u_{m^*} = 1 \tag{2.4}$$

and

$$u_m = e_m \mu_m / e_{m^*} \mu_{m^*}, \qquad \text{for} \quad m \neq m^*. \tag{2.5}$$

By combining (2.1)–(2.5), a general expression for utilization is given by

$$u_m = e_m \mu_m / \max_{l \in [0,1,\cdots,M]} \{e_l \mu_l\}, \tag{2.6}$$

where

$$e_0 = 1. \tag{2.7}$$

The system throughput is usually defined as the number of tasks which can be completed per unit of time. In an open network system, the amount of service that one task requires of processor m is, on the average, equal to $e_m \mu_m$; therefore the system throughput rate, r, is given by $r = u_m / e_m \mu_m$, which is then written, using (2.6) and (2.7), as

$$r = \min_{l \in [0,1,\cdots,M]} \{1/e_l \mu_l\} = \min \{1/\mu_0, 1/e_{m^*} \mu_{m^*}\}. \tag{2.8}$$

Equation (2.8) shows that once the system becomes saturated an increase in the system throughput can be achieved only by increasing the processing speed of bottleneck processor m^*, i.e. by effectively decreasing the value μ_{m^*}.

In a closed network $\mathbf{e}' = [e_1, e_2, \cdots, e_m]$ is a left eigenvector of the Markov matrix $[r^\circ(l, m)]$ associated with eigenvalue being unity, where we assume that this Markov chain is irreducible. (See [5, eq. (4.6)].) Although the solution vector is not unique (since $k\mathbf{e}$ is also a solution for arbitrary k), we can still interpret that e_m is proportional to the frequency of visits to processor m by a task. By imposing an additional constraint $\sum_{m=1}^{M} e_m = 1$, the vector $\mathbf{e}$ is now uniquely defined and it corresponds to the stationary probability vector of the Markov chain. The resource utilization formula (2.6) is, however, not applicable to a closed network model, since $e_0 \mu_0$ is undefined here. The only thing that remains to hold is the property that u_m is proportional to $e_m \mu_m$, i.e. $u_m = e_m \mu_m / L$.

It seems that there is no simple way to determine constant L except for those cases in which we know the queue size distribution exactly. However, we see that L is bounded by

$$L \geq e_{m^*}\mu_{m^*} = \max_{m \in [1,2,\cdots,M]} \{e_m\mu_m\}.$$

If N, the total number of tasks in the system, is sufficiently large, then the processor m^* will be almost always busy; thus the formula (2.5) can be a good approximation in this case. This approximation was in fact used in [5, Sec. 4] to approximate the queue size distribution (see [5, eq. (4.8)].

2.2. The Average Response Time and Response Time Distributions. The average response time, T, in equilibrium state is related to r, the system throughput, and $\bar{n}$, the average number of tasks residing in the system, according to the following simple formula:

$$rT = \bar{n}, \tag{2.9}$$

which is often called Little's formula [7].

In a single server model discussed in [5, Sec. 2], we can compute $\bar{n}$ based on the approximate distribution [5, eq. (2.9)] as

$$\bar{n} = \sum_{n=1}^{\infty} np(n) = \rho/(1 - \hat{\rho}), \tag{2.10}$$

where $\hat{\rho} = \exp\{-2 \mid \beta \mid/\alpha\} = \exp\{-2(1 - \rho)/(c_s + c_a\rho)\}$ when $\rho = \mu_s/\mu_a < 1$. Recall that if $c_s = c_a = 1$, it follows that $\hat{\rho} \simeq \rho$. If c_a and/or c_s is greater than one, then $\hat{\rho} < \rho$ and the system tends to have a longer queue.

Let us consider an M/G/1 system, i.e. Poisson arrivals and a general service time distribution. Then $c_a = 1$ and $\hat{\rho} = \exp\{-2(1 - \rho)/(c_s + \rho)\} \triangleq f(\rho)$. By applying the Taylor series expansion to $f(\rho)$ around $\rho = 1$, we obtain

$$\hat{\rho} = f(\rho) = f(1) + (\rho - 1)f'(1) + \tfrac{1}{2}(\rho - 1)^2 f''(1) + \cdots = 1 + 2(\rho - 1)/(1 + c_s) + \cdots. \tag{2.11}$$

Therefore by substituting (2.11) into (2.10) we have the following approximate result:

$$\bar{n} \simeq \rho(1 + c_s)/2(1 - \rho). \tag{2.12}$$

On the other hand the exact solution of $\bar{n}$ for M/G/1 is known and given by

$$\bar{n} = \rho + \rho^2(1 + c_s)/2(1 - \rho) = \rho(1 + c_s)/2(1 - \rho) + (1 - c_s)/2,$$

which is often called the Pollaczek-Khinchine formula [2]. Therefore we see that (2.12) is a good approximation when either ρ or c_s is close to 1.

The system throughput of a single server system is, from (2.8), given by

$$r = \min\{1/\mu_a, 1/\mu_s\} = 1/\mu_a \tag{2.13}$$

when $\rho < 1$. The average response time, T, is therefore given from (2.9), (2.10), and (2.13) by $T = \bar{n}/r = \mu_s/(1 - \hat{\rho}) = \hat{\rho}/(1 - \hat{\rho})\mu_s + \mu_s$. The first term in the last equation is the average waiting time, and the second term is clearly the average processing time.

For an M/G/1 system we can obtain approximate solutions to waiting and response time distributions which are easier to calculate than a general formula given in terms of Laplace transform (see [2, p. 57]). Let $q(l)$ be the probability of the queue length (excluding a task in service) $l = n - 1$ conditioned that the system is nonempty, i.e. $q(l) = \Pr\{n = l + 1 \mid n > 0\} = p(l + 1)/(1 - p(0))$, $l = 0, 1, 2, \cdots$. Using the approximate distribution $p(n)$ of [5, eq. (2.9)], we obtain $q(l) = (1 - \hat{\rho})\hat{\rho}^l$, $l \geq 0$. We define a probability generating function $Q(z)$ by

$$Q(z) = \sum_{l=0}^{\infty} q(l)z^l = (1 - \hat{\rho})/(1 - \hat{\rho}z). \tag{2.14}$$

Let $w(t)$ be the probability density function of waiting time t. It is clear that $w(t)$ can be represented as

$$w(t) = (1 - \rho)\delta(t) + \rho w_1(t), \tag{2.15}$$

where $w_1(t)$ is the probability density of waiting time conditions that it is nonzero. Since the number of arrivals during $[0, t]$ is Poisson distributed with parameter t/μ_a, an alternative derivation of $q(l)$ is given by $q(l) = \int_0^\infty 1/l!(t/\mu_a)^l \exp(-t/\mu_a) w_1(t)\, dt$, and its generating function is

$$Q(z) = \int_0^\infty \sum_{l=0}^{\infty} 1/l!(tz/\mu_a)^l \exp(-t/\mu_a) w_1(t)\, dt = w_1^*((1 - z)/\mu_a),$$

where $w_1^*(s)$ is the Laplace transform of $w_1(t)$. By equating (2.14) and (2.11) we obtain $w_1^*(s) = [(1 - \hat{\rho})/\hat{\rho}\mu_a]/[s + (1 - \hat{\rho})/\hat{\rho}\mu_a]$. Hence by inverting $w_1^*(s)$ and substituting the result into (2.15) we have the following approximation formula of waiting time:

$$w(t) = (1 - \rho)\delta(t) + [\rho(1 - \hat{\rho})/\hat{\rho}\mu_a] \exp[-(1 - \hat{\rho})t/\hat{\rho}\mu_a], \qquad t \geq 0.$$

The response time density function is therefore given by

$$T(t) = w(t) \otimes s(t) = (1 - \rho)s(t) + [\rho(1 - \hat{\rho})/\hat{\rho}\mu_a]s(t) \otimes \exp[-(1 - \hat{\rho})t/\hat{\rho}\mu_a],$$

where $s(t)$ is the probability density function of service time and $\otimes$ means convolution.

Let us now consider the response time of a network of queues. The expectation of the total number of tasks residing in the system is computable from [5, eqs. (3.8) and (3.13)] as $\bar{n} = \sum_{m=1}^{M} \bar{n}_m = \sum_{m=1}^{M} [\rho_m/(1 - \hat{\rho}_m)]$ when the system is stable. The average response time of the network system is $T = \bar{n}/r = \sum_{m=1}^{M} e_m T_m$, where T_m is the average time that a task spends at processor m and its queue and is given by an expression analogous to (2.13): $T_m = [\hat{\rho}_m/(1 - \hat{\rho}_m)]\mu_m + \mu_m$. The probability density functions of waiting time and response time at each processor can be computed in exactly the same manner as obtained for a single server model:

$$w_m(t) = (1 - \rho_m)\delta(t) + [\rho_m(1 - \hat{\rho}_m)/\hat{\rho}_m\mu_m] \exp[-(1 - \hat{\rho}_m)t/\hat{\rho}_m\mu_m],$$

$$T_m(t) = w_m(t) \otimes s_m(t),$$

where ρ_m and $\hat{\rho}_m$ are defined by [5, eq. (3.10)] and [5, eq. (3.14)], respectively; $s_m(t)$ is the probability density function of processing time at processor m. The waiting time and response time distributions in the *whole* network are difficult to obtain, since the waiting times of, say, two consecutive visits of a task to the same processor are clearly correlated. The problems can be formulated as the first-passage time problems for diffusion processes and will be discussed in a separate report.

3. *Conclusions*

In Section 1 of this paper derivations of transient (nonequilibrium) solutions of queue length distributions in single server and cyclic queue models are presented with some numerical examples. In those cases equilibrium state is almost reached by $\tau = 5$, where time τ is the normalized time defined by (1.5) or (1.7). The scaling transformation indicates that the transient time is longer when the system is heavily loaded ($\rho \simeq 1$) and the distributions of interarrival time and service time have large variances. These results could be very useful in many respects: the sampling rate and observation period can be chosen appropriately in measurements of operational systems or simulation models; adaptive scheduling or resource allocation schemes require an appropriate choice of the observation period (e.g. integration or smoothing time), since the interval should be short enough to follow the change of input processes and yet should be long enough to allow a sufficient number of sampled data.

Section 2 presents how to use the diffusion approximation solutions to performance

evaluation of computer systems. Such performance measures as utilization, throughput, the average and distribution of response time, etc., are clearly defined and their quantitative relationships are investigated.

Appendix A. Derivation of the Transient Solution (1.11)

There are at least three different methods to solve the diffusion equation. They are the method of images [10], the method of separation of variables [8, 11], and the Laplace transform method [3]. Here we use the second method and follow the treatment by Sweet and Hardin [11].

Let us assume the following solution form: $p(y_0, y; \tau) = q(y_0, y)e^{\delta y} \cdot r(\tau)$, and substitute this into (1.8). Then we obtain the following two equations which are interrelated via unknown parameter d^2:

$$q''(y_0, y) + (d^2 - \delta^2)q(y_0, y) = 0 \tag{A-1}$$

and

$$\dot{r}(\tau) + (d^2/2)r(\tau) = 0, \tag{A-2}$$

where $'$ and $\cdot$ denote differentiation with respect to y and τ, respectively. Then by the standard techniques for ordinary differential equations, we find that

$$\phi(y) = Ae^{j\lambda y} + Be^{-j\lambda y} \tag{A-3}$$

satisfies (A-1) where $\lambda^2 = d^2 - \delta^2$. By imposing the boundary condition (1.9) on (A-3), we find that the following equation must be met by λ:

$$(\delta^2 + \lambda^2) \sin \lambda b = 0. \tag{A-4}$$

Furthermore the constants A and B must satisfy $(\delta - j\lambda)A + (\delta + j\lambda)B = 0$. Thus the eigenvalues which satisfy (A-4) are $\lambda = \pm j\delta$, $\lambda = 0$, and $\lambda_n = n\pi/b$, $n = \pm 1, \pm 2, \pm 3, \cdots$. However, the eigenfunctions of the form (A-3) are the same for λ_n and λ_{-n}. For $\lambda = 0$ we find the corresponding function $\phi(y) = 0$. Therefore we need only consider the following eigenvalues: $\lambda_0 \triangleq j\delta$, $\lambda_n = n\pi/b$, $n = 1, 2, 3, \cdots$, and the eigenfunctions associated with these eigenvalues are $\phi_0(y) = [2\delta/(e^{2b\delta} - 1)]^{\frac{1}{2}} e^{\delta y}$ and

$$\phi_n(y) = [2\lambda_n^2/b(\lambda_n^2 + 1)]^{\frac{1}{2}} \{\cos \lambda_n y + (\delta/y_n) \sin \lambda_n y\}, \quad n = 1, 2, 3, \cdots.$$

The functions $\phi_0(y), \phi_1(y), \phi_2(y), \cdots$ are orthonormal and form a complete set for all differentiable functions whose support is $[0, b]$. Once λ is determined, the corresponding solution to (A-2) is given by $r(\tau) = \exp[-\frac{1}{2}(\lambda^2 + \delta^2)\tau]$. Therefore, we arrive at the following form for a general solution:

$$p(y_0, y; \tau) = \sum_{n=0}^{\infty} a_n \phi_n(y) \exp[\delta y - \tfrac{1}{2}(\lambda_n^2 + \delta^2)\tau]. \tag{A-5}$$

The constants $a_0, a_1, a_2, \cdots$ are to be determined by the initial condition $y(0) = y_0$, which is equivalent to

$$p(y_0, p; 0) = \delta(y - y_0), \qquad 0 \leq y \leq b, \tag{A-6}$$

where $\delta(\cdot)$ is the delta function[1] and can be expanded in terms of eigenfunctions as follows [8]:

$$\delta(y - y_0) = \sum_{n=0}^{\infty} \phi_n(y)\phi_n(y_0). \tag{A-7}$$

A careful observation of (A-5) with τ being set to zero suggests replacing (A-6) by the following equivalent condition:

[1] $\delta(\cdot)$ has no relation with the variable defined by (1.10).

$$p(y_0, y; \tau) = \exp(\delta y - \delta y_0)\delta(y - y_0) = \exp(\delta y - \delta y_0)\sum_{n=0}^{\infty}\phi_n(y)\phi_n(y_0). \quad \text{(A-8)}$$

Then by comparing the coefficients of (A-5) and (A-8) we obtain $a_0 = [2\delta/(e^{2\delta b} - 1)]^{\frac{1}{2}}$, $a_n = \exp(-\delta y_0)\phi_n(y_0)$. Therefore the solution of the diffusion equation (1.8) is given by

$$p(y_0, y; \tau) = 2\delta e^{2\delta y}/(e^{2\delta b} - 1) + \exp(\delta y - \delta y_0 - \tfrac{1}{2}\delta^2\tau)\sum_{n=1}^{\infty}\phi_n(y)\phi_n(y_0)\exp(-\tfrac{1}{2}\lambda_n{}^2\tau).$$

Appendix B. Glossary of Symbols for Parts I and II

$A(t)$: cumulative number of arrivals up to time t.
$A_m(t)$: cumulative number of arrivals at station m up to time t.
$\alpha = c_a/\mu_a + c_s/\mu_s$: increasing rate of the variance of diffusion process $x(t)$.
$\boldsymbol{\alpha}$: increasing rate of the covariance matrix of $\mathbf{x}(t)$ in an open network.
$\boldsymbol{\alpha}^\circ$: increasing rate of the covariance matrix of $\mathbf{x}(t)$ in a closed network.
$\boldsymbol{\alpha}^{\circ-}$: pseudo (generalized) inverse of singular matrix $\boldsymbol{\alpha}^\circ$.
$\alpha_{mm'}$: (m, m') element of $\boldsymbol{\alpha}$.
$b = (N + 1)/|\alpha/\beta|$: normalized reflecting barrier.
$\beta = 1/\mu_a - 1/\mu_s$: increasing rate of the mean of diffusion process $x(t)$.
$\boldsymbol{\beta}$: increasing rate of the mean vector of $\mathbf{x}(t)$ in an open network.
$\boldsymbol{\beta}^\circ$: increasing rate of the mean vector of $\mathbf{x}(t)$ in a closed network.
β_m: mth component of $\boldsymbol{\beta}$.
$c_a = \sigma_a{}^2/\mu_a{}^2$: squared coefficient of variation of interarrival time.
$c_s = \sigma_s{}^2/\mu_s{}^2$: squared coefficient of variation of service time in a single server system.
$c_m = \sigma_m{}^2/\mu_m{}^2$: squared coefficient of variation of service time at station m.
$\boldsymbol{\gamma} = 2\boldsymbol{\alpha}^{-1}\boldsymbol{\beta}$: (defined for an open network of queues).
$\boldsymbol{\gamma}^\circ = 2\boldsymbol{\alpha}^{\circ-}\boldsymbol{\beta}^\circ$: (defined for a closed network of queues).
γ_m: mth component of $\boldsymbol{\gamma}$.
$D(t)$: cumulative number of departures up to time t.
$D_l(t)$: cumulative number of departures from station l up to time t.
$D_{l,m}(t)$: cumulative number of jobs up to time t which leave station l and join station m.
$\delta = \operatorname{sgn}(1 - \rho)$: $\delta = +1, 0$, or -1, depending on ρ being less than, equal to, or greater than unity.
$\delta(y)$: delta function.
e_m: average number of visits to station m by a task during its entire life in the system.
K: normalization constant for the queue-length distribution of a closed network of queues.
$\lambda_n = n\pi/b$: eigenvalue of the diffusion equation for a cyclic queue system.
M: number of stations in a network.
μ_a: mean interarrival time.
μ_s: mean service time in a single server system.
μ_m: mean service time at station m in a network of queues.
N: number of tasks in a closed network (or degree of multiprogramming in a closed network model).
$\bar{n}$: average number of tasks in a single server or open network system.
$p(n)$: probability of queue length being n.
$\hat{p}(n)$: approximation of $p(n)$.
$p(n_1, n_2, \cdots, n_M)$: joint probability of queue lengths being $n_1, n_2, \cdots, n_M$ at stations $1, 2, \cdots, M$, respectively.
$\hat{p}(n_1, n_2, \cdots, n_M)$: approximation of $p(n_1, n_2, \cdots, n_M)$.
$p(x_0, x; t)$: probability density function of diffusion process $x(t)$ given that $x(0) = x_0$.
$p(x) = p(x_0, x; \infty)$: probability density function of diffusion process $x(t)$ at equilibrium state.

$p(\mathbf{x}_0, \mathbf{x}; t)$: joint probability density function of M-dimensional diffusion process $\mathbf{x}(t)$ given that $\mathbf{x}(0) = \mathbf{x}_0$.

$p(\mathbf{x}) = p(\mathbf{x}_0, \mathbf{x}; \infty)$: joint probability density function of M-dimensional process $\mathbf{x}(t)$ at equilibrium state.

$p(y_0, y; \tau)$: the probability density function of diffusion process $y(\tau)$ (after scaling transformation) given that $y(0) = y_0$.

$\Phi(x)$: standard normal integral, i.e. integration of unit normal distribution function over $(-\infty, x]$.

$\phi_n(y)$: eigenfunction associated with eigenvalue λ_n.

$Q(t) = A(t) - D(t)$: queue length (including one in service) at time t.

$\Delta Q(t) = Q(T + \Delta) - Q(t)$: change in queue length during $(t, t + \Delta]$.

$Q_m(t) = A_m(t) - D_m(t)$: queue length (including one in service) of station m at time t.

$\Delta Q_m(t) = Q_m(t + \Delta) - Q_m(t)$: change in queue length at station m during $(t, t + \Delta]$.

$Q(z)$: probability generating function of $\{q(l)\}$.

$q(l) = p(l + 1)\{1 - p(0)\}$: probability of queue length l (excluding one in service), conditioned that the system is not idle.

r: system throughput rate.

$r(m, m')$: transition probability from station m to station m' in an open network.

$r^\circ(m, m')$: transition probability from station m to station m' in a closed network.

$\rho = \mu_s/\mu_a$: utilization factor in a single server system.

$\hat{\rho} = \exp\{2\beta/\alpha\}$, $(\beta < 0)$: parameter that characterizes $\hat{p}(n)$ of a single server system.

$\hat{\rho}_m = \exp \gamma_m$: parameter that characterizes the mth marginal distribution of $\hat{p}(n_1, n_2, \cdots, n_m, \cdots, n_M)$.

s: service time (random variable).

σ_a^2: variance of interarrival time.

σ_s^2: variance of service time in a single server system.

σ_m^2: variance of service time at station m in a network of queues.

T: average response time.

T_m: average time that a task spends at station m (including queueing times).

$T(t)$: probability density function of response time t.

$\tau = t/(\alpha/\beta^2)$: normalized time to make the diffusion equation coordinate-free.

u_m: utilization factor of station (or processor) m.

$w(t)$: probability density function of waiting time t.

$w_1(t)$: probability density function of waiting time t conditioned that $t > 0$.

$w_1^*(t)$: Laplace transform of $w_1(t)$.

$x(t)$: diffusion process which approximates $Q(t)$.

x: random variable $x(t)$ with time parameter t being suppressed.

$\mathbf{x}(t)$: M-dimensional diffusion process which approximates joint queue lengths $\{Q_1(t), Q_2(t), \cdots, Q_M(t)\}$.

$x_m(t)$: mth element of $\mathbf{x}(t)$.

$y = x/|\alpha/\beta|$: normalized x to make the diffusion equation coordinate-free.

$z(t)$: white Gaussian process with zero mean and unit variance.

$\mathbf{z}(t)$: M-dimensional white Gaussian process with each component possessing zero mean, unit variance, and zero cross-variance.

REFERENCES

1. ADIRI, I. Queueing models for multiprogrammed computers. In Proc. Symposium on Computer-Communications Networks and Teletraffic, Vol. XXII, Polytechnic Press of the Polytechnic Institute of Brooklyn, Brooklyn, N. Y., April 4–6, 1972, pp. 441–448.
2. COX, D. R., AND SMITH, W. L. *Queues*. Methuen, London, 1961.
3. FELLER, W. *An Introduction to Probability Theory and Its Applications, Vol. II*. Wiley, New York, 1966.
4. KOBAYASHI, H. Applications of the diffusion approximation to queuing networks, Part II. Proc. Seventh Annual Princeton Conference on Information Science and Systems, Princeton

U., Princeton, New Jersey, March 22-23, 1973, pp. 448-454; also IBM Res. Rep. RC-4054, Sept. 1972.

5. KOBAYASHI, H. "Application of the diffusion approximation to queueing networks I: Equilibrium queue distributions. *J. ACM 21*, 2 (April 1974), 316-328.
6. KOBAYASHI, H., AND SILVERMAN, H. F. Some dispatching priority schemes and their effects on response time distribution. IBM Res. Rep. RC 3584, Oct. 1971.
7. LITTLE, J. D. C. A proof for the queueing formula $L = \lambda w$. *Oper. Res. 9* (1961), 383-387.
8. MORSE, P. M., AND FESHBACH, H. *Method of Theoretical Physics*. McGraw-Hill, New York, 1953, Pt. I, Ch. 6.
9. NEWELL, G. F. *Applications of Queueing Theory*. Chapman and Hall, London, 1971.
10. SOMMERFELD, A. *Partial Differential Equations in Physics*. Academic Press, New York, 1949.
11. SWEET, A. L., AND HARDIN, J. C. Solutions for some diffusion processes with two barriers. *J. Appl. Prob. 7* (1970), 423-431.

RECEIVED OCTOBER 1972; REVISED JUNE 1973