AT&T Bell Laboratories Technical Journal
Vol. 63, No. 8, October 1984
Printed in U.S.A.

An Approximate Analysis of Sojourn Times in the M/G/1 Queue With Round-Robin Service Discipline

By P. J. FLEMING*

(Manuscript received February 24, 1984)

In most time-shared computer systems a program is processed by the central processing unit for, at most, a fixed period of time called a time slice, or quantum. If the program requires more processing after it has received its quantum, it is placed at the end of a run queue. This procedure is repeated until the program has finished executing. To the user who submitted the program the two most important performance measures of such a system are the mean and variance of the program's total elapsed time of execution. This total elapsed time is often referred to as the "response time". In this paper we investigate the effect of the quantum size on the mean and variance of the response time.

I. INTRODUCTION

The round-robin queue has been studied by several authors as a model of time-shared computer systems. In a time-shared system, the arrivals of requests for service as well as the service times may be thought of as random variables. From the user's point of view, the two most important measures of performance in such a system are the mean and variance of the response time. The round-robin discipline implicitly favors jobs with shorter service times, in the sense that the mean response time is approximately a linear function of the service time.[1] Thus far, however, the variance has proved to be intractable in

* AT&T Bell Laboratories.

the case of a general service-time distribution. In the case of exponential service times, Muntz has found the Laplace transform of the waiting-time distribution.[2]

The round-robin model can be described as follows: New arrivals join the end of the queue, and all jobs in the queue are served on a first-come first-served basis until they have completed their service requirement or have received one quantum of service. When a job has completed service, it leaves the system, and the next job in the queue begins service immediately. If a job requires more service after receiving its quantum, it rejoins the queue. In this paper we assume that the arrival process is Poisson but the service times are governed by a general distribution. Overhead due to switching between jobs can be included by adjusting the service requirement. For simplicity of exposition we assume that the quantum is constant. Variable quantum sizes also yield to our method of analysis.

In this paper we address the following questions:

1. What value of the quantum minimizes the mean sojourn time of a given class of jobs?

2. For a given class of jobs what is the variance of the sojourn time and what quantum minimizes the variance in the sojourn time?

Here "sojourn time" refers to the total amount of time that a job is in the queueing system, both in the queue and in service. To answer the second question, we use a new light traffic–heavy traffic interpolation (which we will call the RS interpolation) developed by M. Reiman and B. Simon, which makes use of a "light traffic derivative".[3]

In Section II we describe exactly how one calculates the mean waiting time in a round-robin queue. In Section III the results on minimizing the response time and a simple method for finding the optimal quantum are presented. The RS interpolation is described in Section IV, and in Section V we present some numerical examples.

II. THE MEAN WAITING TIME

The mean waiting time in a round-robin queue has been studied by several authors.[4,5] By "waiting" time we mean the total amount of time that the job spends in the queue (but not in service). In this section we describe the authors' analysis and set the notation that will be used throughout the paper. As mentioned above, we assume that the arrival process is Poisson with rate λ and that the service times of newly arriving jobs are independent, identically distributed random variables with distribution function F and density f. Let q denote the quantum size. We say that a job in the system is type j if its service time requirement as a newly arriving job is between $(j-1)q$ and jq, and we say that it is type (i, j) if it is type j and has received between $(i-1)q$ and iq units of service.

The following are additional notations:

m is the mean service time.

ρ is the traffic intensity, λm.

p_j is the probability a newly arriving job is type j.

m_{ij} is the mean amount of time that a type (i, j) job in the queue will occupy the server the next time it receives service.

M_{ij} is the second moment of the amount of time that a type (i, j) job in the queue will occupy the server the next time it receives service.

R_{ij} is the mean forward recurrence time of a type (i, j) job.

Q_{ij} is the mean number of type (i, j) jobs in the queue.

ω_i is the mean amount of time that a job must wait in the queue in the ith time in the queue.

W is the mean amount of time that an arbitrary job must wait in the queue before it completes its total service-time requirement.

2.1 Linear equations for the ω_i

A derivation of the following equations is contained in Ref. 5.

$$\omega_1 = \sum_{\substack{j\geq 1\\ 1\leq i\leq j}} m_{ij}Q_{ij} + \lambda p_j m_{ij} R_{ij}, \tag{1}$$

and for $n \geq 2$

$$\omega_n = \sum_{\substack{j\geq n\\ 1\leq i\leq j-n+1}} m_{ij}Q_{ij} + \lambda p_j m_{ij} m_{n+i-1,j}$$

$$+ \sum_{i=1}^{n-1} \left(\sum_{k\geq i} \lambda p_k m_{ik} \right) (\omega_{n-i} + q). \tag{2}$$

Using Little's Law, which takes the form $Q_{ij} = \lambda p_j \omega_i$ in this case, we can eliminate Q_{ij} from (1) and (2). One can easily verify that the matrix form of the equations for the ω_i is

$$\omega = \rho M \omega + \rho b, \tag{3}$$

where ω is the vector whose ith component is ω_i, and M and b are a matrix and a vector, respectively, that are independent of ρ. Finally, the mean waiting time, W, is given by

$$W = \sum_{i=1}^{\infty} p_i \omega_i. \tag{4}$$

The following fact can be used to simplify the expression for b:

$$b_1 = 2m^{-1} \sum_{\substack{j\geq 1\\ 1\leq i\leq j}} p_j M_{ij},$$

and for $n \geq 2$

$$b_n = q. \tag{5}$$

2.2 *Favoring a class of jobs*

Using the above results we can compute the mean waiting time W^1 of a particular class of jobs if we are given the service-time distribution F_1 of arrivals of that class:

$$W^1 = \sum_{j=1}^{\infty} (1 - F_1((j - 1)q))\omega_j.$$

Figure 1 suggests that, at least in some fairly typical cases, the quantum size that minimizes the mean waiting time of a preferred class of jobs is neither very big [which is, in essence, First-In First-Out (FIFO)] nor very small (processor-sharing), but is just big enough to let all the preferred jobs complete service without having to feed back.

In this example, F_1 is uniform between one and three, F_2 is uniform between four and eight, and $F = 0.75\ F_1 + 0.25\ F_2$. In addition, $\lambda = 0.1$, so $\rho = 0.3$. Note that $q = 3$ minimizes W^1 over all q, and in addition, the mean waiting time of the complementary class decreases with increasing q.

2.3 *A simple method for approximating the optimal quantum size*

Since the expression for W^1 is complicated, it is quite difficult to find the optimal quantum q_0 without significant computational effort. The following observation (from numerical experiments) yields a simple method for finding q_0.

Observation: The value of q that minimizes $[dW^1(0)]/(d\rho)$ is close to the value of q that minimizes W^1 (for arbitrary values of ρ, $0 \leq \rho < 1$).

An easy calculation using (3) implies that

$$\frac{dW^1(0)}{d\rho} = \sum_{j=1}^{\infty} \{1 - F_1[(j - 1)q]\}b_j.$$

Using (5) we see that

$$\frac{dW^1(0)}{d\rho} = b_1 + q \sum_{j=1}^{\infty} [1 - F_1(jq)].$$

III. THE RS INTERPOLATION

3.1 *Description*

We describe the interpolation as it is applied to the steady-state waiting time distribution in the round-robin queue with Poisson arrivals. The interested reader is directed to Ref. 3 for a general

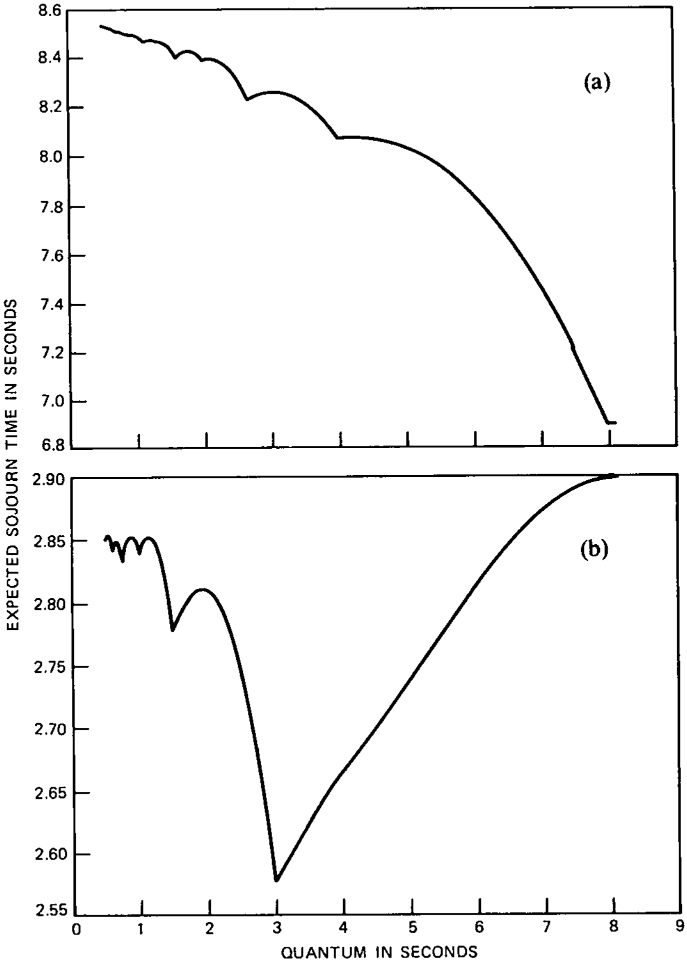

Fig. 1—Expected sojourn times as a function of the quantum for (a) type II jobs and (b) type I jobs.

treatment of the method. Let $W_n(\rho)$, $0 \le \rho \le 1$, be the nth moment of the steady-state waiting time distribution as a function of ρ, the traffic intensity. Let $\overline{W}_n(\rho) = (1 - \rho)^n W_n(\rho)$ when $0 \le \rho < 1$, and let $\overline{W}_n(1) = \lim_{\rho \to 1}(1 - \rho)^n \mathrm{W}_n(\rho)$. $\overline{W}_n(1)$ is well-defined and finite by the argument in Ref. 6.

Note that $\overline{W}_n(0) = W_n(0)$ is zero and $\overline{W}_n(1)$ can be calculated as the heavy traffic limit using a diffusion process. One can interpolate between light and heavy traffic to get the approximation formula:

$$W_n(\rho) \approx \frac{\overline{W}_n(1)\rho + \overline{W}_n(0)}{(1 - \rho)^n} = \frac{\overline{W}_n(1)\rho}{(1 - \rho)^n}.$$

This idea is, of course, well known. The novelty of the RS interpolation is that it makes use of the derivative of $W_n(\rho)$ at $\rho = 0$, which we will denote by $W_n'(0)$. It is clear that $W_n'(0) = \bar{W}_n'(0)$. The RS interpolation is

$$W_n(\rho) \approx \frac{(\bar{W}_n(1) - W_n'(0))\rho^2 + W_n'(0)\rho}{(1 - \rho)^n}.$$

3.2 The light traffic derivative

Let q be a fixed quantum, and let $\sigma(x, a, b)$ be the total waiting time of a tagged job, J_1, that requires a units of service given that in the entire history of the system exactly one other job, J_2, arrives x units of time after J_1 and requires b units of service. Here, a and b are nonnegative real numbers, and x is any real number with negative x, implying that J_2 arrived before J_1. Figure 2 describes σ.

Let

$$\bar{\sigma}_n(a, b) = \int_{-\infty}^{\infty} \sigma(x, a, b)^n dx$$

and let F_i be the service-time distribution of J_i, $i = 1, 2$. The following theorem allows us to calculate $W_n'(0)$.

Theorem 1:

$$W_n'(0) = E(F_2)^{-1} \int_0^{\infty} \int_0^{\infty} \bar{\sigma}_n(a, b) dF_1(a) dF_2(b).$$

The proof of a more general version of this result is contained in Ref. 3. We now present a formula for $W_n'(0)$ using the above theorem. A straightforward calculation yields

$$\begin{aligned}\bar{\sigma}_n(a, b) = {} & \{b - \max[0, (\lfloor b/q \rfloor - \lfloor a/q \rfloor)q]\}^{n+1}/(n + 1) \\ & + \max(0, \lfloor b/q \rfloor - \lfloor a/q \rfloor)[(\lfloor a/q \rfloor + 1)^{n+1} \\ & \quad - \lfloor a/q \rfloor^{n+1}]q^{n+1}/(n + 1) \\ & + \max(0, \lfloor a/q \rfloor - \lfloor b/q \rfloor)qb^n + q^{n+1} \sum_{k=0}^{\min} k^n,\end{aligned}$$

where min = min $(\lfloor a/q \rfloor, \lfloor b/q \rfloor)$ and $\lfloor x \rfloor$ is the greatest integer less than x.

Hence,

$$\begin{aligned}W_n'(0) &= E(F_2)^{-1} \sum_{i,j \geq 0} \int_{jq}^{(j+1)q} \int_{iq}^{(i+1)q} \bar{\sigma}_n(a, b) dF_1(a) dF_2(b) \\ &= E(F_2)^{-1} \sum_{i,j \geq 0} (F_1((i + 1)q) - F_1(iq)) \int_{jq}^{(j+1)q} \bar{\sigma}_n(i, b) dF_2(b).\end{aligned}$$

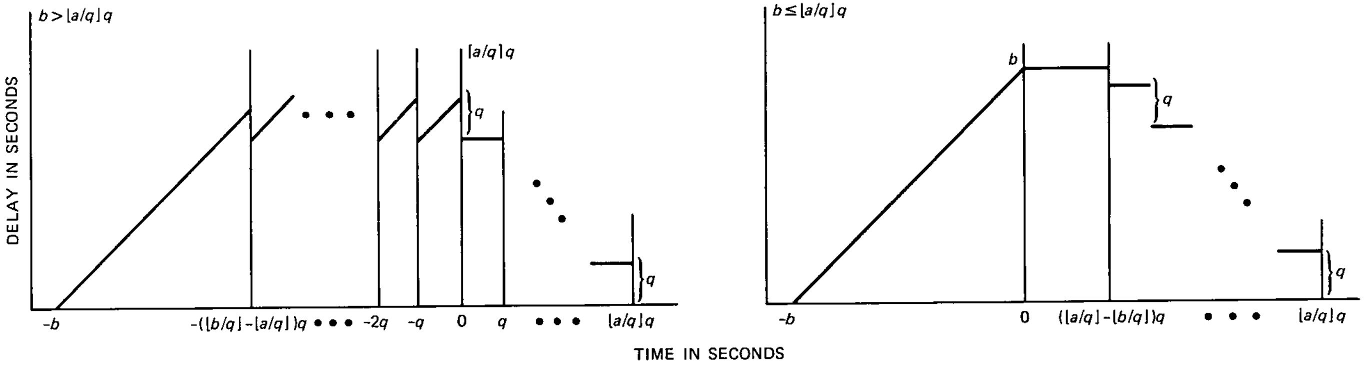

Fig. 2—A description of σ.

And

$$\begin{aligned}\int_{jq}^{(j+1)q} \bar{\sigma}_n(i, b)dF_2(b) = {} & M_{j,n+1}^2[\max(0, j - i)q]/(n + 1) \\ & + M_{j,n}^2(0)\max(0, i - j)q \\ & + M_{j,0}^2\left[q^{n+1} \sum_{k=0}^{\min(i,j)} k^n + \max(0, j - i)\right. \\ & \left.\cdot[(i + 1)^{n+1} - i^{n+1}]q^{n+1}/(n + 1)\right].\end{aligned}$$

Here $M_{j,k}^2(x) = \int_{jq}^{(j+1)q} (b - x)^k dF_2(b)$.

3.3 The heavy traffic limit

The queueing system under consideration in this paper is a special case of the multiclass feedback queue analyzed in Ref. 6. Here we follow the development in Ref. 7 for the readers' convenience.

Let $U(t)$ denote the unfinished work process. Formally,

$$U(t) = V(t) - \inf_{0<s<t} \{V(s)\},$$

where $V(t) = L(t) - t$ and $L(t)$ is the total amount of work entering the system in $[0, t]$. (We assume the system is empty at $t = 0$.) Define a sequence of systems whose parameters, and queue-length and sojourn-time processes are indexed by $n \geq 1$, and consider the normalized processes

$$\hat{U}_{(n)}(t) = \frac{U_{(n)}(nt)}{\sqrt{n}}, \qquad n \geq 1$$

over some finite interval, which we normalize to $[0, 1]$ for convenience. For a heavy traffic limit we assume $\lambda^{(n)} \longrightarrow m^{-1}$ as $n \longrightarrow \infty$. We have the result of D. C. Igelhart and W. Whitt (1971).

Theorem 2: If

$$\lim_{n\to\infty}(\rho^{(n)} - 1)\sqrt{n} = c, \qquad -\infty < c < \infty,$$

then as $n \longrightarrow \infty$

$$\hat{U}^{(n)} \Rightarrow \hat{U} = \mathrm{RBM}[c, m^{-1}(s + m^2)],$$

where $\Rightarrow$ *denotes weak convergence.*

Here s is the variance in the service times, and a is the variance in the interarrival times. $\mathrm{RBM}(d, \sigma^2)$ is one-dimensional reflected Brownian motion with drift d and infinitesimal variance σ^2.

Now consider the normalized queue-length process

$$\hat{Q}^{(n)}_{ij}(t) = \frac{Q^{(n)}_{ij}(nt)}{\sqrt{n}}, \qquad 0 \le t \le 1,\ n \ge 1,\ j \ge 1,\ 1 \le i \le j.$$

We have

$$U^{(n)}(t) \approx \sum_{j \ge 1} \sum_{i=1}^{j} \hat{Q}^{(n)}_{ij}(t) \sum_{k=i}^{j} m_{kj}.$$

Here ≈ means

$$\lim_{n \to \infty} \sup_{0 \le t \le 1} \left| \hat{U}^{(n)}(t) - \sum_{j \ge 1} \sum_{i=1}^{j} \hat{Q}^{(n)}_{ij}(t) \sum_{k=i}^{j} m_{kj} \right| = 0$$

in probability.

Theorem 3: Let

$$Y^{(n)}(t) = \sup | \lambda p_{j'} \hat{Q}_{ij}(t) - \lambda p_j \hat{Q}^{(n)}_{i'j'}(t) |,$$

where the supremum is taken first over all $1 \le i \le j$ and $1 \le i' \le j'$ and then over all $j, j' \ge 1$. If $\lambda^{(n)} \to m^{-1}$, then as $n \to \infty$, $\sup_{0 \le t \le 1} Y^{(n)}(t)$ converges in probability to zero.

Using these results one can prove the following.

Theorem 4: Under the conditions of Theorem 2,

$$\hat{Q}^{(n)}_{ij} \Rightarrow \lambda p_j \Gamma^{-1} \hat{U}$$

as $n \to \infty$, where

$$\Gamma = \sum_{j \ge 1} \lambda p_j \sum_{i=1}^{j} i m_{ij}.$$

Next we consider sojourn times. Again we use the normalization

$$\hat{\omega}^{(n)}_i(t) = \frac{\omega^{(n)}_i(nt)}{\sqrt{n}}, \qquad 0 \le t \le 1,\ n,\ i \ge 1.$$

For single-pass jobs (i.e., $i = 1$),

$$\hat{\omega}^{(n)}_1(t) \approx \sum_{\substack{j \ge 1 \\ 1 \le i \le j}} m_{ij} \hat{Q}^{(n)}_{ij}(t).$$

So from Theorem 4 we see that

$$\hat{\omega}^{(n)}_1(t) \approx \Gamma^{-1} \hat{U}^{(n)}(t) \sum_{\substack{j \ge 1 \\ 1 \le i \le j}} \lambda p_j m_{ij}.$$

Thus, from $\rho^{(n)} \to 1$ and Theorem 3 one can prove Theorem 5.

Theorem 5: Under the conditions of Theorem 2,

$$\hat{\omega}^{(n)}_i \Rightarrow i \Gamma^{-1} \hat{U}$$

as $n \to \infty$.

Let $\hat{\omega}_1(t) = \Gamma\hat{U}(t)$ so that

$$\hat{\omega}_1(t) = \mathrm{RBM}[-\Gamma^{-1}, \Gamma^{-2}(\lambda s + \lambda^{-1})].$$

Since we are interested only in the stationary behavior of the queueing system, we observe that the stationary density of RBM (α, β), $(\alpha < 0)$ is

$$2\,|\,\alpha\,|\,\beta^{-1}\exp(2\alpha\beta^{-1}x).$$

So the kth moment of the stationary distribution of RBM(α, β) is

$$k!\,(2\,|\,\alpha\,|\,\beta^{-1})^{-k}$$

for $k = 1, 2, \cdots$.

Let $\omega_i^k(\rho)$, $0 \le \rho < 1$ be the kth moment of the amount of time a job waits in the queue the ith time in the queue, and let

$$\tilde{\omega}_i^k(\rho) = \begin{cases} \omega_i^k(\rho)(1-\rho)^k & 0 \le \rho < 1 \\ \lim\limits_{\rho\to1}\omega_i^k(\rho)(1-\rho)^k & \rho = 1. \end{cases}$$

The above results yield Theorem 6.

Theorem 6: For all i,

$$\tilde{\omega}_i^k(1) = \omega_1^k(1) = k![2\Gamma/(\lambda s + \lambda^{-1})]^{-k}.$$

Furthermore, $\tilde{W}_n(1)$, is given by

$$\tilde{W}_k(1) = \tilde{\omega}_1^k(1) \sum_{j\ge1} j^k p_j.$$

V. SOME NUMERICAL EXAMPLES

In this section, we present numerical examples of two sorts. Figures 3 and 4 demonstrate the accuracy of the interpolation when applied to the mean sojourn time. Here we are comparing the interpolation with the exact formula given in Section II. In addition, these examples tell us that for some typical load situations, the quantum that minimizes the sojourn time for the type I jobs does not seriously degrade sojourn time for the type II jobs.

Figures 5 through 10 present the approximation to the variance in the waiting time for some typical choices of service times using the RS interpolation to approximate the second moment $W_2(\rho)$ of the waiting time. An interesting thing here is the similarity between the mean and variance regardless of the service-time distribution. Notice that the qualitative properties of the mean and variance of the sojourn time as a function of the quantum size are quite sensitive to the service-time distribution. In the case of exponential service times it appears that the waiting time gets smaller as the quantum gets smaller,

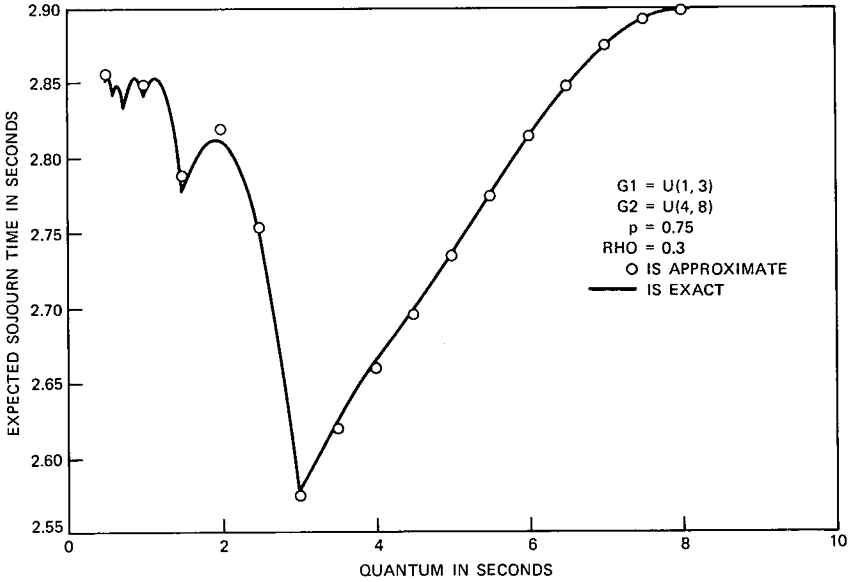

Fig. 3—Expected sojourn time for type I jobs.

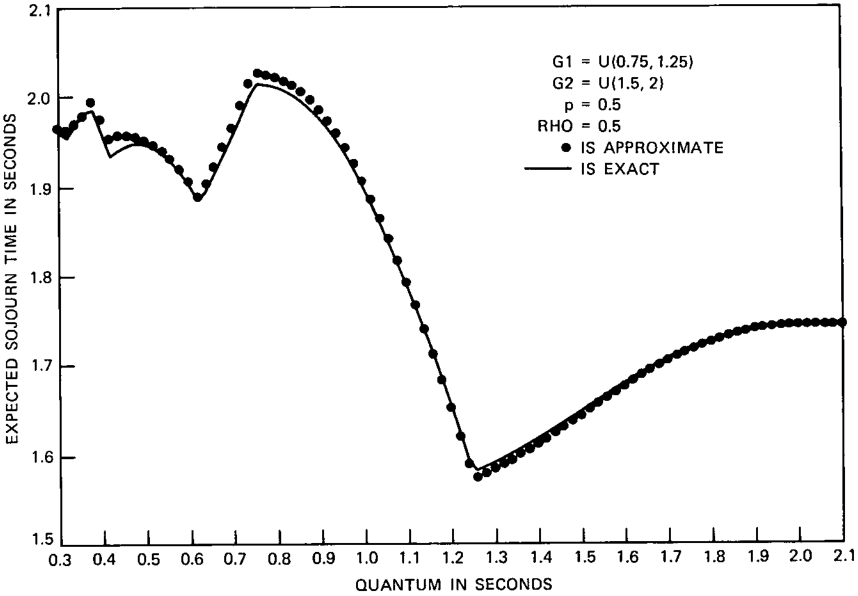

Fig. 4—Expected sojourn time for type I jobs.

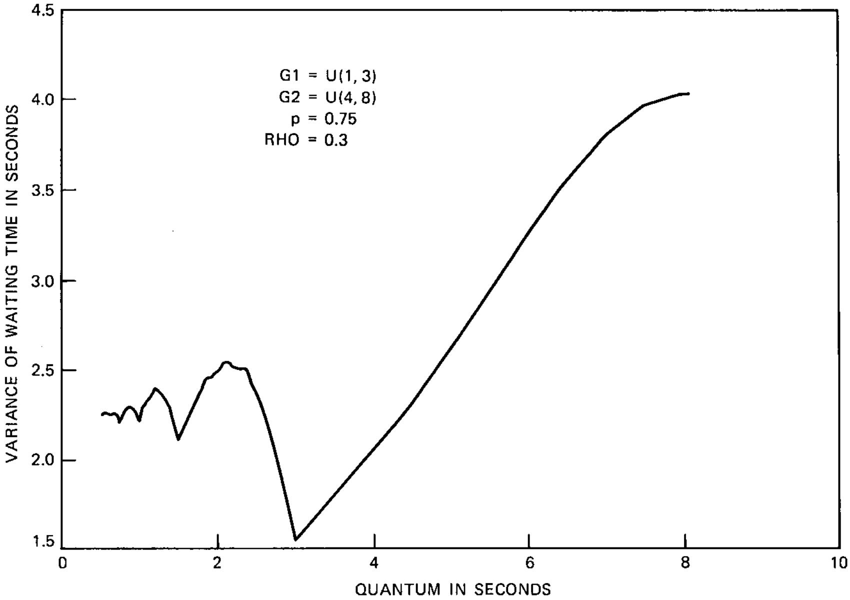

Fig. 5—Variance of waiting time for type I jobs.

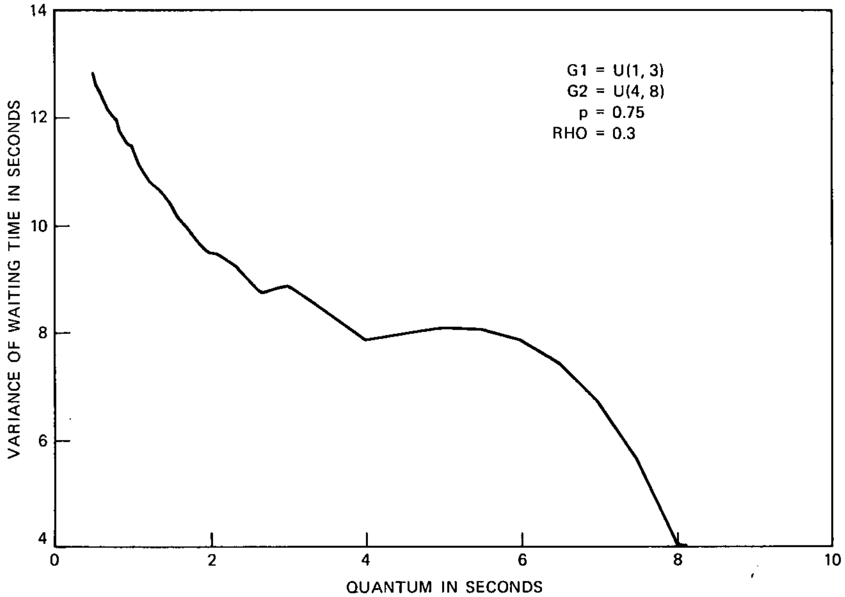

Fig. 6—Variance of waiting time for type II jobs.

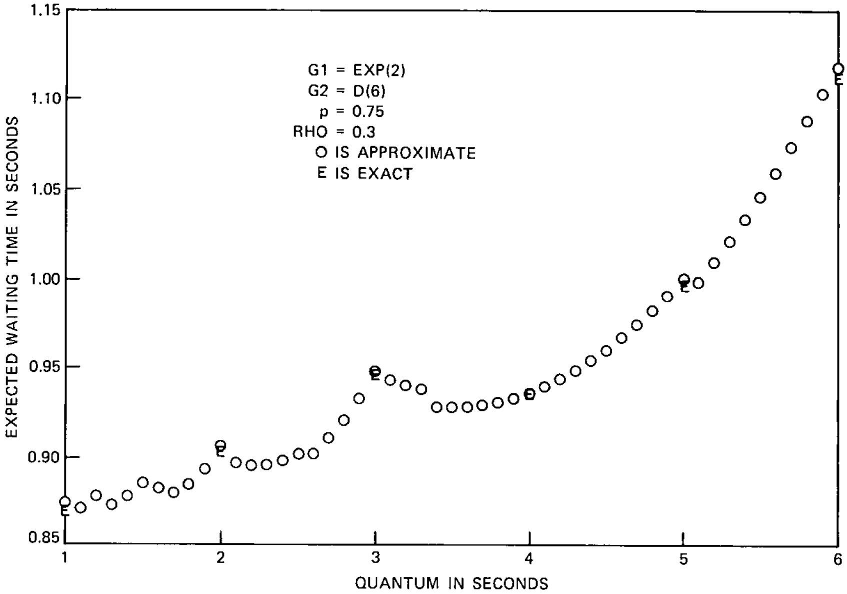

Fig. 7—Expected waiting time for type I jobs.

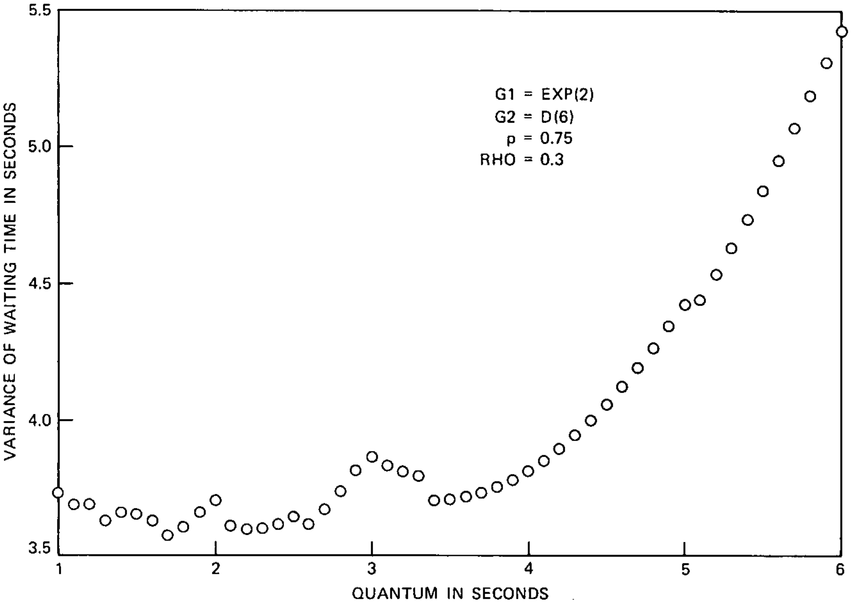

Fig. 8—Variance of waiting time for type I jobs.

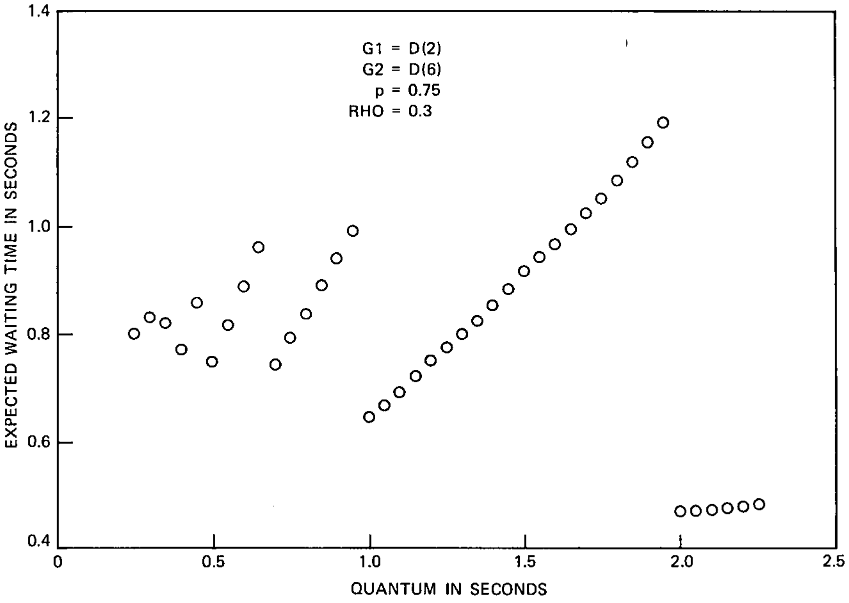

Fig. 9—Expected waiting time for type I jobs.

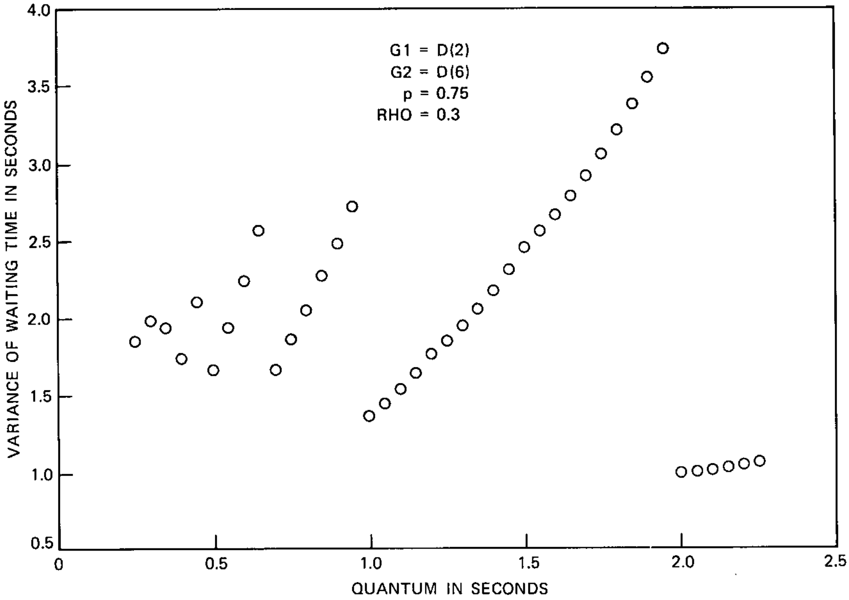

Fig. 10—Variance of waiting time for type I jobs.

whereas in the deterministic and uniform cases it is best to allow all the favored jobs to finish in one quantum.

The following notations are used in Figs. 3 through 10.

rho is the overall traffic intensity of the example queue.

p is the proportion of jobs that are type I.

G1 is the service-time distribution of the type I jobs.

G2 is the service-time distribution of the type II jobs.

EXP(X) means an exponential distribution with mean X.

D means a deterministic distribution with mean X.

U(X,Y) means a distribution that is uniform between X and Y.

VI. ACKNOWLEDGMENT

I would like to thank Marty Reiman and Burt Simon for many useful suggestions.

REFERENCES

1. L. Kleinrock, *Queueing Systems*, Vol. 2, New York: Wiley, 1976, pp. 166–70.
2. R. Muntz, "Waiting Time Distribution for Round-Robin Queueing Systems," Symp. on Computer-Communications Networks and Teletraffic, Polytechnic Institute of Brooklyn (April 4–6, 1972), pp. 429–39.
3. M. I. Reiman, and B. Simon, unpublished work.
4. M. Sakata, S. Noguchi, and J. Oizumi, "An Analysis of the M/G/1 Queue Under Round-Robin Scheduling," Oper. Res., *19*, No. 1 (March–April 1971), pp. 317–85.
5. R. W. Wolff, "Time Sharing With Priorities," SIAM J. Appl. Math., *19*, No. 3 (November 1970), pp. 566–74.
6. M. I. Reiman, unpublished work.
7. E. G. Coffman, Jr. and M. I. Reiman, "Diffusion Approximations for Computer/Communication Systems," in *Mathematical Computer Performance and Reliability*, G. Iazeolla, T. J. Courtois, and A. Hortijk, eds., New York: North Holland, pp. 33–53.
8. D. W. Igelhart and W. Whitt, "Multiple Channel Queues in Heavy Traffic," Advance. Appl. Probab., *2* (1970), pp. 150–77, 355–64.

AUTHOR

Philip J. Fleming, B.A. (Mathematics), 1974, Wayne State University; M.A. and Ph.D. (Mathematics), University of Michigan, Ann Arbor, Michigan, 1977 and 1981, respectively; AT&T Bell Laboratories, 1982—. Before joining AT&T Bell Laboratories, Mr. Fleming was Assistant Professor of Mathematics and Statistics at Case Western Reserve University in Cleveland, Ohio, from 1981 through 1982. Mr. Fleming joined AT&T Bell Laboratories as a Member of Technical Staff in the Systems Design and Exploratory Development department. His work has been in the area of computer performance modeling and analysis, queueing theory, and cryptography as it relates to computer security. He is currently a member of the Operating System Architecture department. Member, AMS and ORSA.