M. Reiser
H. Kobayashi

Accuracy of the Diffusion Approximation for Some Queuing Systems

Abstract: This paper presents the results of a rather extensive study of the accuracy of the diffusion approximation technique as applied to queuing models. The motive for using the diffusion process approximation here is to develop realistic analytical models of computing systems by considering service time distributions of a general form. We first review the theory of the diffusion approximation for a single server and then develop a new and simplified treatment of a queuing network system. The accuracy of this approximation method is then considered for a wide class of distributional forms of service and interarrival times and for various queuing models. The approximate solutions and exact (or simulation) solutions are compared numerically in terms of the means and variances of queue sizes, server utilizations, the asymptotic decrements of the distributions, and the queue size distributions themselves.

The accuracy of the diffusion approximation is found to be quite adequate in most cases and is considerably higher than that obtained by an exponential server model that is prevalent in computer system modeling.

1. Introduction

The diffusion approximation is an attempt to overcome the limitations of exponential server queuing models by considering both the mean and variance of the service time distributions. It is based on the assumption that queues are almost always nonempty. The central limit theorem is then applied to characterize the fluctuations in the queue lengths, and the discrete-valued queuing process is replaced by a continuous-path Markov process (also called a diffusion process) with a similar distribution of the infinitesimal increments. The probability distribution of this continuous process is then described by a diffusion equation, which has to be solved with appropriate boundary conditions. Applications of the diffusion approximation to queuing systems have been discussed by Cox and Miller [1], Gaver [2], Newell [3], Gaver and Shedler [4], and Kobayashi [5].

This paper presents a simplified treatment of queuing networks but concentrates primarily on an investigation of the accuracy of the diffusion approximation by means of a consistent set of examples. The single server queue is considered in detail, since it is the element used in the treatment of networks. Specifically, exact analytical solutions for the steady state queue size distributions of the M/G/1 and the GI/M/1 queues are derived. In the case of the M/G/1 queue, the Erlangian distribution is used as a model for service more regular than completely random, whereas the hyperexponential distribution is used as a model for long-tailed distributions. In general, the diffusion approximation leads to favorable results. Its accuracy increases as the traffic intensity approaches one, which is not unexpected from the assumptions made.

Analytical solutions are difficult to obtain for queuing networks with nonexponential holding times. An exact solution for cyclic tandem queues with one general server is reported in Appendix D, and it is used for comparison purposes. In general, however, one has to resort to simulation. To keep the simulation time within reasonable bounds, relatively simple examples of networks are used. Again, in most cases a satisfactory agreement between simulation and the diffusion approximation was observed.

In the next section the diffusion approximation of the single server queue is briefly reviewed. Then a theory for queuing networks is introduced that is based on the additional assumption that each server may be treated independently of the others. The accuracy of the diffusion approximation of the queue size distribution of the single server queue is evaluated next. Simple examples of networks are discussed, and results using the diffusion approximation are compared with some analytical results and also with the results of simulations. Finally, the results are summarized and discussed from a unifying point of view.

Appendix A summarizes definitions and properties of the Erlangian and the hyperexponential distributions. Explicit closed-form formulas for the steady state queue size of the M/G/1 queue with Erlangian and with hyper-

exponential service time distributions are given in Appendix B. In Appendix C Erlangian input and general service time distribution are considered. A simple closed-form solution is given for the $E_m/G/1$ queue. In the last Appendix, D, the analytical solution for cyclic tandem queues with one general server is given in terms of the results of Appendices B and C.

2. Diffusion approximation for the single server queue

We briefly review here the basic assumptions leading to the diffusion approximation for the GI/G/1 (i.e., general independent interarrival time distribution/general service time distribution/a single server) queuing system. For a more detailed treatment, see [5].

• *Assumption of a normal distribution for queue size fluctuations*

Let $\Delta Q(t)$ be the change of queue length between times t and $t+\Delta$. Then, for Δ sufficiently large, $\Delta Q(t)$ should be approximately normally distributed with

$$E[\Delta Q(t)] = (\lambda - \mu)\Delta = \beta\Delta; \tag{1}$$

$$\text{var}[\Delta Q(t)] \cong (C_a\lambda + C_s\mu)\Delta = \alpha\Delta, \tag{2}$$

where λ is the arrival rate, μ is the processing rate (or the inverse of the mean holding time), C_a is the squared coefficient of variation of the interarrival time τ_a, i.e., $C_a = \text{var}[\tau_a]\lambda^2$, and C_s is the squared coefficient of variation for the service time τ_s, i.e., $C_s = \text{var}[\tau_s]\mu^2$.

• *Replacement of the discrete process by a continuous process*

The discrete-valued queueing process $Q(t)$ is approximated by a continuous-path process $x(t)$ with incremental changes $dx(t)$ that are normally distributed with mean βdt and variance αdt, i.e.,

$$dx(t) = \beta dt + z(t)(\alpha dt)^{\frac{1}{2}}, \tag{3}$$

where $z(t)$ is a white Gaussian process. If there is no boundary condition imposed on $x(t)$, then $x(t)$ is a Brownian motion with drift, which has a probability distribution $p(x_0, x; t)$ satisfying

$$\frac{\partial p(x_0, x; t)}{\partial t} = \frac{\alpha}{2}\left(\frac{\partial^2 p(x_0, x; t)}{\partial x^2}\right) - \beta\left(\frac{\partial p(x_0, x; t)}{\partial x}\right), \tag{4}$$

where x_0 is the initial value, and $p(x_0, x; t)dx = P\{x \le x(t) \le x + dx | x(0) = x_0\}$.

• *Introduction of appropriate boundary conditions*

The diffusion equation is now solved with the boundary condition $x(t) \ge 0$ (reflecting barrier) or $p(x_0, x; t) = 0$ for $x < 0$. For the stationary case, the time derivative in Eq. (4) is set to zero. Then the obvious requirement $\int_0^\infty p(x_0, x; \infty)dx = 1$ leads to the well-known stability condition $\beta < 0$ or $\lambda < \mu$ and to the boundary condition [1]

$$\frac{\alpha}{2}\left(\frac{dp(x_0, x; \infty)}{dx}\right) - \beta p(x_0, x; \infty) = 0 \text{ at } x = 0. \tag{5}$$

With this boundary condition, the steady state solution of Eq. (4), which is subsequently called $p(x)$, is uniquely determined to be

$$p(x) = \frac{2|\beta|}{\alpha}\exp\left(-\frac{2|\beta|x}{\alpha}\right). \tag{6}$$

• *Interpretation of the diffusion process and adjustment for small queue sizes*

The steady-state solution of the diffusion process is the exponential distribution of Eq. (6). We now go back to the discrete-valued queuing process for which we interpret Eq. (6) as a geometrical distribution of the queue size variable n with the same decrement factor $\exp(-2|\beta|/\alpha)$. By the very nature of the basic assumption (i.e., the use of the central limit theorem), we cannot expect meaningful results for small queue size n. For general interarrival and service time distributions, however, the probability of an empty queue is known to be exactly $1 - \lambda/\mu$. We then adjust the geometrical distribution at $n = 0$ and use $\exp(-2|\beta|/\alpha)$ as the decrement factor. If we denote the approximate queue size distribution constructed accordingly by $\hat{p}(n)$, we get

$$\hat{p}(n) = \begin{cases} 1 - \rho & \text{if } n = 0 \\ \rho(1-\hat{\rho})\hat{\rho}^{n-1} & \text{if } n \ge 1, \end{cases} \tag{7}$$

with

$$\hat{\rho} = \exp\left[-\frac{2(\mu - \lambda)}{\mu C_s + \lambda C_a}\right] = \exp\left[-\frac{2(1-\rho)}{C_s + \rho C_a}\right], \tag{8}$$

where $\rho = \lambda/\mu$.

3. Diffusion approximation for queuing networks

We consider a network with M single server stations in which:

1. The holding time distribution at each station $m \varepsilon [1, M]$ has the mean μ_m^{-1} (μ_m is the processing rate) and squared coefficient of variation C_m.
2. Customers (or jobs) make instantaneous transitions from station m to station m' with probability $\theta_{mm'}$. Probability $\theta_{mm'}$ is independent of the state of the system (i.e., the routing of each customer is generated by a Markov chain with transition matrix $[\theta]$).
3. In the case of an open network, a customer arrives at the network with rate μ_0, and the squared coefficient of variation is C_0. A customer joins the mth station with probability θ_{0m} and leaves the system from the m'th

station with probability $\theta_{m',M+1}$ (for notational convenience, the source is treated as station 0 and the sink as station $M+1$).

If all service time distributions are exponential, then Jackson's theorem [6] applies, which states that the joint queue size distribution $p(n_1, n_2, \cdots, n_M)$ is the product of the M marginal distributions $p_m(n_m)$:

$$p(n_1, n_2, \cdots, n_M) = \prod_{m=1}^{M} p_m(n_m). \tag{9}$$

In his recent paper [5], Kobayashi proposed that queuing processes of a general queuing network be approximated by a vector-valued diffusion process. The interactions among different queuing processes are explicitly considered in the diffusion equation in terms of the variance-covariance matrix. He derived the joint queue size distribution, which is expressed in a product form of the marginal queue size distributions. This solution form suggests that we may treat each server independently, provided that the interactions among different server queues are appropriately taken into account. In this section we develop a computationally simpler (but perhaps less exact) approach than the method discussed in [5]: an approximate solution to the marginal queue size distribution is computed by applying Eq. (7) individually to each server and then deriving the joint queue size distribution.

• *Departure and arrival processes*

To apply formula (7) to each station $m \varepsilon [1, M]$, we have to know the rate of the arrival process $\lambda^{(m)}$ and the squared coefficient of variation $C_a^{(m)}$ of the interarrival time, as well as the processing rate $\mu^{(m)}$ and the squared coefficient of variation for the processing time $C_s^{(m)}$. Clearly

$$\mu^{(m)} = \mu_m \quad C_s^{(m)} = C_m. \tag{10}$$

Subsequently we want to determine $\lambda^{(m)}$ and $C_a^{(m)}$. We first concentrate on the departure process of station m and then turn our attention to the arrival process, which is the sum of M ($M-1$ for closed networks) departure processes weighted with the routing probabilities.

Departure process of station m

During busy periods, the rate of the departure process is μ_m, and its squared coefficient of variation is C_m. But the server is busy with probability u_m only (u_m is server utilization). Accordingly, the mean and variance have to be weighted with u_m: thus

$$\frac{1}{\Delta} E[\Delta D_m] = \text{departure rate} = u_m \mu_m, \tag{11}$$

$$\frac{1}{\Delta} \text{var}[\Delta D_m] = \text{variance of the number of departures per unit time} \cong u_m C_m \mu_m, \tag{12}$$

where ΔD_m denotes the increase in the cumulative number of departures during the interval $(t, t+\Delta)$.

Arrival process at station m

The arrival process is the superposition of the departure process of those servers m' that have nonzero routing probabilities $\theta_{m'm}$. Therefore, the arrival rate is

$$\frac{1}{\Delta} E[\Delta A_m] = \sum_{m'=0 \text{ (or 1)}}^{M} u_{m'} \mu_m \theta_{m'm}, \tag{13}$$

where ΔA_m is the change in the cumulative number of arrivals during the interval $(t, t+\Delta)$. Note that in the lower index of the summation, $m'=0$ applies for an open network, whereas $m'=1$ is for a closed network. The expression for the variance is complicated by the fact that the randomness of the routing is an additional source of variation. We have

$$\frac{1}{\Delta} \text{var}[\Delta A_m] = \text{variance of the number of arrivals during unit time,}$$

$$= \sum_{m'=0 \text{ (or 1)}}^{M} \text{var}[\Delta D_{m'm}] u_{m'}, \tag{14}$$

where $\Delta D_{m'm}$ is that part of the output stream of station m' that is routed to station m. The expression for $\text{var}[\Delta D_{m'm}]$ (its derivation is in Appendix A of [5]) is

$$\frac{1}{\Delta} \text{var}[\Delta D_{m'm}] = \text{variance of the number of arrivals from } m' \text{ to } m \text{ during unit time}$$

$$= [(C_{m'} - 1)\theta_{m'm} + 1]\theta_{m'm}\mu_{m'}. \tag{15}$$

The rate and the squared coefficient of variation of the interarrival time can now be expressed as

$$\lambda^{(m)} = \sum_{m'} u_{m'} \mu_{m'} \theta_{m'm} \quad \text{and} \tag{16}$$

$$C_a^{(m)} \cong \frac{\text{var}[\Delta A_m]}{E[\Delta A_m]} = \frac{1}{\lambda^{(m)}} \sum_{m'} [(C_{m'} - 1)\theta_{m'm} + 1]\, \mu_{m'} \mu_{m'} \theta_{m'm}, \tag{17}$$

where the approximation used in Eq. (17) is, as is that of Eq. (2), based on the central limit theorem as applied to the number of arrival epochs [1].

• *Open networks*

For open networks in equilibrium, the arrival rate at station m [Eq. (16)] is completely determined by the arrival rate $\mu_0 = \lambda^{(0)}$ and the routing probabilities $\{\theta\}$, namely

$$\lambda^{(m)} = \lambda^{(0)} e_m, \tag{18}$$

where e_m is the average number of vistis to station m by a job during its lifetime in the system. If the Markov chain $[\theta]$ is irreducible, the quantities e are uniquely determined by

$$e_m = \theta_{0m} + \sum_{m'=1}^{M} e_{m'}\theta_{m'm}. \tag{19}$$

With the arrival rate $\lambda^{(m)}$ at server m determined, we obtain for the server utilization

$$u_m = e_m\lambda^{(0)}\mu_m^{-1}. \tag{20}$$

(Note that this result holds exactly independently of the forms of interarrival times and service times.) By means of Eq. (19) and Eq. (20), the expression for $C_a^{(m)}$ can be simplified to

$$C_a^{(m)} \cong 1 + \sum_{m'=0}^{M} (C_{m'} - 1)\theta_{m'm}^{\ 2} e_{m'}e_m^{-1}. \tag{21}$$

Now we apply Eq. (21) to station m, giving the following expression for the queue size distribution for this station

$$\hat{p}_m(n) = \begin{cases} 1 - u_m & \text{if } n = 0, \\ u_m(1 - \hat{\rho}_m)\hat{\rho}_m^{\,n-1} & \text{if } n \geq 1, \end{cases} \tag{22}$$

where

$$\hat{\rho}_m = \exp\left(-\frac{2(\mu_m - \lambda^{(m)})}{C_a^{(m)}\lambda^{(m)} + C_m\mu_m}\right). \tag{23}$$

• *Closed networks*

Two basic problems exist in the analysis of closed networks:

1. The server utilization can no longer be simply determined via Eq. (20).
2. The distribution is over a finite population N.

Closely related to problem 1 is the fact that the parameters e are no longer uniquely determined, since the system of linear equations

$$e_m = \sum_{m'=1}^{M} e_{m'}\theta_{m'm} \tag{24}$$

has clearly nonunique solutions even if the matrix $[\theta]$ is irreducible. If a set $\{e\}$ forms a solution, so does $\{\gamma e\}$ for any scalar constant γ. At best we can say that

$$u_m = \gamma e_m\mu_m^{-1}. \tag{25}$$

There is no simple way of determining the constant γ. If we assume that at least one of the parameters e_m/μ_m is larger than the others, then the server with this service rate is the bottleneck of the system. In such a system, the utilization of the bottleneck server goes to 1 as $N \to \infty$. For a closed system with bottleneck server k and sufficiently large population N, utilization of the server is well approximated by

$$u_m = e_k^{-1}\mu_k e_m\mu_m^{-1}, \text{ where } m = 1, 2, \cdots, M. \tag{26}$$

A different approach to obtaining approximate values for the utilizations is to assume an exponential server network with the same $\{\mu\}$ and $\{\theta\}$. The problems associated with the estimations of the utilizations are discussed further in the examples of section 5. It should be noted here that Gaver and Shedler discuss in a recent paper [7] a different way of "fitting" the diffusion approximation.

To deal with the finite population N, we make the assumption that for $n \leq N$ the marginal distributions differ from the limiting case $N = \infty$ only by a proportionality constant.

The suggested treatment of closed networks may now be summarized as follows:

Step 1 Estimate the server utilization $\tilde{u}_m$, $m \in [1, M]$, and use these estimates to compute $\lambda^{(m)}$ according to Eq. (16) and $C_a^{(m)}$ according to Eq. (17).

Step 2 Compute the improper distributions for $n \in [0, N]$:

$$\tilde{p}_m(n) = \begin{cases} 1 - \tilde{u}_m & \text{for } n = 0, \\ \tilde{u}_m(1 - \hat{\rho}_m)\hat{\rho}_m^{\,n-1} & \text{for } n \geq 1. \end{cases} \tag{27}$$

Step 3 Compute the approximative joint distribution according to

$$\hat{p}(n_1, n_2, \cdots, n_M) = \pi\tilde{p}_M\left(N - \sum_{m=1}^{M-1} n_m\right)\prod_{\ell=1}^{M-1}\tilde{p}(n_\ell), \tag{28}$$

where π is a normalization constant to be chosen so that Eq. (28) is a proper distribution.

Before we close this section, it is worthwhile to give an interesting interpretation of the formalism introduced for the treatment of networks (open or closed). It is easy to verify that the assumption of independent marginal distributions of the form of Eq. (27) is equivalent to the solution of an exponential server network with processing speed dependent on the local queue size n_m. Such networks have been analyzed by Jackson [6]. The speed of server m is then found as a function of the local queue size n as follows

$$\mu_m(n) = \begin{cases} \mu_m(1 - u_m)(1 - \hat{\rho}_m)^{-1} & \text{if } n = 0, \\ \mu_m u_m\hat{\rho}_m^{-1} & \text{if } n \geq 0. \end{cases} \tag{29}$$

Thus the diffusion approximation has led us to a *replacement of the original network by an exponential server network with suitably chosen queue-dependent processing rates*. These effective processing rates are determined by the diffusion approximation and depend not only on the processing rates $\{\mu\}$ but also on the routing $\{\theta\}$ and the variance of the service time distributions expressed by $\{C\}$.

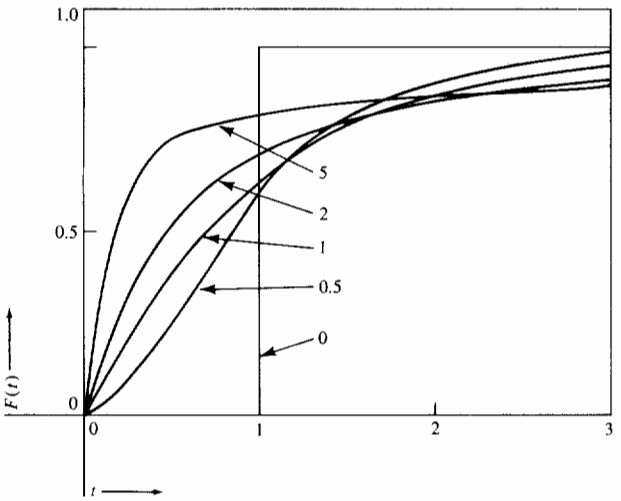

Figure 1 Cumulative service time distributions used in the examples given subsequently. The parameter represents C_s, the squared coefficient of variation.

Table 1 Summary of parameters used in sample distributions depicted in Fig. 1. Parameter C_s = the squared coefficient of variation.

No.	*Service time distribution*	C_s
1	Hyperexponential	5
2	Hyperexponential	2
3	Exponential	1
4	Erlang $m = 2$	0.5
5	Constant (Erlang with $m = \infty$)	0

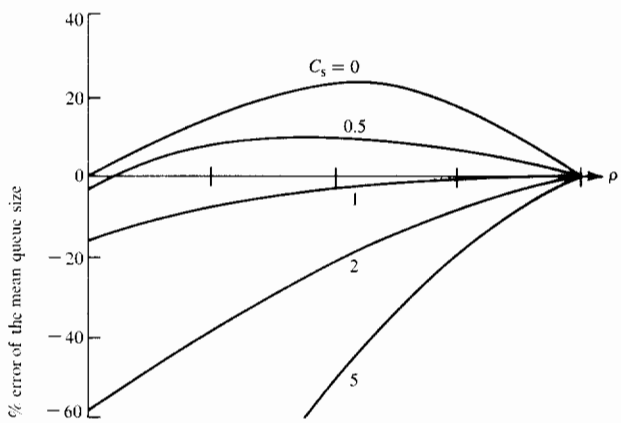

Figure 2 Relative error of the mean queue size ε_r vs the server utilization ρ in the M/G/1 system, $\varepsilon_r = (E[n] - E[\hat{n}])/E[n]$, where the parameter is C_s, the squared coefficient of variation of service time.

4. Accuracy of the diffusion approximation for the single server queue

In this section we first discuss the errors in the mean and variance of queue size and subsequently give comparisons of the approximate solution with some known analytical results. The assumed distribution functions for the service times are the Erlang distribution as a model for service times more regular than exponential (i.e., $C < 1$) and the hyperexponential distribution for service times with $C > 1$.

These distributions are defined in Appendix A. Examples that are used consistently throughout the following sections are depicted in Fig. 1, and their parameters are summarized in Table 1.

- *M/G/1 queue*

Error in mean queue size

The mean queue size of the M/G/1 queue is well known [8] as

$$E[n] = \rho + \frac{\rho^2}{2}\left(\frac{1 + C_s}{1 - \rho}\right), \tag{30}$$

where $\rho = \lambda/\mu$ is the server utilization and $C_s = \text{var}[\tau_s]/E[\tau_s]^2$ is the squared coefficient of variation of the service time distribution. The mean queue size $E[\hat{n}]$ obtained by the approximation (7) is

$$E[\hat{n}] = \rho/1 - \hat{\rho}. \tag{31}$$

A plot of the relative error ε_r of the mean queue size is shown in Fig. 2. We make the following observations:

- The relative error of the mean vanishes as $\rho \to 1$.
- The mean queue size $E[\hat{n}]$ of the diffusion approximation tends to be an underestimate for cases with $C_s < 1$ and an overestimate with $C_s > 1$.
- Although the relative error of the mean queue size may sometimes be quite large, the absolute error is always small. In fact it is not difficult to show that the absolute error $|\varepsilon_a|$ is bounded by $|\varepsilon_a| \leq |1 - C_s|/2$ for all $C_s \geq 0$ and for all values of utilization factor ρ. The upper bound is achieved when $\rho = 1$. The above inequality does not hold only for the interval $C_s \in [0.08, 1.06]$, for which it can be shown that $|\varepsilon_a| \leq 0.08$.
- Highest positive values of ε_r are found for $C_s = 0$ and $\rho \approx 0.5$, whereas for $C_s > 1$ the maximum of $|\varepsilon_r|$ is found at $\rho = 0$ and increases with C_s.

- *Error in the variance of queue size*

The variance of the M/G/1 queue size distribution [8] is

$$\text{var}[n] = \frac{\rho^3(1 + 3C_s - D_s)}{3(1 - \rho)} + \frac{3\rho^2(1 + C_s)}{2(1 - \rho)} + \rho - E[n]^2, \tag{32}$$

Table 2 Relative error (in %) of variance of the queue size $\varepsilon_r' = (\mathrm{var}[n] - \mathrm{var}[\hat{n}])/\mathrm{var}[n]$, where ρ is the server utilization and C_s the squared coefficient of variation of service time.

ρ	C_s = 5	2	1	0.5	0
0.4	−195	−71	−11	23	51
0.6	−85	−35	−4	19	50
0.8	−30	−12	−0.8	10	28
0.9	−13	−5	−0.2	4	14

Table 3 Relative error of the asymptotic decrement: $(r - \hat{\rho}/r)$

ρ	C_s = 5	2	1	0.5	0
0.4	−17	−15	−6	15	75
0.6	−5	−4	−1	6	32
0.8	−1	−0.7	0.01	1.3	7
0.9	−0.2	−0.2	0.001	0.3	1.5

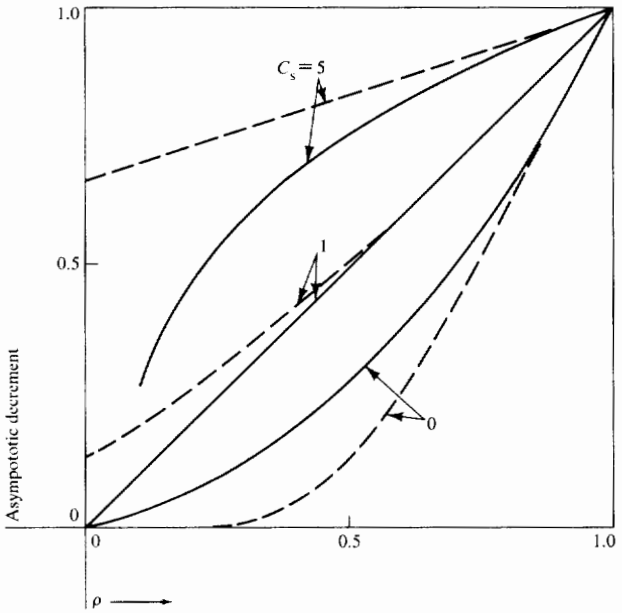

Figure 3 The asymptotic decrement r and its approximation $\hat{\rho}$ vs the server utilization ρ in the M/G/1 system. The solid lines are exact values (r), and the dashed ones are by the diffusion approximation ($\hat{\rho}$). The parameters 0, 1, and 5 represent C_s.

where $D_s = E[(\tau_s - \lambda)^3]/E[\tau_s]^3$ and as before $C_s = \mathrm{var}[\tau_s]/E[\tau_s]^2$. The variance of queue size obtained from the diffusion approximation solution is

$$\mathrm{var}[\hat{n}] = \rho(1 - \Delta\rho)/(1 - \hat{\rho})^2, \tag{33}$$

with $\Delta\rho = \rho - \hat{\rho}$. The error of $\mathrm{var}[\hat{n}]$ follows a pattern similar to the error of the mean discussed above. Some values of the relative error of variance, ε_r', for various utilizations ρ are summarized in Table 2. The magnitude of the relative error $|\varepsilon_r'|$ is found to be consistently higher than corresponding values $|\varepsilon_r|$ for the mean.

Asymptotic slope of the queue size distribution

For a wide class of holding time distributions, the resulting queue size distribution has an exponential tail, i.e., $p_{n+1}/p_n \to r$ as $n \to \infty$. In the argument of section 2, the quantity $\hat{\rho}$ was introduced as an approximation of r. For the distributions of Table 1, the exact values of r have been obtained as a function of ρ. The graphs of r and $\hat{\rho}$ vs ρ are shown in Fig. 3. The relative errors of $\hat{\rho}$ are summarized in Table 3. We find that

- The quantity $\hat{\rho}$ tends to be an overestimate of r for $C_s > 1$ and an underestimate of r for $C_s < 1$.
- Relative and absolute errors vanish as $\rho \to 1$.
- For $C_s < 1$ the error is largest for $C_s = 0$, and for $C_s > 1$ the error increases with C_s.

Comparison of queue size distributions

Analytic expressions for the queue size distributions $p(n)$ have been obtained for the distributions of Table 1. In the cases of the 2-stage Erlang and the 2-stage hyperexponential distributions for the holding times, the exact queue size distribution is of the form

$$p(n) = \alpha_1 r_1^n + \alpha_2 r_2^n, \tag{34}$$

where r_1 and r_2 are the roots of a second-order polynomial (see Appendix B), and α_1 and α_2 are chosen such that $p(0) = 1 - \rho$ and $\Sigma p(n) = 1$.

The observations reported above imply that $\hat{\rho}$ is a good approximation for the larger of the roots, r_1, which determines the asymptotic behavior. Here we are interested in how well the diffusion approximation of the queue size distribution, $\hat{p}(n)$ of Eq. (7), fits the analytic form for $p(n)$. First we note that $\hat{p}(n)$ has only one geometric term, whereas $p(n)$ for m-stage Erlang and the m-stage hyperexponential services has in general m geometric terms. Thus, the question is how fast the dominant term takes over. Inspection of the graphs in Fig. 4 shows that this is usually the case for $n \geq 3$. The examples of Fig. 4 show quantitatively what has been discussed above, especially that the error gets worse for small ρ and extreme deviations from the exponential distribution (e.g., $C_s = 0$ and $C_s \geq 5$).

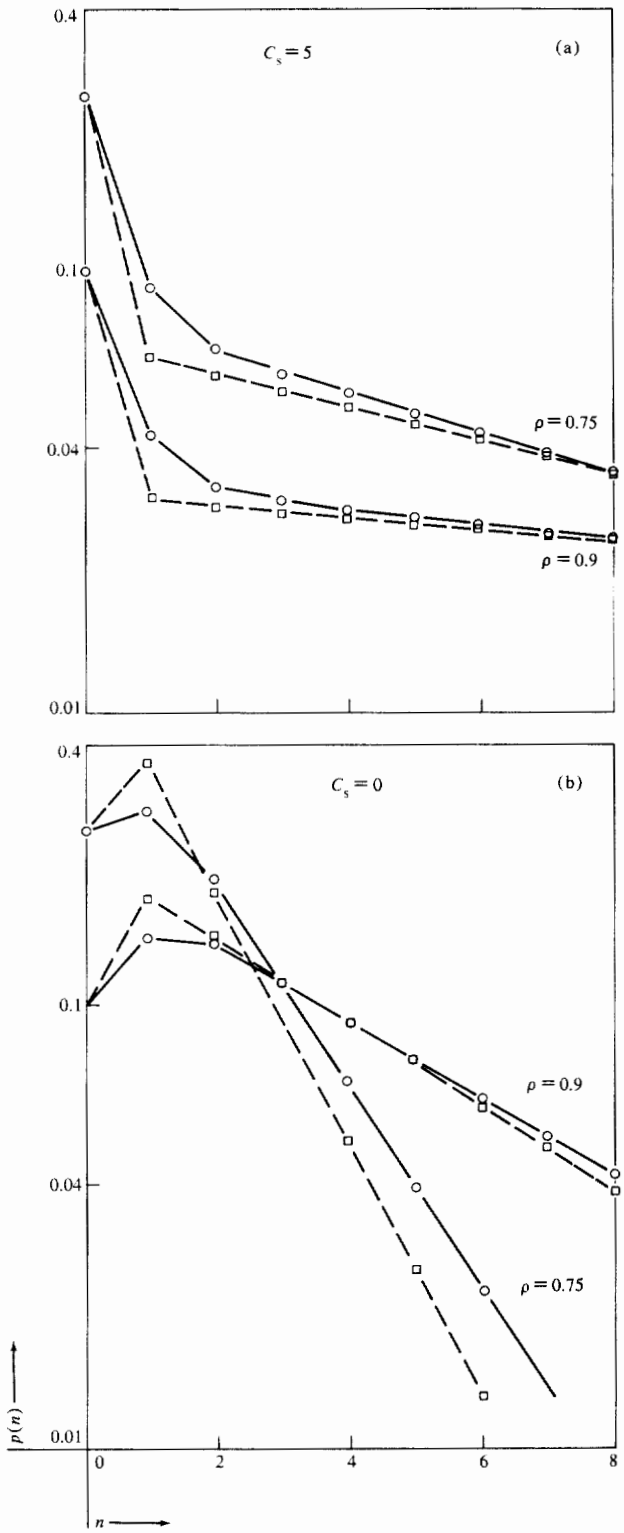

Figure 4 (a) The queue size distribution (in logarithmic scale) for the M/G/1 queue (a), where the service time is hyperexponential with $C_s = 5$. The solid lines are exact values, and the dashed ones are by the diffusion approximation. (b) The queue size distribution (in logarithmic scale) for the M/D/1, viz the service time is constant, i.e., $C_s = 0$.

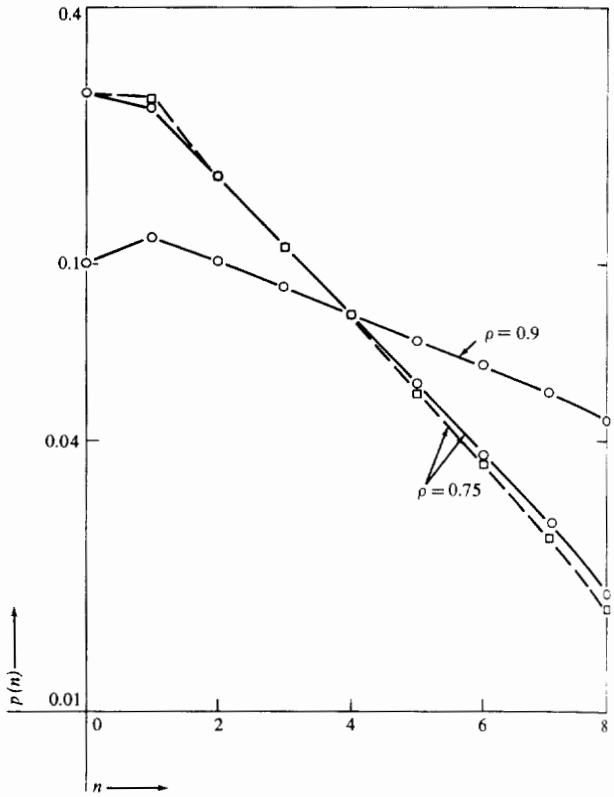

Figure 5 The queue size distribution (in logarithmic scale) for the $E_2/M/1$, i.e., $C_a = 0.5$. For $\rho = 0.9$ the diffusion approximation solution (dashed lines) and the exact solution (solid lines) are indistinguishable.

It may be interesting to note the principal difference between the cases $C_s < 1$ and $C_s > 1$. Compared to exponential service, more regular service with $C_s < 1$ favors the state $n = 1$ (one customer being serviced), and the tail of the queue distribution falls off more rapidly. The opposite is true for service with $C_s > 1$; here $n = 1$ is considerably less probable than in the case of random service, but the tail falls off less rapidly.

• *$E_2/M/1$ queue*

Even for the single server queue, the solution for non-Poisson input is difficult. However, a relatively simple solution is available for the case of Erlang distributed interarrival times (see Appendix C). A closed form solution for the queue size distribution exists for 2-stage Erlang input and exponentially distributed holding times (for more general cases polynomial equations have to be solved). This solution is of the form

$$p(n) = \begin{cases} 1 - \rho & \text{for } n = 0 \\ \rho(1 - r)r^{n-1} & \text{for } n \geq 1, \end{cases} \tag{35}$$

Table 4 Asymptotic decrement r vs server utilization ρ in $E_2/M/1$ and $M/E_2/1$ systems. Relative error is defined by $(r-\hat{\rho})/r$

	$E_2/M/1$		$M/E_2/1$	
ρ	r	error (%)	r	error (%)
0.4	0.312	15	0.275	−33
0.6	0.515	6	0.496	−9
0.8	0.745	1.4	0.740	−1.5
0.9	0.870	0.3	0.868	−0.33

Table 5 Exact asymptotic decrement r compared to $\hat{\rho}$ of the diffusion approximation such that $\ln \hat{\rho} = -2(1-\rho)/(C_s + \rho C_a)$ = constant for each row. (The actual value of ρ is adjusted and varies from column to column.)

	$E_2/M/1$	$M/G/1$		
	$C_a = 0.5$	$(C_a = 1)$		
$\hat{\rho}$	$C_s = 1$	$C_s = 2$	$C_s = 0.5$	$C_s = 0$
0.4	0.414	0.376	0.351	0.326
0.6	0.609	0.596	0.582	0.569
0.8	0.803	0.800	0.796	0.793
0.9	0.901	0.900	0.899	0.898

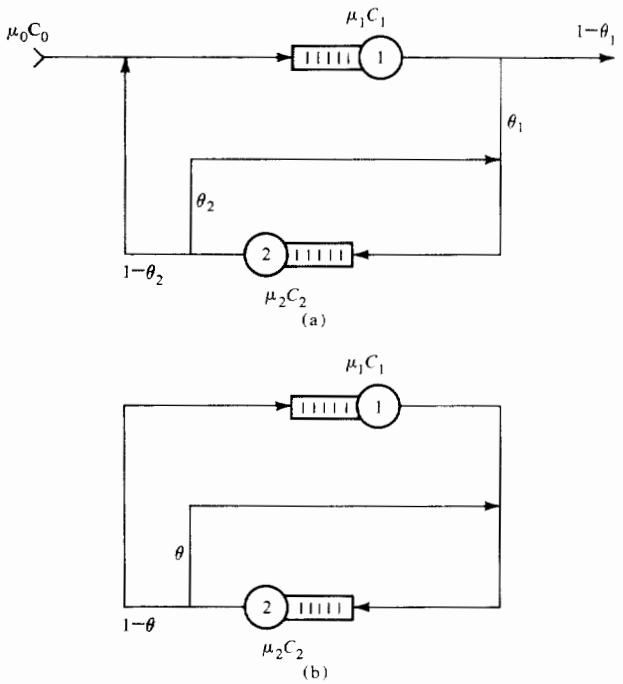

Figure 6 An open server network (a) and a closed network (b).

where r is obtained as the larger root of a quadratic equation (see Appendix C). Equation (35) is analogous to the diffusion approximation solution of Eq. (7). A graph of $p(n)$ vs n is shown in Fig. 5. Values of r and the relative error of $\hat{\rho}$ compared to r are summarized in Table 4 for both the $E_2/M/1$ and the $M/E_2/1$ queues. In general we make findings similar to the case of the $M/G/1$ queue.

It is interesting to observe that

- The effect of Erlangian input on the queue size distribution is similar to that of Erlang distributed holding times. The asymptotic decrements of the two systems converge to the common value as $\rho \to 1$.
- The errors in the mean and asymptotic decrement obtained by the diffusion approximation for the $E_2/M/1$ and the $M/E_2/1$ queues have different signs.

Before we close the section on the single server queue, we want to answer the question of whether the invariance of the diffusion approximation (Eq. 7) with respect to changes in C_a and C_s such that $C_s + \rho C_a$ = constant is supported by the analytical results. For this purpose, Table 5 gives the asymptotic decrement r for various systems such that $2|\beta|/\alpha = \ln \hat{\rho}$ = constant for each row (note that the utilization ρ itself is different for each column). We find from the data of Table 5 that indeed the exact solution is quite similar if the quantity $2|\beta|/\alpha$ is kept constant.

5. Some examples of networks

In this section we consider the two server networks of Fig. 6. Such networks may be useful for various applications. For example server 1 may represent the CPU and server 2 a swapping device. An interactive computer system that may be modeled by this queuing system is described in [9], where it is pointed out that the CPU time distribution has a long tail, i.e., $C_s \gg 1$ (hyperexponential service time distribution). Alternatively, server 1 may be the paging device and server 2 the CPU, with each task requiring a (random) number of time slices.

Subsequently, we compare some analytical results, as well as some simulation results, with the diffusion approximation to investigate the validity of the approach described in section 3, i.e., separate treatment of each server.

- *Open network*

Let us consider the open network of Fig. 6(a). For con-

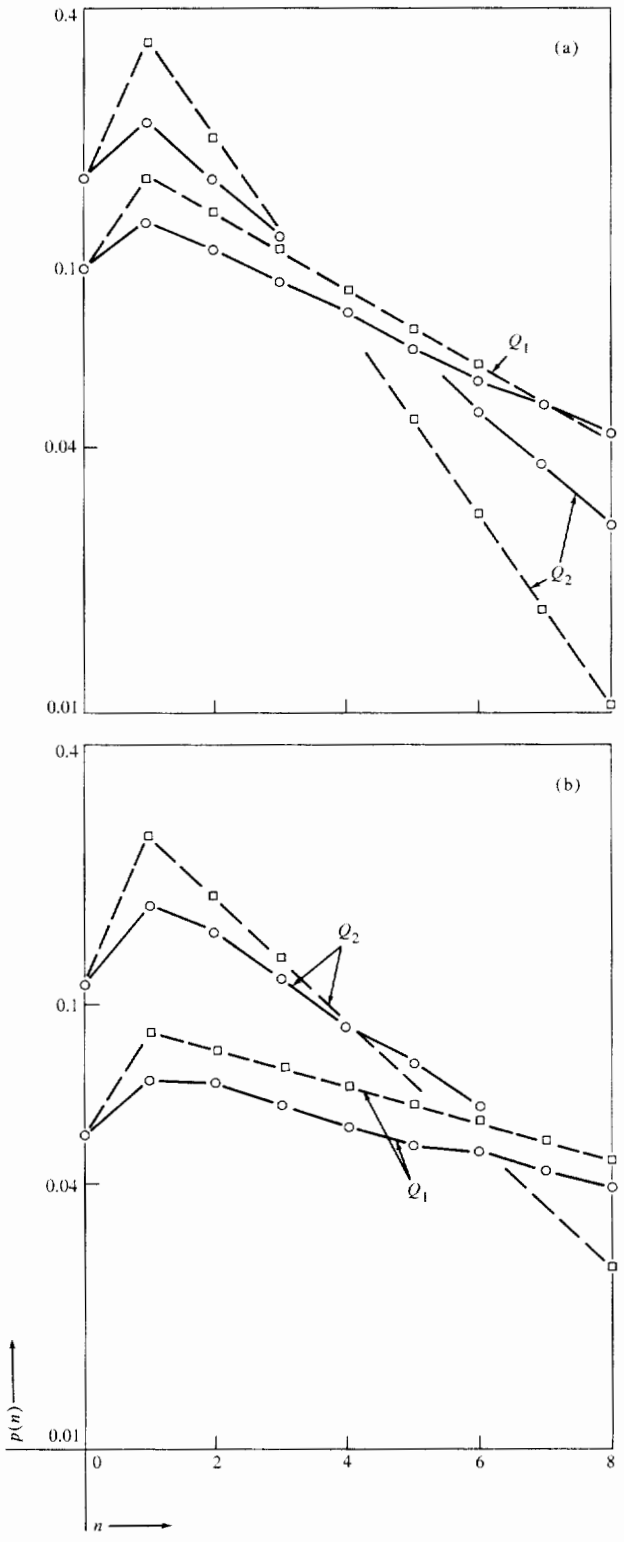

Figure 7 Queue size distributions at servers 1 and 2 in the open network of Fig. 6(a) when the system parameters are those of Case A of Table 6(a). Solid lines are simulation results; dashed lines, diffusion approximation solutions. Queue size distribution for the open network of Fig. 6(a), when the system parameters are those of Case B of Table 6(b).

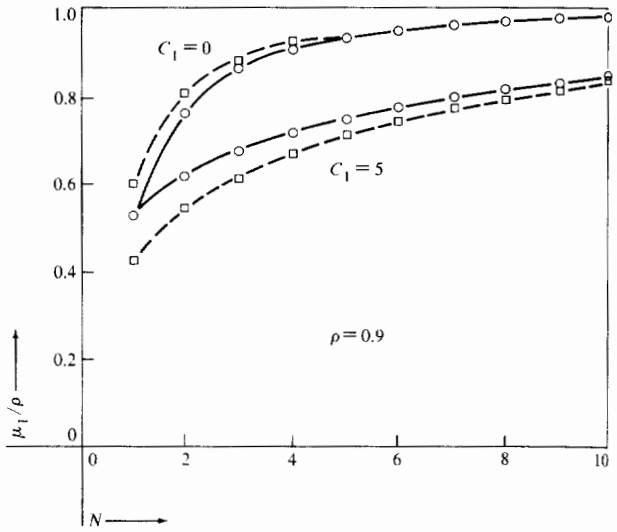

Figure 8 Utilization of server 1 (normalized by ρ) vs the number of jobs N in the closed network of Fig. 6(b). Server 2 is exponential, i.e., $C_2 = 1$ and $\rho = \mu_2(1-\theta)/\mu_1$. Solid lines are exact solutions; dashed lines, diffusion approximation solutions.

Table 6 Average queue size for the open network of Fig. 6(a).

	Case A $\mu_1 = 0.9$,	$\mu_2 = 0.84$	*Case B* $\mu_1 = 0.95$,	$\mu_2 = 0.89$
	queue 1	*queue 2*	*queue 1*	*queue 2*
simulation	6.84	3.22	13.3	4.5
diffusion approx.	6.76	2.70	14.3	4.52
error	1.5%	15%	6%	1%
expon. servers	9	5.25	19	9
error	30%	65%	30%	100%

venience we assume that $\mu_0 = 1$ and $C_0 = 1$ (i.e., Poisson input). We then find

$$\lambda^{(1)} = e_1 = 1/(1-\theta_1), \tag{35}$$

$$\lambda^{(2)} = e_2 = \theta_1/[(1-\theta_1)(1-\theta_2)], \tag{36}$$

$$C_a^{(1)} = 1 + (C_2 - 1)(1-\theta_2)\theta_1, \tag{37}$$

$$C_a^{(2)} = 1 + (C_1 - 1)\theta_1(1-\theta_2) + (C_2 - 1)\theta_2^2. \tag{38}$$

We do not know of any analytic solution for this network. Consequently we compare the diffusion approximation with a few simulation results. The number of examples is restricted because simulation is costly and time-consuming. The following set of parameters was chosen:

Routing: $\theta_1 = 0.5$, $\theta_2 = 0.1$;
Server 1: 2-stage Erlang, i.e., $C_1 = 0.5$;
Server 2: Constant Service, i.e., $C_2 = 0$.

Results for two different sets of values for μ_1 and μ_2 are given in Table 6 and in Fig. 7, leading to the following observations:

- Mean queue sizes (Table 6) agree well. For comparison, results with an equivalent network having exponential servers are also given.
- Comparison of the queue size distributions shows good agreement for the highly utilized server. The fit is less satisfactory for the second server, which has constant service time distribution. Accuracy gets consistently better as the utilizations of both servers are increased (e.g., Fig. 7(a) vs Fig. 7(b)).

- *Closed network*

We now turn our attention to the closed network of Fig. 6(b). The formulas for arrival rate and squared coefficient of variation, Eq. (16) and Eq. (17), can be made more specific, i.e.,

$$\lambda^{(1)} = u_2\mu_2\ (1-\theta) \tag{40}$$

$$\lambda^{(2)} = u_1\mu_1 + (C_2 - 1)(1-\theta) \tag{41}$$

$$C_a^{(1)} = 1 + (C_2 - 1)(1-\theta) \tag{42}$$

$$C_a^{(2)} = 1 + (C_1 - 1)(1-\theta) + (C_2 - 1)\theta. \tag{43}$$

As already mentioned, the difficulty in applying the diffusion approximation to closed networks is the estimation of the server utilizations, which have to be known a priori to evaluate these formulas. One possibility is to take as estimates the values obtained by an exponential server network with identical routing and the same processing rates. For example, we have

$$\tilde{u}_1 = \rho(1-\rho^N)/(1-\rho^{N+1}); \tag{44}$$

$$\tilde{u}_2 = (1-\rho^N)/(1-\rho^{N+1}), \tag{45}$$

where

$$\rho = \mu_2(1-\theta)/\mu_1, \tag{46}$$

and N is the number of customers. In the sequel, we assume that server 2 is the bottleneck server (i.e., $\rho < 1$). Then, apparently, $\tilde{u}_1 \to \rho$ and $\tilde{u}_2 \to 1$ as $N \to \infty$.

We first consider the example in which the second server has exponentially distributed holding times. An analytical solution for this case is known (see Appendix D). This solution has the property that the queue size distribution is invariant to changes in θ and μ_2 such that $(1-\theta)\mu_2 =$ constant. The diffusion approximation as computed with the above formulas does not show this

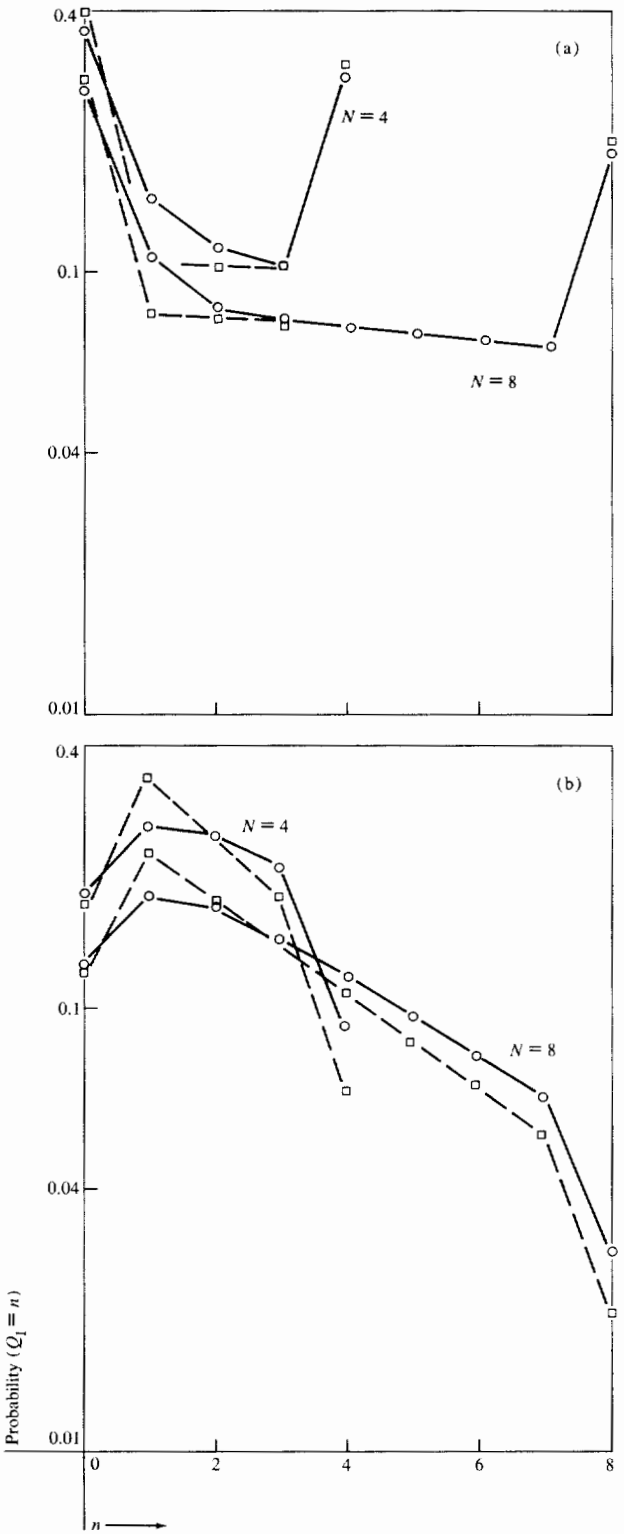

Figure 9 Queue size distribution of server 1 for the closed network of Fig. 6(b) with $\theta = 0$ and $\rho = 0.9$(a). Server 2 is exponential, i.e., $C_2 = 1$, and server 1 is hyperexponential with $C_1 = 5$. Solid lines are exact solutions; dashed lines, diffusion approximations. Queue size distribution of server 1 for the closed network of Fig. 6(b) with $\theta = 0$ and $\rho = 0.9$(b). Server 2 is exponential, i.e., $C_2 = 1$, and server 1 is regular, i.e., $C_1 = 0$.

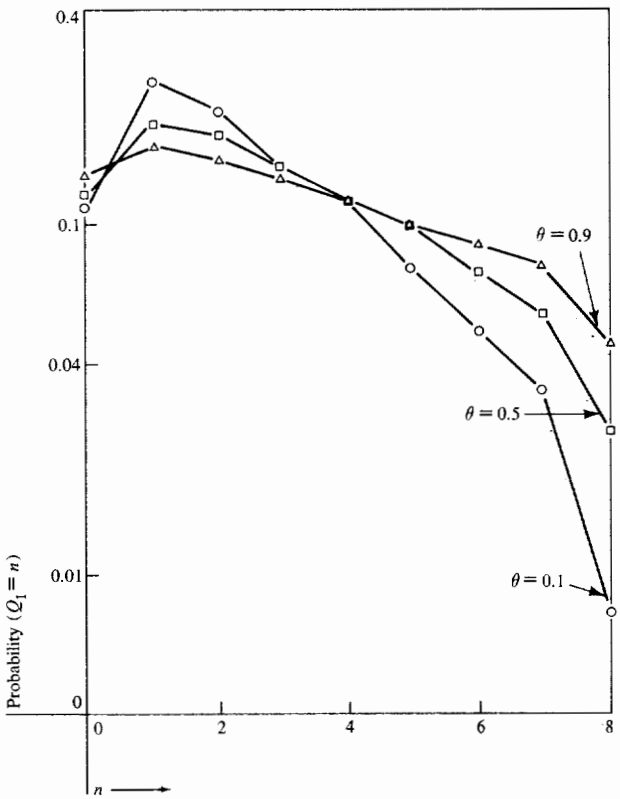

Figure 10 Queue size distribution of server 1 for the closed network of Fig. 6(b) obtained by simulation for different values of θ. The system parameters are $C_1 = 0.5$, $C_2 = 0$, $\rho = 0.9$, and $N = 8$.

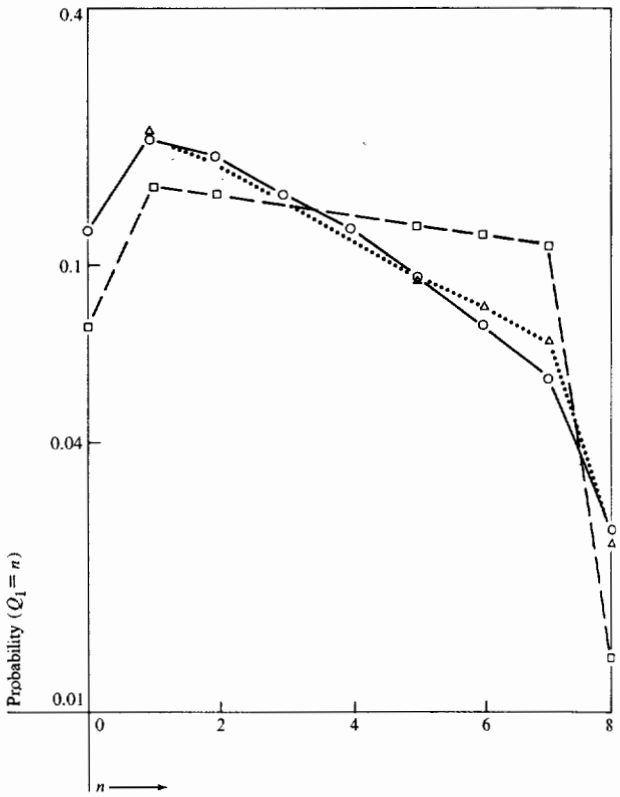

Figure 11 Queue size distribution of server 1 for the closed network of Fig. 6(b). The system parameters are $C_1 = 0.5$, $C_2 = 0$, $\rho = 0.9$, $N = 8$, and $\theta = 0.5$. Solid line is the simulation result; dashed line is the diffusion approximation with $\tilde{u}_1$ and $\tilde{u}_2$ according to Eqs. (43) and (44); dotted line is the diffusion approximation with $\tilde{u}_1 = 0.9$ and $\tilde{u}_2 = 0.99$.

Table 7 Relative error (in %) of the mean queue size and the utilization of server 1. The holding time distributions of server 1 are those of Table 1 and Fig. 1. Server 2 has exponential holding times. The estimated utilizations are obtained by the exponential server network. Restuls are for the diffusion approximation (a) and the exponential server approximation (b).

(a)

	N					
	mean			*utilization*		
C_1	*1*	*4*	*8*	*1*	*4*	*8*
5	20	0.75	−2.5	20	6	3
1	0.3	−0.025	−0.05	0.35	0.1	0.02
0	−15	7.5	8	−15	−1.6	−0.5

(b)

	N					
	mean			*utilization*		
C_1	*1*	*4*	*8*	*1*	*4*	*8*
5	50	16	6	50	16	2.5
1	0.3	−0.025	−0.05	0.35	0.1	0.02
0	−30	12	15	−30	2.5	2

invariance. Generally, we find the best results for $\theta = 0$ (i.e., cyclic tandem queues) and observe an increase in the error as θ approaches one.

We now consider the results given in Fig. 8, Fig. 9, and Table 7:

- The utilization of server 1 as a function of the number of customers is plotted in Fig. 8. As mentioned above, $u_1 \to \rho$ as $N \to \infty$. The approach to the limiting value is fastest for $C_1 = 0$ and slows with increasing values of C_1. In addition the more balanced the system (i.e., with ρ close to 1), the slower this convergence.
- The most accurate results are found for ρ close to 1 (heavy traffic), C_1 close to 1 (exponential service), θ close to 0 (no feedback), and N large. The errors increase as these quantities deviate from the ideal values.
- The error in the mean changes its sign as N increases. For sufficiently large N, the mean queue size of the diffusion approximation is lower than the true value for $C_1 < 1$ and higher for $C_1 > 1$.

- Exact distributions and diffusion approximations are shown in Fig. 9. An excellent fit is found for $\theta = 0$ and $\rho = 0.9$. It is quite remarkable how well the special behavior of $p(N)$ is reproduced in the approximate solution. However, as mentioned above, the fit gets worse as $\theta \to 1$ and the special feature of $p(N)$ tends to disappear.
- The accuracy of the diffusion approximation increases as the exact values of the utilizations are substituted for the approximations found by the exponential server network.

As the last example, we assume nonexponential holding times for both servers; specifically

Server 1: 2-stage Erlangian service times with $C_1 = 0.5$
Server 2: constant service times with $C_2 = 0$.

That portion of the output of the second server routed to the first server has interdeparture times with a geometric distribution that approaches the exponential distribution as $\mu_2 \to \infty$ and $\theta \to 1$ in such a way that $(1 - \theta)\mu_2$ remains constant. Thus for this limit we find again the solution of Appendix D. This is indeed found in the simulation results of Fig. 10. Note that this is an example of how the routing process generates additional variation. The diffusion approximation as depicted in Fig. 11 gives best results for $\theta = 0$. However, the loss of accuracy is more serious than in the previously discussed example with $C_2 = 1$. This is due to a much higher sensitivity of the approximate solution to the error in the imperfectly estimated utilizations $\tilde{u}_1$ and $\tilde{u}_2$. A good fit can be obtained by more accurate values for $\tilde{u}_1$ and $\tilde{u}_2$, as shown by the graphs of Fig. 11. An unsolved problem is how such values may be obtained.

6. Conclusions

The main purpose of this paper is to assess the accuracy of the diffusion approximation to queuing systems. Generally, satisfactory results have been found for reasonably highly utilized systems. Throughout the examples we find that:

1. Accuracy is highest for $C \approx 1$, i.e., the exponential server case. It is of course an important requirement that the approximation, designed to handle general distributions, reproduces the one case for which an exact solution is known. The errors grow as C deviates from one.
2. The errors go to zero as the utilizations approach one. This is a consequence of the use of the central limit theorem leading to the diffusion approximation.
3. In all examples with nonexponential distributions, the diffusion approximation yielded a mean and a variance of the queue sizes that are considerably more realistic than those obtained by an exponential server model. As a rule the mean queue size tends to be low for cases with $C < 1$ and high for cases with $C > 1$.
4. The treatment of open networks, which is based on the assumption that each server may be treated separately, has proved successful. The most accurate values for mean and variance are found for the highly utilized servers. Of course, higher errors have to be tolerated for the less utilized servers in a network.
5. In the case of closed networks with a small number of customers, the estimation of adequately accurate utilizations remains an unsolved problem. In some cases (e.g., when the bottleneck server is exponential) utilizations obtained by an exponential server model work satisfactorily.

Finally it should be mentioned that the computational complexity is small and that a general computer program capable of handling any reasonable number of servers may be easily implemented (e.g., in APL). There is therefore considerable hope that such a program may prove a useful tool for design and analysis of systems.

Acknowledgments

The authors are grateful to J. Griesmer for his support in carrying out the symbolic computations of Appendix B. They are also indebted to G. S. Graham for his careful reading of the manuscript.

Appendix A: Definition and properties of the m-stage Erlang and the m-stage hyperexponential distribution

The m-stage Erlang distribution for a random variable τ is

$$F(t) = \text{prob}\,\{\tau \leq t\} = 1 - e^{-\mu m t} \sum_{k=0}^{m-1} \frac{(\mu m t)^k}{k!}, \qquad \text{(A1)}$$

with

$$E[\tau] = 1/\mu, \qquad \text{(A2)}$$

$$\text{var}\,[\tau] = 1/m\mu^2, \qquad \text{(A3)}$$

$$C = \text{var}\,[\tau]/E[\tau]^2 = 1/m \leq 1. \qquad \text{(A4)}$$

Note that the squared coefficient of variation is less than (or equal) to one. The exponential distribution is the special case $m = 1$, whereas τ becomes completely regular as $m \to \infty$.

The hyperexponential distribution of a random variable τ is

$$F(t) = \text{prob}\,\{\tau \leq t\} = 1 - \sum_{k=1}^{m} \pi_k e^{-\mu_k t} \qquad \text{(A5)}$$

with $\pi_k \geq 0$ for all $k \varepsilon [1, m]$ and $\Sigma \pi_k = 1$.

Mean, variance, and squared coefficient of variation are found to be

$$E[\tau] = \Sigma_k \pi_k / \mu_k = \mu^{-1}, \qquad \text{(A6)}$$

$$\mathrm{var}[\tau] = \mu^{-2} + \Sigma_k \Sigma_{k'} \pi_k \pi_{k'} (\mu_k^{-1} - \mu_{k'}^{-1})^2, \tag{A7}$$

$$C = 1 + \mu^2 \Sigma_k \Sigma_{k'} \pi_k \pi_{k'} (\mu_k^{-1} - \mu_{k'}^{-1})^2 \geq 1. \tag{A8}$$

Appendix B: Queue size distributions of the M/G/1 queuing system

The M/G/1 queue is among the oldest solved queuing problems with nonexponential distributions. One solution method is known as the imbedded Markov chain method and derives the queue size distribution via the generating function $U(z) = \Sigma_n p(n) z^n$, which is given as [8, 10]

$$U(z) = (1 - \rho) \frac{(1 - z)\psi[\lambda(1 - z)]}{\psi[\lambda(1 - z)] - z}, \tag{B1}$$

where $\psi(s)$ is the Laplace-Stieltjes transform of the service time distribution

$$\psi(s) = \int_0^\infty e^{-st} dF(t). \tag{B2}$$

For the m-stage Erlang service-time distribution, we have

$$\psi[\lambda(1 - z)] = \left[\left(\frac{\rho}{m} + 1\right) - \frac{\rho}{m} z\right]^{-m}, \tag{B3}$$

which in the case $m \to \infty$ (i.e., constant service) converges to

$$\psi[\lambda(1 - z)] = \exp[-\rho(1 - z)]. \tag{B4}$$

In the case of the m-stage hyperexponential service time distribution defined by Eq. (A5), $\psi[\lambda(1 - z)]$ becomes

$$\psi[\lambda(1 - z)] = \sum_{k=1}^{m} \frac{\pi_k}{(\rho_k + 1) - \rho_k z}, \tag{B5}$$

where $\rho_k = \lambda/\mu_k$. Generating function $U(z)$ is in both cases (except for constant service) a rational function in z.

To obtain the queue size distribution, the generating function $U(z)$ has to be expanded into a power series. For rational functions, such an expression is most easily obtained from its representation as a partial fraction

$$U(z) = \sum_{k=1}^{m} \frac{\alpha_k}{1 - r_k z} = \sum_{i=0}^{\infty} z^i \sum_{k=1}^{m} \alpha_k r_k^i, \tag{B6}$$

where $r_k = z_k^{-1}$ and $|z_k| > 1$ are roots of the following characteristic equations:

$$1 - z\left[\left(\frac{\rho}{m} + 1\right) - \frac{\rho}{m} z\right]^m = 0 \text{ for Erlang service,} \tag{B7}$$

$$\Sigma_k \pi_k \left[\prod_{i \neq k} (\rho_i + 1 - \rho_i z)\right] - z \prod_i (\rho_i + 1 - \rho_i z) = 0$$

for hyperexponential service. (B8)

Both polynomials are of degree $m + 1$ and have $z_0 = 1$ as a root. Then, by Rouche's theorem, there are exactly m other roots outside the unit circle, i.e., $|z_k| > 1$ for $k \varepsilon [1, m]$. But $(z - 1)$ is a common factor of the numerator and the denominator of $U(z)$, which can be cancelled. Therefore we find for the probability distributions

$$p(n) = \sum_{k=1}^{m} \alpha_k r_k^n, \tag{B9}$$

i.e., a superposition of m different geometric distributions.

Closed form solutions may be easily obtained for 2-stage Erlang and 2-stage hyperexponential service time distributions, because then the polynomials of Eq. (B7) and (B8) are both third order and reduce to second order after cancellation of the common factor $(1 - z)$. Some elementary computations show these quadratic equations to be

$$\rho^2 z^2 - \rho(\rho + 4)z + 4 = 0$$

for 2-stage Erlang service, and (B10)

$$\rho_1 \rho_2 - [(\rho_1 + 1)(\rho_2 + 1) - 1]z + (\rho_1 + \rho_2 - \rho + 1) = 0$$

for 2-stage hyperexponential service. (B11)

The queue size distribution then becomes

$$p(n) = \alpha_1 r_1^n + \alpha_2 r_2^n, \tag{B12}$$

with $r_1 = z_1^{-1}$ and $r_2 = z_2^{-1}$, and

$$\alpha_1 = (r_2 + \rho)(1 - r_1)(r_1 - r_2)^{-1}, \tag{B13}$$

$$\alpha_2 = (r_1 + \rho)(1 - r_2)(r_2 - r_1)^{-1}. \tag{B14}$$

In the case of constant holding times, $U(z)$ is a transcendental function. The queue size distribution is found as the coefficients of the Taylor series around the origin $z = 0$. Repetitive derivation of the generating function is a tedious task. The solution, which has been obtained by the symbolic computation system SCRATCHPAD [11], is

$$p(n) = (1 - \rho) \sum_{j=1}^{n} (-1)^{n-j} \frac{(j\rho)^{n-j-1}(j\rho + n - j)}{(n - j)!} e^{j\rho}. \tag{B15}$$

Appendix C: Queue size distribution of the E_m/G/1 queuing system

The single server queue with m-stage Erlang distributed interarrival times and holding times with general distribution $F(t) = \text{prob}\{\tau \leq t\}$ is equivalent to the following queuing system:

1. Interarrival times are exponentially distributed with mean $1/\lambda$.
2. The server accepts only batches of m customers. Service time distribution for servicing a batch is $F(t)$. If

after service completion the queue size n' is less than m, the server idles until the queue has grown to m.

The solution of the latter system by the method of the imbedded Markov chain is similar to the solution of the M/G/1 system. The generating function for the queue size probabilities $p'(n)$ (the primed quantities are for the equivalent system) is found to be

$$U(z) = \frac{(1 - z^m)\psi[m\lambda(1 - z)] \sum_{j=0}^{m-1} p'(j)z^j}{\psi[m\lambda(1 - z)] - z^m}, \qquad \text{(C1)}$$

where $\psi(s)$ is the Laplace-Stieltjes transform of the service time distribution. The queue size probabilities $p'(j)$, $j\varepsilon[0, m-1]$, are m unknown parameters of $U(z)$. Since $U(z)$ must be analytic in the unit circle (i.e., $|z| \leq 1$), these parameters are determined by equating the numerator of $U(z)$ to zero at all the roots of the denominator that are inside the unit circle. According to Rouche's theorem, there are exactly m such roots, one being $z = 1$. The parameters $p'(j)$, $j\varepsilon[0, m-1]$, are therefore uniquely determined. The queue size $p(n)$ of the original problem is obtained by the relation $n = \left[\frac{n'}{m}\right]$ (where n is the original queue size, n' is the queue size of the equivalent system and $[x]$ is the greatest integer less than x). In terms of the probabilities, this relation becomes

$$p(n) = \sum_{j=n}^{n+m-1} p'(j). \qquad \text{(C2)}$$

A closed-form solution is found for 2-stage Erlang input and exponential service, in which case the polynomial equation is of second order. The solution, which is easily obtained by elementary algebra, is

$$p(n) = \begin{cases} 1 - \rho & \text{for } n = 0, \\ \rho(1 - r)r^{n-1} & \text{for } n \geq 1, \end{cases} \qquad \text{(C3)}$$

where

$$r = 4\rho[1 + (1 + 8\rho)^{\frac{1}{2}}]^{-1}. \qquad \text{(C4)}$$

Appendix D: Analytic solution for a cyclic network with one general and one exponential server

A simple analytic solution for the cyclic network of Fig. 6(b) is available by the equivalency principle due to Kobayashi and Silverman [12]. This principle states that the cyclic tandem system with one exponential server and one server with general distribution of service times is equivalent to the M/G/1 queue with finite waiting room. If the parameters of the cyclic queuing system are

$F(t)$: service time distribution of server 1
μ_1 : rate of server 1
μ_2 : rate of server 2 (exponential)
θ : feedback around server 2
N : number of customers:

then the equivalent M/G/1 queue is given by

$F(t)$: service time distribution
$\mu = \mu_1$: rate of server
$\lambda = (1 - \theta)\mu_2$: arrival rate (Poisson arrivals)
N : size of waiting room.

The analytic solution of the M/G/1 queue with finite waiting room is known (see, for example, Keilson [13]). If $\rho = (1 - \theta)\mu_2/\mu_1$, then the queue size distribution with finite waiting room N, $p_N(n)$, is simply expressed in terms of the queue size distribution $p_\infty(n)$ of the unconstrained M/G/1 queue. The result may be cast into the following principles:

1. The shape of the distribution is not affected for $n < N$: more precisely,

$$p_N(n) = K_N p_\infty(n), \qquad \text{(D1)}$$

where K_N is a proportionality factor.

2. For equilibrium the inflow rate equals the outflow rate, i.e.,

$$\lambda[1 - p_N(N)] = \mu[1 - p_N(0)]. \qquad \text{(D2)}$$

3. $p_N(n)$, $0 \leq n \leq N$, is a probability distribution (i.e., adds up to one).

These three conditions uniquely specify the solution as:

$$p_N(n) = K_N p_\infty(n), \qquad \text{for } 0 \leq n < N, \qquad \text{(D3)}$$

$$p_N(N) = 1 - [1 - K_N(1 - \rho)]/\rho, \qquad \text{(D4)}$$

$$K_N = \left\{1 - \rho\left[1 - \sum_{i=0}^{N-1} p_\infty(i)\right]\right\}^{-1}. \qquad \text{(D5)}$$

References

1. D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*, John Wiley and Sons, Inc., New York 1965, Chapters 2 and 5.
2. D. P. Gaver, "Diffusion Approximations and Models for Certain Congestion Problems," *Jour. Appl. Prob.* **5**, p. 607, (1968).
3. G. F. Newell, *Applications of Queuing Theory*, Chapman and Hall, Ltd., London, 1971, Chapter 6.
4. D. P. Gaver and G. S. Shedler, *Multiprogramming System Performance via Diffusion Approximations*, IBM Research Report RJ-938, 1971, Yorktown Heights, N.Y.
5. H. Kobayashi, "Applications of the Diffusion Approximation to Queuing Networks: Parts I and II," to appear in the April and July, 1974, issues of *Jour. ACM*. Also IBM Research Reports RC 3943, 1972, and RC 4054, 1972, Yorktown Heights, N.Y.
6. J. R. Jackson, "Jobshop-Like Queuing Systems," *Management Science* **10**, p. 131-142, (1973).
7. D. P. Gaver and G. S. Shedler, "Approximate Models for

Processor Utilization in Multiprogrammed Computer Systems," *SIAM J. Computing* **2**, No. 3, p. 183-192 (September 1973).
8. L. Takacs, "A Single-Server Queue with Poisson Input," *Operations Research* **10**, p. 388 (1962).
9. H. A. Anderson and R. Sargent, "The Statistical Evaluation of the Performance of an Experimental APL/360 System" *Statistical Computer Performance Evaluation*, W. Freiberger (ed.), p. 73-98, Academic Press, 1972.
10. D. G. Kendall, "Some Problems in the Theory of Queues", *J. Roy. Stat. Soc.* **B13**, No. 2, p. 151-185 (1951).
11. J. H. Griesmer and R. D. Jenks, "Experience with an Online Symbolic Mathematics System," *Proc. of the ONLINE72 Conf.* **1**, Brunel University, Uxbridge, Middlesex, England, 1972, p. 457-476. (Also available as IBM Research Report RC 3925, 1972, Yorktown Heights, N.Y.)
12. H. Kobayashi and H. F. Silverman, *Some Dispatching Priority Schemes and their Effects on Response Time Distribution*, IBM Research Report RC 3584, 1971, Yorktown Heights, N.Y.
13. J. Keilson, "The Role of Green's Functions in Congestion Theory," *Proceedings of the Symposium on Congestion Theory*, W. L. Smith and W. E. Wilkinson (ed.) University of North Carolina Press, 1965, p. 43-71.

Received July 27, 1973

Dr. Reiser, a member of the staff of the IBM Research Laboratory in Zurich, is on temporary assignment at the IBM Research Division Laboratory, Monterey and Cottle Roads, San Jose, California 95193. Dr. Kobayashi is located at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.