INFORMATION PROCESSING 74 - NORTH-HOLLAND PUBLISHING COMPANY (1974)

THE EFFECTS OF SERVICE TIME DISTRIBUTIONS ON SYSTEM PERFORMANCE

Martin REISER and Hisashi KOBAYASHI

Computer Sciences Department, International Business Machines Corporation
Thomas J. Watson Research Center
Yorktown Heights , New York, USA

A cyclic queuing system is often used to model a multiprogrammed computing system. Most of known results, however, are based on the assumption that service times at CPU and I/O device are both exponentially distributed.

This paper derives a closed form solution for the queue size distribution in such a model with one general server. The derivation is based on Green's function method for the imbedded Markov chain analysis and on the equivalence between the assumed model and an M/G/1 queue with finite capacity. We also derive simple algebraic expressions for the cases of hyperexponential and Erlangian distributions. The numerical evaluation results obtained provide useful information about how critically the system performance is affected by the distributional form of service demands.

1. INTRODUCTION

In performance analysis of a multiprogrammed system, a cyclic queueing model is often used [1-4], in which the central processing unit (CPU) and input-output (I/O) device are treated as two independent servers. The jobs in the model correspond to those programs which are allocated some portion of main memory. The number of such programs is called the degree of multiprogramming, and is treated as a constant, which is a reasonable assumption to make when the system is heavily loaded.

Despite its great simplification, very little is known about the behavior of a cyclic queueing system, except for a special case where service times at both servers are exponentially distributed [3,4]. Applications of these exponential server queueing models to computer system modelling are in prevalent use [5] because of their mathematical tractability, although some of the recent empirical studies indicate that service time distributions in real computer systems are often quite far from exponential [5,6].

In the present paper we discuss the cyclic queueing model of fig. 1 in which one of two servers (i.e., either CPU or I/O device) is allowed to have a service time distribution of general form. An analysis of such a model is given by Gaver [1]. Shedler and Lewis [2] consider the effect of system overhead. These studies, however, are primarily concerned with the server utilization and to the present authors' knowledge, no prior work discusses such quantities as queue size distribution, response time, which are important measures in system performance analysis.

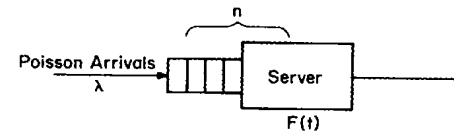
In section 2, we derive a closed form solution for the queue size distribution. The result is based on Green's function method for the imbedded Markov chain analysis discussed by Keilson [7,8] and on the equivalence between the system of fig. 1 and an M/G/1 queue with finite capacity [9].

In section 3, taking as special cases hyperexponential and Erlangian distributions, we obtain the corresponding queue size distributions in the form of finite sums. The solution for a multi-stage Erlangian case is derived by using a symbolic computation system [10].

Section 4 presents various numerical results. Such performance measures as utilization, the average queue size and response time are plotted for different configuration parameters. These results clearly show that most of performance measures are sensitive to the distributional form of service times, and thereby emphasize the relevance of our study.

2. GENERAL SOLUTION

We consider the cyclic queueing system of fig. 1 with (1) A fixed number N of jobs (or customers) circulating in the network. (2) Server 1 having service times $\{\tau\}$ with distribution $F(t)=\text{Prob}\{\tau \le t\}$ and mean $\mu^{-1}=E[\tau]$. (3) Server 2 having exponentially distributed service times $\{\tau\}$, i.e., $\text{Prob}\{\tau \le t\}=1-e^{-\lambda t}$.



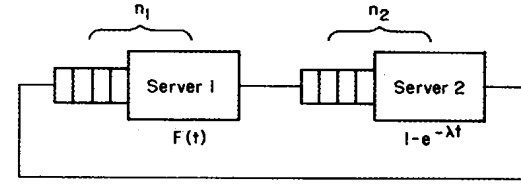
$n \le N$: Finite Capacity

Fig. 1. A cyclic queueing system

Kobayashi and Silverman [9] have shown the following equivalence principle between the cyclic queueing system and in general server with finite capacity: The joint probability distribution of this queueing system of fig. 1 is

$$p(n,N-n) = \text{Prob}\{n \text{ jobs in queue 1 and } N-n \text{ jobs in queue 2}\} = p_N(n)$$

where $p_N(n)$ is the queue size distribution of the M/G/1 queue with finite capacity of size N (including the one being served), service time distribution F(t) and Poisson arrival rate λ as shown in fig. 2.



$n_1 + n_2 = N$

Fig. 2. The M/G/1 queue with finite capacity

The validity of this principle follows from the fact that jobs arrive at queue 1 according to a Poisson process with parameter λ whenever the second queue is not empty. If the second queue is empty then the first queue has length N and the Poisson arrivals are interrupted. This corresponds exactly to the finite capacity of size N of the equivalent M/G/1 queue. Note that this relationship is based on the so-called "memoryless property" of an exponential server and a Poisson process.

The solution of the M/G/1 queueing problem is based on the imbedded Markov chain. If the number of jobs in the waiting room is observed at successive service-initiation times, then the observed sequence X_k is a sample from a Markov chain X, which is said to be imbedded in the queueing process. In case of the finite capacity size N, the chain X assumes the value 0,1,2,...,N-1. If we denote by α_k the number of arrivals during service of customer k, then the sequence $\{X_k\}$ is governed by the recursive law

$$X_{k+1} = \begin{cases} 0 & \text{if } X_k - 1 + \alpha_k < 0 \\ X_k - 1 + \alpha_k & \text{if } 0 \le X_k - 1 + \alpha_k \le N-1 \\ N-1 & \text{if } X_k - 1 + \alpha_k > N-1 \end{cases} \quad (1)$$

The random variable α_k has the probability density

$$a(n) = \text{Prob}\{\alpha_k = n\} = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} dF(t) \quad (2)$$

where λ is the arrival rate. Hence, the generating function for a(n) is

$$A(z) = \sum_0^\infty a(n) z^n = \psi(\lambda - \lambda z) \quad (3)$$

where $\psi(s)$ is the Laplace-Stieltjes transform of the service time distribution, i.e., $\psi(s) = \int_0^\infty e^{-st} dF(t)$.

The Markov chain X of eq. (1) is a random walk on the integers with two reflecting barriers at zero and N-1. Such processes are well investigated, and one known technique uses Green's function method [7,8]. Below, we give a summary statement of the results which are of our immediate interest.

If A(z) has an annulus of convergence which includes all the roots of z-A(z)=0, then the queue size (including the one being served) distribution $p_N(n)$ of the M/G/1 queue with finite capacity is determined by

$$p_N(n) = \begin{cases} K_n \hat{p}(n), & 0 \le n < N \quad (4a) \\ 1-[1-K_N(1-\rho)/\rho], & n = N \quad (4b) \end{cases}$$

where $\rho=\lambda/\mu$, μ^{-1} is the mean of F(t) and

$$\hat{p}(n) = \frac{1}{n!} \frac{d^n}{dz^n} \left[(1-\rho) \frac{(1-z)A(z)}{A(z)-z}\right] \Big|_{z=0} \quad (5)$$

$$K_n = \{1-\rho[1- \sum_{n=0}^{N-1} \hat{p}(n)]\}^{-1} . \quad (6)$$

For $\rho<1$ we recognize the quantities $\hat{p}(n)$ as the queue size probabilities of unrestricted M/G/1 queue. In this case we find the following interesting interpretation of the above result: (1) Except for a multiplicative constant K_N, the boundary at n=N has no influence on the distributional form of $P_n(n)$ for n<N. (2) In the steady state the inflow rate must equal the outflow rate, thus

$$[1-p_n(N)]\lambda = [1-p_N(0)]\mu ; \quad (7)$$

(3) The probabilities sum up to one, $\sum_0^N p_n(n)=1$.

These three principles completely specify the solution as given above. The major difficulty in applying this simple result is to get the quantities $\hat{p}(n)$ explicity. Closed form solutions for several cases of interest are derived in the next section.

The joint distribution p(n,N-n) in the cyclic queueing system of fig. 1 follows directly from the equivalence principle stated earlier. Eq. (7) relates the utilizations of the servers, i.e.,

$$\frac{u_1}{u_2} = \frac{\lambda}{\mu} = \rho \quad (8)$$

where $u_1=1-p(0,N)$ is the utilization of server 1 and $u_2=1-p(N,0)$, that of server 2.

3. THE UNRESTRICTED M/G/1 QUEUE WITH ERLANGIAN AND HYPEREXPONENTIAL SERVICE TIME DISTRIBUTION

The main result in section 2 was expressed in terms of the quantity p(n) which, in case of $\rho<1$, is the queue size distribution for the unrestricted problem. This section is devoted to the derivation of simple closed form solutions for $\hat{p}(n)$ in the case of the m-stage Erlangian distribution

$$F(t) = 1 - e^{-\mu m t} \sum_{k=1}^{m} \frac{(\mu m t)^k}{k!} \quad (9)$$

and of the hyperexponential distribution

$$F(t) = 1 - \sum_{k=1}^{m} \pi_k e^{-\mu_k t} \quad (10)$$

where $\sum \pi_k = 1$ and $\frac{1}{\mu} = \sum_k \frac{\pi_k}{\mu_k}$. The generating function of the arrival probabilities of eq. (3) is

$$A(z) = \{(\frac{\rho}{m} + 1) - \frac{\rho}{m} z\}^{-m} \quad (11)$$

for the case of Erlangian service times and

$$A(z) = \sum_{k=1}^{m} \rho_k \{(\rho_k+1) - \rho_k z\}^{-1} \quad (12)$$

for the case of hyperexponential service times where $\rho_k=\lambda/\mu_k$. The quantities p(n) are defined by (5) or

$$\hat{p}(n) = \frac{1}{n!} \frac{d^n U(z)}{dz^n} \Big|_{z=0} \quad (13)$$

with

$$U(z) = (1-\rho) \frac{(1-z)A(z)}{A(z)-z} \quad (14)$$

which we recognize as the probability generating function of the unrestricted M/G/1 queue. For the distributions considered here, U(z) is a rational function in z. The derivatives in eq. (13) are most easily found through a partial fraction expansion of U(z); this, however, requires knowledge of the roots of the characteristic equation A(z)-z=0. The asymptotic behavior of $\hat{p}(n)$ is determined by the smallest real root z_1 of the characteristic equation z-A(z) = 0 such that $z_1 \neq 1$ ($z_0=1$ is always a root), i.e.,

$$\hat{p}(n) = O(z_1^{-n}) \quad (15)$$

Thus in the case of Erlangian or hyperexponential service times, $\hat{p}(n)$ has always a geometric "tail." In fact this property holds for a much larger class of distributions for which A(z) has an annulus of convergence which is sufficiently large to enclose z_1.

For the two-stage Erlangian or hyperexponential distribution the characteristic equation is of the third order and reduces to the second order after cancellation of the common factor (1-z). We then find for both cases

$$\hat{p}(n) = C_1 z_1^{-n} + C_2 z_2^{-n} \tag{16}$$

where

$$C_1 = (1 - \rho z_2)(1 - z_1)/(z_2 - z_1) \tag{17}$$

$$C_2 = (1 - \rho z_1)(1 - z_2)/(z_1 - z_2) \tag{18}$$

with z_1 and z_2 being the roots of the following characteristic equations

$$\rho^2 z^2 - \rho(\rho+4)z + 4 = 0 \quad \text{for the 2-stage Erlangian} \tag{19}$$

$$\rho_1\rho_2 z^2 - (\rho_1+\rho_2+\rho_1\rho_2)z + 1 + \rho_1 + \rho_2 - \rho = 0 \tag{20}$$

for the 2-stage hyperexponential where the parameter ρ of eq. (20) is given by $\rho=\pi_1\rho_1+\pi_2\rho_2$. The hyperexponential distribution with $m=2$ is sufficiently general to treat most cases of practical interest since it has three free parameters and can produce any coefficient of variation greater than 1, i.e., $c^2>1$. This is not so for the corresponding Erlang case, which is specified up to the mean and for which $c^2=0.5$.

We are therefore interested in the solution for the Erlangian distribution with general m, which we have obtained in the form of a finite sum, i.e.,

$$\hat{p}(n) = (1-\rho) \sum_{j=0}^{n} (-1)^{n-j} r^{n-j-1} \left\{ \binom{jm}{n-1-j} + \binom{jm}{n-j} r \right\} R^{jm}, \quad n \geq 0 \tag{21}$$

with $R=(\rho/m)+1$ and $r=\rho/(\rho+m)$. In the limit $m\to\infty$, which corresponds to constant service times, eq. (21) specializes to

$$\hat{p}(n) = (1-\rho) \sum_{j=0}^{n} (-1)^{n-j} \frac{(j\rho)^{n-j-1}(j\rho+n-j)}{(n-j)!} e^{j\rho} . \tag{22}$$

The above results, which can easily be programmed in a suitable computer language, have been obtained by forming successive derivatives of U(z) and inducing the form of the general term. The computation of the derivatives may become quite tedious if we attempt to do it manually. We have obtained the results by means of symbolic computations available on a digital computer [13].

4. NUMERICAL RESULTS

We now present numerical results to illustrate how the distributional form of service time affects system's performance in terms of server utilization, average queue size, average response time, etc.

First, we show typical examples of the queue size distribution in figs. 3(a) and (b), in which $\rho = \frac{\lambda}{\mu} = 0.75$ is chosen, and the total number of jobs (i.e., the degree of multiprogramming in a computing system model) is set to N=4,6, or 8. Fig. 3(a) shows the case where service times at server 1 are constant, i.e., the coefficient of variation c=0. The queue size distribution is calculated from eqs. (4) and (22). Fig. 3(b) plots the case where the distribution F(t) is the two-stage hyperexponential distribution with $\pi_1=0.744$ and $\mu_2/\mu_1=20$, which give the squared coefficient of variation $c^2=5$.

Note that in figs. 3(a) and (b) $p_N(n)$ is plotted in logarithmic scale. If F(t) were an exponential distribution (i.e., c=1), then the curve would be a straight line with slope log p, whereas figs. 3(a) and (b) indicate that $p_N(n)$ is concave for Erlangian distribution, and convex for hyperexponential distribution. Furthermore, $p_N(0)<p_N(1)$ for Erlangian distributions and $p_n(N)>p_N(N-1)$ for hyperexponential distributions. In fact, these properties hold for a much broader class of distributions than Erlangian or hyperexponential distribution; namely, the queue size distribution (in logarithmic scale) is concave if F(t) has coefficient variation c<1, and is convex if c>1. A theoretical basis for such a statement is provided, for example, by the diffusion approximation method [11, 12].

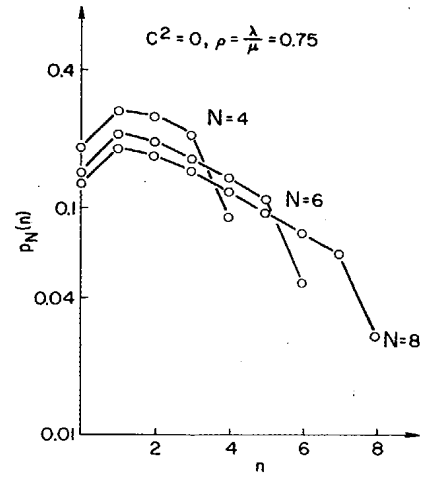


Fig. 3a. The queue size distribution of server 1, when server 1 has a constant service time $\frac{1}{\mu}$.

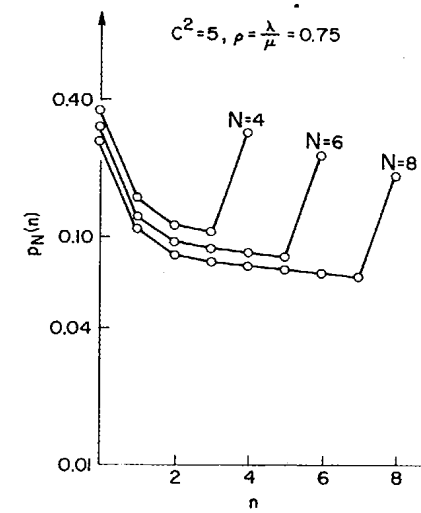


Fig. 3b. when server 1 has a hyperexponential distribution with the mean $\frac{1}{\mu}$ and the squared coefficient variation $c^2=5$.

The utilization or productivity u_1 of the general server as a function of the number of jobs N is depicted in figs 4(a) and (b), where the curves are grouped according to ρ, the ratio of the mean service times. Within each group c^2, the squared coefficient of variation of service time at server 1, ranges from 0 to 100. The case $c^2=0$ represents the constant service time, and $c^2=5$, 20, 100, are realized by appropriate choices of parameters in hyperexponential server with m=2, which are tabulated in table 1. The dotted chain curve in each group represents $c^2=1$, exponential service time. Note that u_2 is simply proportional to u_1 according to eq. (8). As $N\to\infty$, the solution approaches the result for the unconstrained M/G/1 queue, and therefore we find that

$$u_1 \to \min\{\rho,1\} , \quad u_2 \to \min\{\rho^{-1},1\} \quad \text{for } n\to\infty . \tag{23}$$

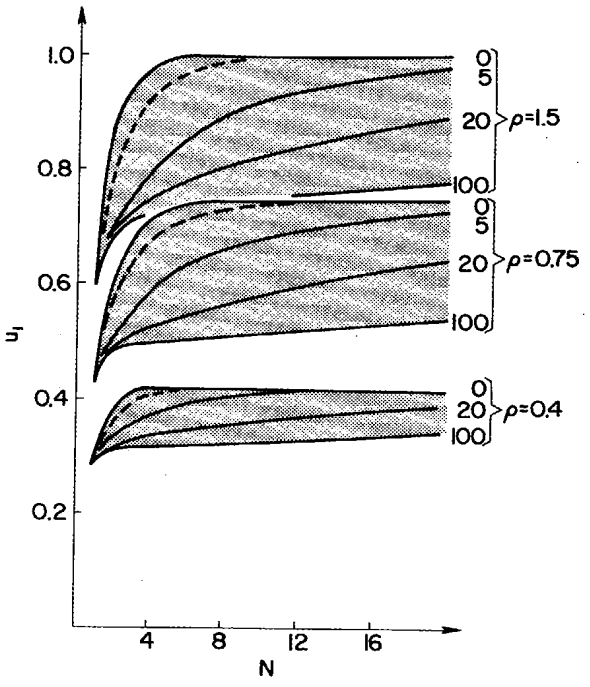


Fig. 4a. The utilization of server 1 vs. multiprogramming N,
$\rho = \frac{\lambda}{\mu}$ is 1,5, 0.75 or 0.4

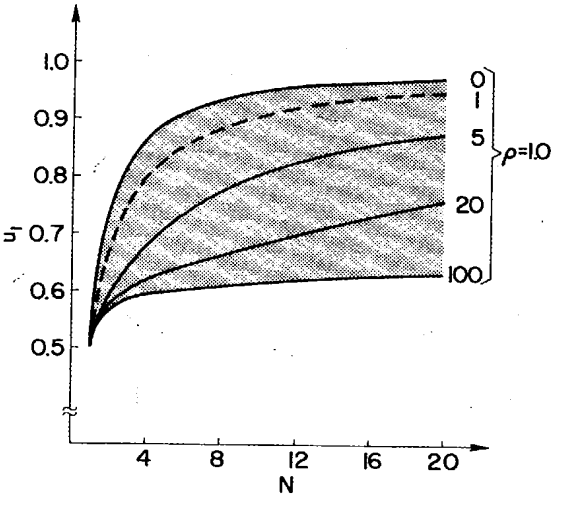


Fig. 4b. ρ=1

Table 1
The parameters of the hyperexponential distributions used

d^2	μ_2/μ_1	π_1
5	20	0.744
20	50	0.949
100	250	0.99

The value of u_1 for N=1 is

$$u_1 = (1+\rho)^{-1} \quad , \quad u_2 = \rho(1+\rho)^{-1} \quad . \tag{24}$$

This is clearly the smallest value of u_1 and therefore we find the following lower and upper bounds

$$(1+\rho)^{-1} \leq u_1 \leq \min\{\rho,1\} \tag{25a}$$

$$\rho(1+\rho)^{-1} \leq u_2 \leq \min\{\rho^{-1},1\} \quad . \tag{25b}$$

From figs. 4(a) and (b) we see that the smaller c^2 is and the more ρ derivates from unity (i.e., the more unbalanced the system is), the faster u_1 approaches the asymptotic limit. Compared to exponential service with the same mean, the utilization is larger for $c^2<1$ and smaller for $c^2>1$, which was reported earlier by Gaver [1]. For large c^2, the convergence to the asymptotic value may become extremely slow and throughout the practical range for N, U_1 may be close to the lower limit. In such cases, the lower limit of eq. (25) should be used for design estimate rather than the values obtained under the exponential server assumption.

The average queue size of the general server, $\bar{n}_1$, are depicted in fig. 5 as a function of N with ρ and c^2 chosen as parameters. Asymptotically,

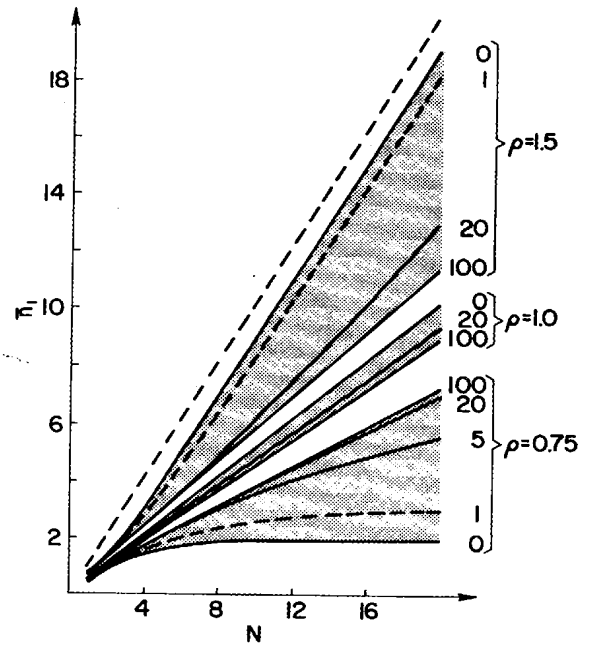


Fig. 5. The average queue size of server 1 vs. degree of multiprogramming N.

$\bar{n}_1$ approaches a constant value

$$\bar{n}_1 \to \rho + \frac{\rho^2(1+c^2)}{2(1-\rho)} \quad \text{for} \quad N \to \infty \quad \text{if} \quad \rho < 1 \tag{26}$$

which is a well-known formula in the theory of the M/G/1 queue. If ρ>1, the server 1 is the bottleneck, therefore, $\bar{n}_1$ increases without bound in parallel with the straight line $\bar{n}_1$=N. It is interesting that $\bar{n}_1$ is the least sensitive to changes in c^2 if ρ=1.

The average response time (sum of waiting time and service time) $\bar{T}_1$ at the general server is obtained by Little's formula

$$\mu_1 \bar{T}_1 = \bar{n}_1/u_1 \tag{27}$$

and the corresponding curves are similar to those of $\bar{n}_1$ but show more variation with c^2 [13].

ACKNOWLEDGEMENT

The authors want to thank Dr. W. D. Frazer for his careful reading of the manuscript.

REFERENCES

[1] D. P. Gaver, Probability models for multiprogramming computer systems, Journal of the ACM vol. 14, no. 3, July 1967, 423-438.

[2] P. A. W. Lewis and G. S. Shedler, A cyclic-queue model of system overhead in multiprogrammed computer systems, Journal of the ACM, vol. 18, no. 2, April 1971, 199-220.

[3] J. Mitrani, Nonpriority multiprogramming systems under heavy demand conditions, Journal of the ACM, vol 19, no. 3, July 1972, 445-452.

[4] I. Adiri, Queueing models for multiprogrammed Computers, in Procedings of the Symposium on Computer-Communication Networks and Teletraffic, Polytechnic Press, New York, 1972, 441-448.

[5] H. Kobayashi, Some recent progress in analytic studies of system performance, in Proceedings of the First USA-Japan Computer Conference, Tokyo, Japan, October 3-5, 1972, 130-138.

[6] H. A. Anderson, Jr. and R. G. Sargent, A statistical evaluation of the scheduler of an experimental interactive computing system, in Statistical computer performance evaluation, edited by W. Freiberger, Academic Press, New York, 1972, 73-98.

[7] J. Keilson, Green's function methods in probability theory, Hafner Publishing Company, New York, 1965, Chapter 6, 147-172.

[8] J. Keilson, The role of Green's functions in congestion theory, in Proceedings of the Symposium on Congestion Theory, edited by W. L. Smith and W. E. Wilkinson, University of North Carolina, Chapel Hill, North Carolina, August 24-26, 1964, The University of North Carolina Press, 43-71.

[9] H. Kobayashi and H. F. Silverman, Some dispatching priority schemes and their effects on response time distribution, Research Report RC-3584, IBM T. J. Watson Research Center, Yorktown Heights, New York, October 1971.

[10] J. H. Griesmer, and R. D. Jenks, Experience with an on-line symbolic mathematics system, in Proceedings of the ONLINE 72 Conference, Brunel University, Middlesex, England, September 4-7, 1972, vol. 1, 457-476. (Also available as IBM Research Report RC-3975, entitled "The SCRATCHPAD System").

[11] H. Kobayashi, Applications of the diffusion approximation to queueing networks: Parts I and II. To appear in the Journal of ACM, vol. 21, no. 2, April 1974 and vol. 21, no. 3, July 1974.

[12] M. Reiser and H. Kobayashi, On the accuracy of the diffusion approximation for queueing systems, IBM Journal of Research and Development, vol. 18, no. 2, March 1974.

[13] M. Reiser and H. Kobayashi, The effects of service time distributions on system performance in a multi-rogrammed computer system model, Research Report RC-4671, IBM T. J. Watson Research Center, Yorktown Heights, New York, November 1973.