AT&T Technical Journal
Vol. 64, No. 2, February 1985
Printed in U.S.A.

Analysis of a Multistage Queue

By B. T. DOSHI and K. M. REGE*

(Manuscript received June 12, 1984)

Multistage queueing mechanisms with quantum service are suitable in various computer and communication systems to guarantee small delays to short jobs without first knowing the service requirement of any job. In this paper we analyze the efficacy of one such scheme—a two-stage First-In First-Out (FIFO) and Round Robin (RR)—in discriminating between short and long jobs. We obtain the distribution of the delay for short jobs, the cycle time in the RR queue for long jobs, and the number of messages in the FIFO and the RR queues. For the specific parameters used in our numerical results, the two-queue scheme seems to discriminate effectively between the long and short jobs.

I. INTRODUCTION

In computer systems as well as data communication systems, it is frequently desirable to guarantee that short jobs see small delay even under a high load. This may be done at the expense of long jobs. It is also true that in many of these systems the time required to do a job is not known beforehand. Thus simple priority schemes based on the service requirements of jobs are not possible. If the jobs are served in order of arrival, First-In First-Out (FIFO), then all jobs will see long delays at high load. To discriminate between short and long jobs without knowing the type of a job beforehand, various schemes based on quantum service are used. The simplest of these is a Round Robin (RR) scheme. Here, when a job arrives, it is put behind all the waiting

* Authors are employees of AT&T Bell Laboratories.

jobs. When it reaches the server it gets at most Δ time units of service. If its service requirement is smaller than Δ, then the job leaves before the quantum Δ expires. Otherwise, after getting Δ units of service, it is put at the back of the queue and waits for the next pass at the server. Since a shorter job requires fewer passes, its delay is smaller. For this scheme, Wolff[1] obtained the mean delay conditioned on the service requirement as the solution of an infinite system of linear equations. Other schemes are possible if more discrimination is desired between short and long jobs. At one extreme we have a scheme based on infinite number of queues (IQ). In this scheme the server keeps an infinite number of queues numbered 1, 2, ... On arrival a job is placed at the back of the queue numbered "1." When the server completes a service, it takes the first job from the lowest numbered nonempty queue. If this job is from queue n, then it gets at most Δ_n time units of service. If its service is not complete by then, it is put at the back of the queue numbered $n + 1$. Schrage analyzed this scheme and derived the mean and Laplace-Stieltjes transform of the delay conditioned on the service requirement.[2] Somewhere between RR and IQ schemes are the schemes based on a finite number $(N + 1)$ of queues. The first N queues behave as they do in the IQ scheme, while the last one can be either FIFO or RR. In this paper we study one such scheme. In particular, we consider the case where $N = 1$ and the second queue is served round robin. We call this queueing system a FIFO-RR system. The analysis for general N is almost identical but the resulting expressions and notation are more complex.

Fraser and Morgan[3] have analyzed this FIFO-RR discipline as the model of the trunk service discipline in *Datakit*™ Virtual Circuit Switch (VCS) (see Ref. 3 for details of the trunk module operation in *Datakit* VCS). They obtain the mean delay for various classes of jobs under fairly general assumptions, essentially by extending the results in Wolff[1] to the FIFO-RR system. They also use simulation to obtain the percentiles of the delay distributions. In this paper we focus on analytical methods to obtain information about the delay distributions. In particular, we derive a simple expression for the transform of the delay distribution for short jobs under the assumption of Poisson arrivals and general service time distribution. This transform is inverted numerically to obtain the delay distribution. This enables us to get the delay distribution for one-character typed messages and short control messages in communication applications. For jobs long enough to require service in both FIFO and RR queues, the analysis is more difficult. Under more restrictive assumptions, we get the marginal generating function of the number of jobs in the FIFO and RR queues, the transform of the cycle time in the RR queue, and the mean sojourn time in essentially closed form. We illustrate our analysis with nu-

merical results from a data communication application such as the trunk service in *Datakit* VCS. In particular, we show that extremely short jobs see very short delay even under very high overall load. We also discuss how our model may differ from the actual service discipline in data communication applications and the performance implications of these differences.

The analysis presented here for the RR queue uses busy cycle analysis to derive quantities of interest. Recently, Ramaswami showed that some of these quantities can also be derived using matrix methods.[4]

This paper is organized as follows: In Section II we define the model formally and introduce the notation. The delay in the FIFO queue is analyzed in Section III. In Section IV we derive the performance measures for the RR queue. Finally, in Section V we illustrate our results with an application from communication over a 56-kb/s link.

II. MODEL

In this section we formally define the model of the FIFO-RR queues, which we will analyze in Sections III and IV. The analysis of Section III is for the FIFO queue and thus will give the delay distribution for the jobs with service time less than or equal to the quantum size in the FIFO queue. This will be done under fairly weak assumptions. In Section IV we will analyze the RR queue under more restrictive assumptions.

Assume that the arrival process of the jobs is Poisson at rate λ. Let H be the distribution function of the service time. Let Δ_1 and Δ_2 denote, respectively, the quanta of service in the FIFO and the RR queue.

In Section III we will let H be general. In Section IV we will assume that there are two types of jobs. A fraction p of the jobs are short enough to be completed within one quantum in the FIFO queue. Thus, if H_1 is the distribution function of the service time of the short jobs, then

$$H_1(\Delta_1) = 1. \tag{1}$$

The other fraction, $(1 - p)$, of jobs may be long and has distribution function

$$H_2(x) = 1 - e^{-\mu x} \qquad 0 \le x < \infty \tag{2}$$

for some $\mu > 0$. Thus

$$H(x) = pH_1(x) + (1 - p)(1 - e^{-\mu x}), \qquad 0 \le x < \infty. \tag{3}$$

Let h_{i1} and h_{i2} denote the first two moments of H_i, $i = 1, 2$. Of course, $h_{21} = 1/\mu$ and $h_{22} = 2/\mu^2$.

When a job arrives it is put at the back of the FIFO queue. When its turn arrives it gets up to Δ_1 units of service. If the complete service is not rendered by then, the job moves to the RR queue. The RR queue is served in a round robin way with quantum size Δ_2. The FIFO queue has priority over the RR queue to the extent that after each quantum of service, the next service is from the FIFO queue as long as there is work in the FIFO queue.

III. ANALYSIS OF THE FIFO QUEUE

Let

$$q_1 = H(\Delta_1), \tag{4}$$

and for $i \geq 1$,

$$r_i = \frac{H(\Delta_1 + i\Delta_2) - H[\Delta_1 + (i-1)\Delta_2]}{1 - q_1}. \tag{5}$$

Let

$$\bar{N}_2 = \sum_{i=1}^{\infty} i r_i. \tag{6}$$

Thus $\bar{N}_2$ is the expected number of passes at the server in the RR queue given that a job enters the RR queue. Let

$$Q_1(t) = H(t) \qquad 0 \leq t < \Delta_1, \tag{7}$$

and

$$Q_2(t) = \sum_{i=1}^{\infty} \frac{\{H[\Delta_1 + (i-1)\Delta_2 + t] - H[\Delta_1 + (i-1)\Delta_2]\}}{1 - q_1}, \qquad 0 \leq t < \Delta_2. \tag{8}$$

Then the rate of service completions in the FIFO queue is $\lambda_1 = \lambda$, and the distribution of the service time, X_1, in the FIFO queue is given by

$$P\{X_1 \leq t\} = F_1(t) = Q_1(t) = H(t), \qquad 0 \leq t < \Delta_1, \tag{9}$$

and

$$P\{X_1 = \Delta_1\} = F_1(\Delta_1) - F_1(\Delta_1^-) = 1 - H(\Delta_1^-). \tag{10}$$

The rate of service completions in the RR queue is

$$\lambda_2 = \lambda \bar{N}_2 (1 - q_1) \tag{11}$$

and the distribution of the amount of service, X_2, in a typical service in the RR queue is

$$P\{X_2 \leq t\} = F_2(t) = Q_2(t)/\bar{N}_2, \qquad 0 \leq t < \Delta_2, \tag{12}$$

$$P\{X_2 = \Delta_2\} = F_2(\Delta_2) - F_2(\Delta_2^-) = \frac{\bar{N}_2 - 1}{\bar{N}_2} + \frac{1 - Q_2(\Delta_2^-)}{\bar{N}_2}. \quad (13)$$

Now consider a nonpreemptive priority queueing system with two FIFO queues and one server. The arrival rate and the service time distribution in queue i and λ_i are F_i, respectively, $i = 1, 2$. It can be shown using level-crossing arguments (see Refs. 5 and 6) that the distribution of the waiting time in the high-priority queue does not depend on the actual dynamics of arrivals in the low-priority queue. Thus, the waiting time distribution for an arbitrary arrival in queue 1 for this system is the same as that for an arbitrary arrival in the FIFO queue in the original FIFO-RR system. Thus, let $\tilde{f}_1$ and $\tilde{f}_2$ be the Laplace-Stieltjes transforms of F_1 and F_2, respectively, and let $\tilde{W}_1$ be the Laplace-Stieltjes transform of the waiting time in the FIFO queue. Let

$$\zeta_1 = \lambda_1 \int_0^{\Delta_1^+} t dF_1(t), \quad (14)$$

$$\zeta_2 = \lambda_2 \int_0^{\Delta_2^+} t dF_2(t). \quad (15)$$

Then, from Ref. 7,

$$\tilde{W}_1(s) = \begin{cases} \dfrac{s(1 - \zeta_1 - \zeta_2) + \lambda_2[1 - \tilde{f}_2(s)]}{s - \lambda_1 + \lambda_1 \tilde{f}_1(s)} \\ \qquad\qquad \text{if} \quad \zeta_1 + \zeta_2 < 1 \\ \dfrac{1 - \zeta_1}{s - \lambda_1 + \lambda_1 \tilde{f}_1(s)} \cdot \dfrac{\lambda_2[1 - \tilde{f}_2(s)]}{\zeta_2} \\ \qquad\qquad \text{if} \quad \zeta_1 + \zeta_2 \geq 1, \quad \zeta_1 < 1. \end{cases} \quad (16)$$

Equation (16) can be inverted using a method of Jagerman[8] to obtain the waiting time distributions numerically.

Let us now consider the total sojourn time (waiting time + service time) for a job in the FIFO queue. Its Laplace-Stieltjes transform is given by

$$\begin{aligned} \tilde{D}_1(s) &= \tilde{W}_1(s) \tilde{f}_1(s) \\ &= \frac{s(1 - \zeta_1 - \zeta_2) + \lambda_2[1 - \tilde{f}_2(s)]}{s - \lambda_1 + \lambda_1 \tilde{f}_1(s)} \cdot \tilde{f}_1(s). \end{aligned} \quad (17)$$

Also, the transform of the total time in the system for a job that has service requirement $x \leq \Delta_1$ is given by

$$\tilde{D}_{1,x}(s) = \tilde{W}_1(s) e^{-sx}. \quad (18)$$

The FIFO queue is essentially an M/G/1 queue with arrival rate λ and service time distribution F_1. Thus, from Ref. 9, the distribution of the number in the system at an arbitrary instant is the same as that at an arbitrary arrival epoch and is the same as that seen by a random departure from the FIFO queue (either exiting the system or going to the RR queue). Also, the distribution $\{P_{1,K}\}$ of the number in the FIFO queue at a random departure epoch is related to the sojourn time distribution by

$$\hat{P}_1(z) = \sum P_{1,K} z^K = \tilde{D}_1[\lambda(1 - z)]. \tag{19}$$

Thus the generating function of the number in the FIFO queue at an arbitrary instant is given by

$$\hat{P}_1(z) = \tilde{D}_1[\lambda(1 - z)], \tag{20}$$

where $\tilde{D}_1$ is given by eq. (17).

IV. ANALYSIS OF THE RR QUEUE

In this section we mainly derive the expressions for various quantities of interest for the RR queue. However, in that process we also obtain some additional quantities related to the FIFO queue. As mentioned earlier, in this section we will use the following distribution function of the service time:

$$H(x) = pH_1(x) + (1 - p)(1 - e^{-\mu x}), \qquad 0 \le x < \infty,$$

with

$$H_1(\Delta_1) = 1.$$

As in Section III, let X_1 and X_2 denote the service times in typical chunks of service in the FIFO and the RR queues, respectively. Let F_1 and F_2 be the distribution functions of X_1 and X_2, respectively. Then, from Section III, we have

$$F_1(x) = \begin{cases} pH_1(x) + (1 - p)(1 - e^{-\mu x}) & 0 \le x < \Delta_1 \\ 1 & x \ge \Delta_1, \end{cases}$$

$$F_2(x) = \begin{cases} 1 - e^{-\mu x} & 0 \le x < \Delta_2 \\ 1 & x \ge \Delta_2, \end{cases}$$

$$\tilde{f}_1(s) = p\tilde{h}_1(s) + \frac{(1 - p)}{\mu + s}\{\mu + se^{-(s+\mu)\Delta_1}\},$$

$$\tilde{f}_2(s) = \frac{1}{\mu + s}\{\mu + se^{-(\mu+s)\Delta_2}\},$$

$$\zeta_1 = \lambda\left[ph_1 + \frac{(1 - p)(1 - e^{-\mu\Delta_1})}{\mu}\right],$$

and

$$\zeta_2 = \frac{\lambda(1-p)e^{-\mu\Delta_1}}{\mu}.$$

We begin by defining and analyzing various busy periods and cycles associated with the FIFO-RR system. These will be used subsequently to derive quantities of interest. The system is said to be busy as long as a job is being processed at high or low priority. The continuous interval of time during which the system is busy is called a system-busy period. A 1-busy period is started by a job arriving at the system while the server is idle and lasts until no job is left in the FIFO queue (so that the server moves to the RR queue). A 2-busy period is started by a service quantum in the RR queue and lasts until the end of this quantum and the time required to empty the FIFO queue. Note that each service quantum in the RR queue generates a 2-busy period and that a system-busy period consists of exactly one 1-busy period, which triggers off the system-busy period and is followed by zero or more 2-busy periods.

Let $\beta(x, k)$ denote the joint probability that the length of the system-busy period is less than or equal to x and that during this busy period exactly k jobs get routed to the RR queue after completing their service quanta in the FIFO queue. Let

$$\beta(s, z) = \int_{0^-}^{\infty} \sum_{k=0}^{\infty} e^{-sx} z^k dB(x, k) \tag{21}$$

denote the joint transform of $B(x, k)$.

Similarly, let $B_1(x, k)[B_2(x, k)]$ denote the joint probability that the length of a 1-busy period (2-busy period) is less than or equal to x and that during this busy period exactly k jobs are moved to the back of the RR queue after receiving one service quantum during that cycle. In the case of a 2-busy period, k includes the job in the RR queue that started this busy period if it was routed to the back of the RR queue. Let $\beta_1(s, z)$ and $\beta_2(s, z)$ denote the joint transforms of $B_1(x, k)$ and $B_2(x, k)$, respectively:

$$\beta_i(s, z) = \int_0^{\infty} e^{-sx} \sum_{k=0}^{\infty} z^k dB_i(x, k), \qquad i = 1, 2. \tag{22}$$

In the Appendix we obtain expressions for these quantities in the form of functional equations.

We will now derive the expressions for the cycle time, the distribution of the number in the RR queue at an arbitrary instant, and the mean sojourn time in the RR queue. The actual distribution function of the sojourn time does not seem to lead to a simple form.

First we consider the number in the RR queue at special time points. We look at the points in time when a service quantum has just completed and the FIFO queue is empty. The interarrival times of the new arrivals and the remaining service requirements of the jobs in the RR queue are independent random variables with exponential distributions. Thus the number in the RR queue at these imbedded instants forms a Markov chain. Let n_k denote the number in the RR queue at the kth such instant. Then we have the following transition mechanism:

$$n_{k+1} = \begin{cases} n_k - 1 + \ell_k & \text{if } n_k \geq 1 \\ \ell_k' & \text{if } n_k = 0, \end{cases} \tag{23}$$

where ℓ_k denotes the number of jobs sent to the back of the RR queue during a 2-busy period (including the message in the RR queue that started this 2-busy period if it gets sent to the back of the RR queue after completing its service quantum), and ℓ_k' denotes the number sent to the RR queue during a 1-busy period.

Let $\Psi_k(z)$ be the generating function of n_k. Then eq. (23) can be rewritten as

$$\Psi_{k+1}(z) = \frac{[\Psi_k(z) - \Psi_k(0)]\beta_2(0, z)}{z} + \Psi_k(0)\beta_1(0, z), \tag{24}$$

and the equilibrium generating function $\Psi(z) = \lim_{k\to\infty} \Psi_k(z)$ is given by

$$\Psi(z) = \frac{\Psi(0)[z\beta_1(0, z) - \beta_2(0, z)]}{z - \beta_2(0, z)}. \tag{25}$$

Equating $\Psi(1)$ with 1, we get the unknown $\Psi(0)$ as

$$\Psi(0) = \frac{1 - \tilde{b}_2}{1 + \tilde{b}_1 - \tilde{b}_2},$$

where

$$\tilde{b}_i = \left.\frac{d\beta_i(0, z)}{dz}\right|_{z=1} \qquad i = 1, 2. \tag{26}$$

We now evaluate the Laplace-Stieltjes transform of the cycle time defined as the time interval between two successive passes through the server by a job in the RR queue. Let t_1 and t_2 be the instants at which the server begins to provide two successive service quanta to a tagged job in the RR queue. (In case the tagged job leaves the system after receiving the first quantum, t_2 is the instant at which the job would have begun to receive the second quantum had it still been in the system.) Then $t_2 - t_1$ is the cycle time.

Let m denote the number of messages in the RR queue at time t_1. Then the generating function of m is given by

$$E[z^m] = E[z^n \mid n \geq 1], \tag{27}$$

where n is the number in the RR queue at the imbedded instants discussed above. Thus

$$E[z^m] = \frac{\Psi(z) - \Psi(0)}{1 - \Psi(0)}, \tag{28}$$

where Ψ is as given by eq. (25). Now, because of the memoryless property of the service requirements of jobs in the RR queue, the cycle time $t_2 - t_1$ is the sum of m independent and identically distributed 2-busy periods. Thus

$$\begin{aligned} \chi(s) = E[s^{(t_2 - t_1)}] &= E[\beta_2(s, 1)^m] \\ &= \frac{\Psi(\beta_2(s, 1)) - \Psi(0)}{1 - \Psi(0)}. \end{aligned} \tag{29}$$

We can now obtain an expression for the generating function of the number in the RR queue at an arbitrary instant.

Let $\tilde{n}$ and $\hat{n}$ denote, respectively, the number of jobs in the RR queue just after an arbitrary departure from and just before an arbitrary arrival to the RR queue. Then $\tilde{n}$ and $\hat{n}$ have the same distribution. Let n denote, as before, the number in the RR queue at an instant when a service quantum has just completed and the FIFO queue is empty. Then

$$\begin{aligned} P\{\hat{n} = k\} &= P\{\tilde{n} = k\} \\ &= \frac{\begin{array}{c} P\{n = k + 1 \text{ and a departure occurs at} \\ \text{the end of this service quantum}\} \end{array}}{\begin{array}{c} P\{n \geq 1 \text{ and a departure occurs at the} \\ \text{end of this service quantum}\} \end{array}} \\ &= \frac{P\{n = k + 1\}(1 - e^{-\mu\Delta_2})}{P\{n \geq 1\}(1 - e^{-\mu\Delta_2})} \\ &= \frac{P\{n = k + 1\}}{P\{n \geq 1\}}. \end{aligned} \tag{30}$$

Now, the number of jobs in the RR queue just before a randomly selected arrival to that queue is the same as the number in the RR queue when this tagged job began to receive its first (and only) service quantum in the FIFO queue. This number is a function only of the arrivals prior to the arrival of the tagged job. Also, this number is independent of the tagged job's service time in the FIFO queue and,

in particular, is independent of whether or not the tagged job enters the RR queue. Therefore, the number in the RR queue when an arbitrary job completes its service in the FIFO queue has the same distribution as $\hat{n}$. Its generating function is given by

$$\begin{aligned}\xi(z) &= \sum_{K=0}^{\infty} P(\hat{n} = k)z^K \\ &= \sum_{K=0}^{\infty} \frac{P(n = k+1)z^K}{P(n \geq 1)} \\ &= \frac{(\Psi(z) - \Psi(0))}{z(1 - \Psi(0))}. \end{aligned} \tag{31}$$

We can now derive the generating function of the number in the RR queue at an arbitrary instant. If the observation instant lies in an interval of time during which the server is serving the FIFO queue, the number of jobs in the RR queue is the same as when the job being served finishes its service quantum in the FIFO queue, that is, it has the generating function $\xi(z)$. If the server is working on a job in the RR queue, then the number in the RR queue has the same distribution as the variable m defined above, that is, it has the generating function

$$\frac{\Psi(z) - \Psi(0)}{1 - \Psi(0)}.$$

Finally, if at the observation instant the system is empty, the generating function of the number in the RR queue is 1. Thus the generating function of the number in the RR queue at an arbitrary instant is given by

$$\hat{P}_2(z) = \zeta_1 \xi(z) + \zeta_2 \frac{\Psi(z) - \Psi(0)}{1 - \Psi(0)} + (1 - \zeta_1 - \zeta_2). \tag{32}$$

The average number in the RR queue at an arbitrary instant is given by

$$L_2 = \left. \frac{d\hat{P}_2(z)}{dz} \right|_{z=1}. \tag{33}$$

Finally, the mean sojourn time in the RR queue can be obtained by using Little's law for that queue:

$$S_2 = \frac{L_2}{\lambda_2} = \frac{\left. \dfrac{d\hat{P}_2(z)}{dz} \right|_{z=1}}{\lambda(1 - \rho)e^{-\mu\Delta_1}}. \tag{34}$$

V. SPECIAL CASES AND NUMERICAL EXAMPLES

We now consider two special cases of the general model analyzed in Sections III and IV. These examples are typical of some communication applications. The service times here will correspond to the number of characters in the message.

The first case corresponds to three types of jobs: one time unit long, Δ_1 time units long, and $\Delta_1 + n\Delta_2$ time units long ($n \geq 1$). Let λ be the total arrival rate. Let X denote the service time of a job. Let

$$q_{11} = P\{X = 1\}, \tag{35}$$

$$q_{12} = P\{X = \Delta_1\}, \tag{36}$$

$$q_1 = q_{11} + q_{12}, \tag{37}$$

$$r_n = \frac{P\{X = \Delta_1 + n\Delta_2\}}{1 - q_1}, \quad n \geq 1. \tag{38}$$

Then

$$\bar{N}_2 = \sum_{n=1}^{\infty} nr_n \tag{39}$$

$$\begin{aligned} \tilde{f}_1(s) &= q_{11}e^{-s} + q_{12}e^{-s\Delta_1} + (1 - q_1)e^{-s\Delta_1} \\ &= q_{11}e^{-s} + (1 - q_{11})e^{-s\Delta_1}, \end{aligned} \tag{40}$$

$$\zeta_1 = \lambda[q_{11} + \Delta_1(1 - q_{11})] \tag{41}$$

$$\tilde{f}_2(s) = e^{-s\Delta_2}, \tag{42}$$

$$\lambda_2 = \lambda\bar{N}_2(1 - q_1), \tag{43}$$

and

$$\zeta_2 = \lambda_2\Delta_2. \tag{44}$$

Thus,

$$\begin{aligned} \tilde{W}_1(s) &= \frac{s(1 - \zeta_1 - \zeta_2) + \lambda_2[1 - \tilde{f}_2(s)]}{s - \lambda_1 + \lambda_1\tilde{f}_1(s)} \\ &= \frac{\begin{array}{l}(s\{1 - \lambda[q_{11} + (1 - q_{11})\Delta_1] - \lambda\bar{N}_2\Delta_2(1 - q_1)\} \\ \qquad + \lambda\bar{N}_2(1 - q_1)(1 - e^{-s\Delta_2}))\end{array}}{s - \lambda + \lambda(q_{11}e^{-s} + (1 - q_{11})e^{-s\Delta_1})}. \end{aligned} \tag{45}$$

The second example corresponds to the traffic mixture assumed in eqs. (1) and (2), that is, a proportion p of the jobs have service time less than or equal to Δ_1 and others have service time exponentially distributed with mean $1/\mu$. Thus,

$$q_1 = p + (1 - p)(1 - e^{-\mu\Delta_1}), \tag{46}$$

$$1 - q_1 = (1 - p)e^{-\mu\Delta_1}, \tag{47}$$

$$Q_1(t) = pH_1(t) + (1 - p)(1 - e^{-\mu t}), \quad 0 \le t < \Delta_1, \tag{48}$$

$$F_1(t) = Q_1(t) = pH_1(t) + (1 - p)(1 - e^{-\mu t}), \quad 0 \le t < \Delta_1, \tag{49}$$

and

$$F_1(\Delta_1) - F_1(\Delta_1^-) = (1 - p)(1 - e^{-\mu\Delta_1}). \tag{50}$$

Let $\tilde{h}_1$ be the Laplace-Stieltjes transform of H_1. Then, from eqs. (1) and (2), we get

$$\begin{aligned}\tilde{f}_1(s) &= p\tilde{h}_1(s) + (1 - p)\int_0^{\Delta_1} e^{-si}\mu e^{-\mu t}dt + (1 - p)e^{-s\Delta_1 - \mu\Delta_1} \\ &= p\tilde{h}_1(s) + (1 - p)\left[\frac{\mu}{s + \mu}(1 - e^{-\Delta_1(s+\mu)}) + e^{-(s+\mu)\Delta_1}\right] \\ &= p\tilde{h}_1(s) + \frac{(1 - p)}{s + \mu}(\mu + se^{-(s+\mu)\Delta_1}).\end{aligned} \tag{51}$$

Also,

$$\begin{aligned}r_i &= \frac{(e^{-\mu(\Delta_1+(i-1)\Delta_2)} - e^{-\mu(\Delta_1+i\Delta_2)})[1 - p]}{1 - q_1} \\ &= \frac{e^{-\mu\Delta_2(i-1)}e^{-\mu\Delta_1}(1 - e^{-\mu\Delta_2})(1 - p)}{1 - q_1}.\end{aligned} \tag{52}$$

Thus

$$\bar{N}_2 = \frac{e^{-\mu\Delta_1}(1 - p)}{(1 - q_1)(1 - e^{-\mu\Delta_2})} = \frac{1}{(1 - e^{-\mu\Delta_2})}, \tag{53}$$

$$\lambda_2 = \frac{\lambda e^{-\mu\Delta_1}(1 - p)}{(1 - e^{-\mu\Delta_2})}. \tag{54}$$

Finally,

$$F_2(t) = 1 - e^{-\mu t}, \quad 0 \le t < \Delta_2, \tag{55}$$

and

$$F_2(\Delta_2) - F_2(\Delta_2^-) = e^{-\mu\Delta_2}. \tag{56}$$

Thus,

$$\tilde{f}_2(s) = \int_0^{\Delta_2} \mu e^{-\mu t} e^{-st} dt + e^{-\mu\Delta_2} e^{-s\Delta_2}$$

$$= \frac{\mu}{s+\mu}(1 - e^{-(s+\mu)\Delta_2}) + e^{-(s+\mu)\Delta_2}$$

$$= \frac{1}{s+\mu}(\mu + se^{-(s+\mu)\Delta_2}). \tag{57}$$

For ζ_1 and ζ_2, we get

$$\zeta_1 = \lambda\left(ph_1 + \frac{(1-p)(1-e^{-\mu\Delta_1})}{\mu}\right), \tag{58}$$

and

$$\zeta_2 = \frac{\lambda e^{-\mu\Delta_1}(1-p)}{(1-e^{-\mu\Delta_2})} \frac{1-e^{-\mu\Delta_2}}{\mu}$$

$$= \frac{\lambda(1-p)e^{-\mu\Delta_1}}{\mu}. \tag{59}$$

Thus, from (16) we get

$$\tilde{W}_1(s) = \frac{s(1-\zeta_1-\zeta_2) + \lambda_2[1-\tilde{f}_2(s)]}{s - \lambda_1 + \lambda_1\tilde{f}_1(s)}$$

$$= \frac{s\left(1 - \lambda p h_1 - \lambda\dfrac{(1-p)}{\mu}\right) + \dfrac{\lambda_2 s}{s+\mu}(1 - e^{-(s+\mu)\Delta_2})}{s - \lambda p(1-\tilde{h}_1(s)) - \dfrac{\lambda(1-p)s}{\mu+s}(1-e^{-(s+\mu)\Delta_1})}. \tag{60}$$

In communication applications there is usually an overhead associated with each segment of transmitted data. Thus, with a typical segment of X_1 characters transmitted from the FIFO queue, δ_1 overhead characters are added. Similarly, δ_2 characters are added to each segment of data transmitted from the RR queue. The effective numbers of characters sent in a typical service segment from the FIFO and the RR queue are then $X_1' = X_1 + \delta_1$ and $X_2' = X_2 + \delta_2$, respectively. The corresponding transforms are then $\tilde{f}_i'(s) = e^{-\delta_i s}\tilde{f}_i(s)$. If we replace $\tilde{f}_1$ and $\tilde{f}_2$ by $\tilde{f}_1'$ and f_2' and adjust the occupancy numbers accordingly in all the waiting time transforms, the resulting expressions will give transforms of the waiting time in the presence of the overhead characters. Besides these overhead characters associated with each service segment, there are usually overhead characters associated with frames,

that is, data from various service segments are combined into frames of some maximum size and, at the end of each frame, framing protocol characters are added. The exact analysis of the waiting time in presence of the framing overhead is not easy, but good approximations can be obtained by distributing the framing overhead over all transmitted characters. We have chosen to exclude the framing overhead in our numerical calculations.

Next we numerically evaluate performance measures for short and long messages for a few traffic mixes. We assume that the communication link runs at 56 kb/s. Thus, each character corresponds to 1/7 ms of delay. We use $\delta_1 = \delta_2 = 2$ in all the cases described below.

First consider short messages ($\leq \Delta_1$ characters). The Laplace-Stieltjes transform of the waiting time is given by eq. (16) with appropriate modifications to account for the overhead characters. We numerically inverted this transform using the inversion algorithm of Jagerman[8] for the following traffic mixes and quanta sizes (these traffic mixes are selected to give the same mean number of characters per message):

1. $P(X = 1) = 100/111$, $P(X = \Delta_1) = 10/111$, $P(X = \Delta_1 + n\Delta_2) = r_n \times 1/111$, where $\sum_{n=1}^{\infty} r_n = 1$ and $\bar{N}_2 = \sum_{n=1}^{\infty} nr_n = 99/6$. Also, $\Delta_1 = 16$, $\Delta_2 = 48$. This will be called the traffic mix M_1. Note that $\bar{N}_2$ uniquely defines the delay distribution irrespective of the individual values of r_ns.

2. $P(X = 1) = 100/111$ and with probability $11/111$, X is exponentially distributed with mean $912/11$. $\Delta_1 = 16$, $\Delta_2 = 48$. This will be called the traffic mix M_2.

3. For the third traffic mix, M_3, we assume that $P\{X = 1\} = 100/111$, $P\{$is exponentially distributed with mean $40\} = 10/111$ and $P\{X$ is exponentially distributed with mean $512\} = 1/111$. $\Delta_1 = 16$ and $\Delta_2 = 48$.

4. The traffic mix here is the same as in M_3 but we use $\Delta_1 = \Delta_2 = 16$ and $\Delta_1 = \Delta_2 = 64$. These cases will be denoted by M_3' and M_3'', respectively. With these sizes of the quanta the cases M_3' and M_e'' correspond to the cases studied in Refs. 3 and 10 with and without the framing overhead, respectively. We will use the results in Ref. 10 to cross check our calculations.

Figures 1 through 3 show the tails of the delay distribution for short messages under the traffic mixes M_1, M_2, and M_3, respectively. Two occupancy numbers are given for each curve in these figures. The lower number is the raw occupancy, while the larger number indicates the total occupancy including the overhead characters. A number larger than one indicates the saturation of the RR queue and an indication that not all the offered messages will be completely trans-

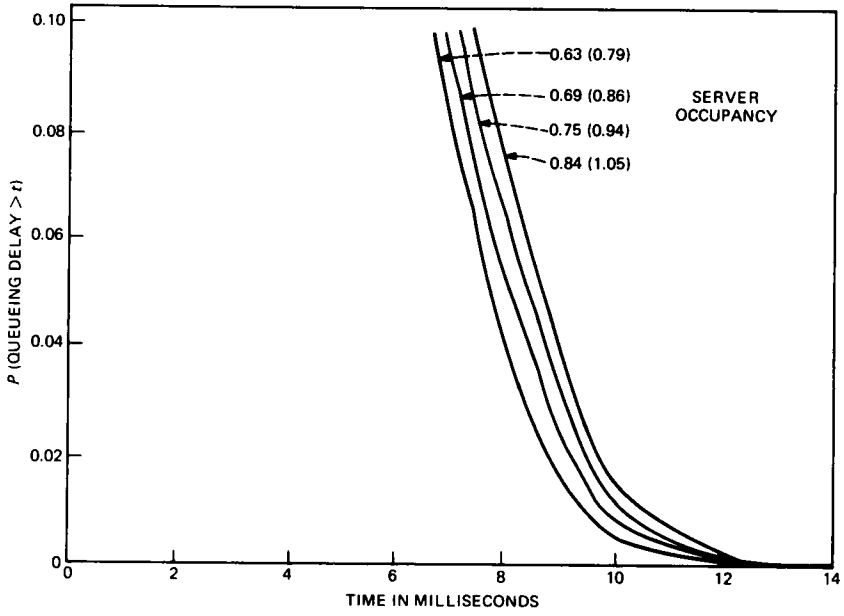

Fig. 1—Delay distribution for traffic mix $M1$: short messages.

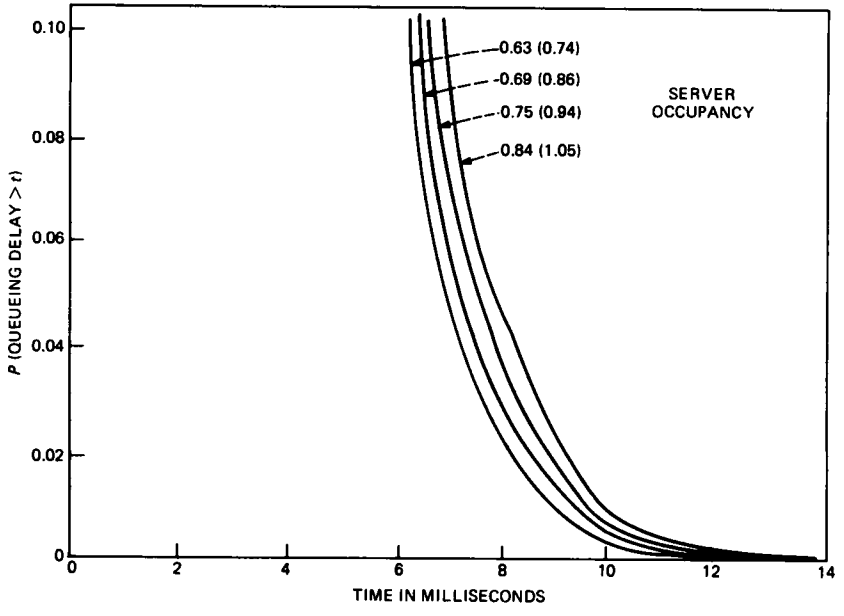

Fig. 2—Delay distribution for traffic mix $M2$: short messages.

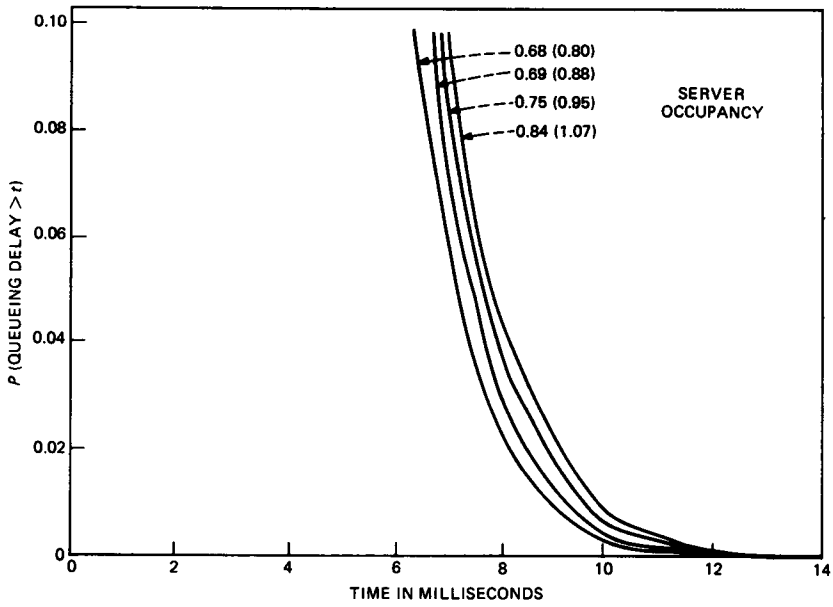

Fig. 3—Delay distribution for traffic mix $M3$: short messages.

mitted. Note, however, that the occupancy of the server due to the FIFO queue is still well below one for these curves.

It is clear from these figures that, for the parameters chosen, the short messages will see very short delays due to queueing even when the long messages see essentially infinite delays. Of course, if the occupancy of the server due to the FIFO is close to one, the short messages will also see long delays.

As mentioned earlier, cases M_3' and M_3'' were selected to match the traffic mix and the quanta sizes of those in Refs. 3 and 10 so that a cross check can be carried out. In these references Fraser and Morgan get the delay distribution (in particular, the 95th percentile of the delay distribution) for short messages via simulation. Since the results in Ref. 3 are obtained in presence of the framing overhead, they cannot be compared directly with our results. However, in their unpublished work Fraser and Morgan[10] obtain the 95th percentile of the delay distribution without the framing overhead. In Fig. 4 we plot the 95th percentile of the delay as a function of the raw occupancy of the server for M_3' and M_3''. The simulation points from Ref. 10 are superimposed on these curves and the agreement looks very good.

We next look at other performance measures studied in Section IV.

We have chosen two different traffic mixes to illuminate the effects of various load parameters on the performance measures relevant to queue 2. The traffic mixes are:

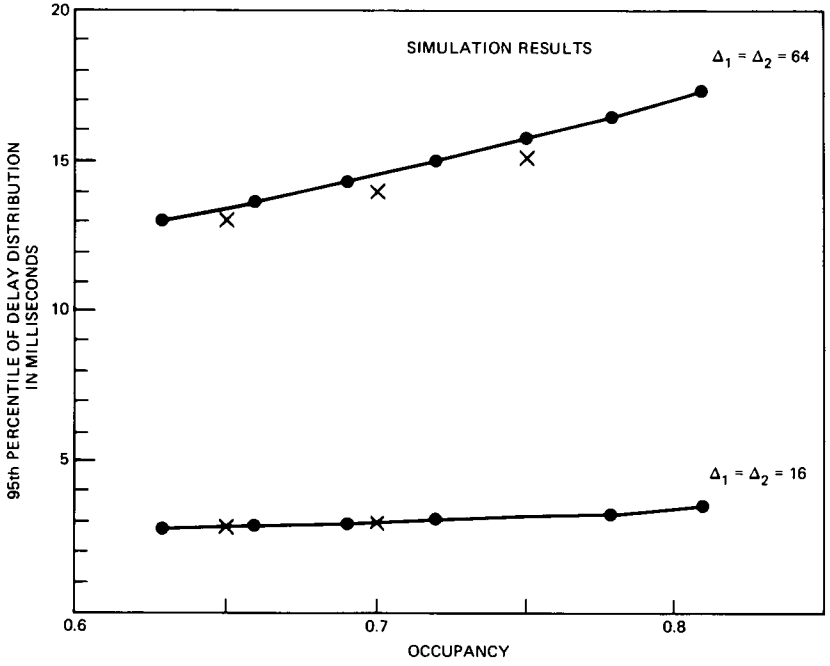

Fig. 4—95th percentile of the distribution for short messages.

1. A message has a single character in it with probability 0.98, and with probability 0.02 it is exponentially distributed with a mean length of five hundred characters. We call this traffic mix M_4.

2. The probability of a message being a single character one is 0.9, and with probability 0.1 it is exponentially distributed with a mean length of one hundred characters. This traffic mix will be referred to as M_5.

In both cases we assume that the "quantum overhead" (i.e., $\delta_1 = \delta_2$) is two characters. The quantum size Δ_1 is assumed to be 16 characters and Δ_2 to be 48 characters. The performance measures relevant to queue 2 include the mean cycle time, the mean sojourn time (for messages entering queue 2), and the mean queue length. These are plotted as functions of the overall occupancy (or, equivalently, the arrival rate) for both traffic mixes in Figs. 5 and 6.

From Figs. 5 and 6 it can be seen that the mean sojourn time for messages entering queue 2 varies essentially in direct proportion to the mean message length of type 2 jobs and in inverse proportion to $(1 - \rho)$, where ρ is the overall utilization of the server. The average cycle time also appears to vary inversely in proportion to $(1 - \rho)$; moreover, it is insensitive to the mean length of type 2 messages as long as it is large compared to Δ_2. The mean queue length also displays

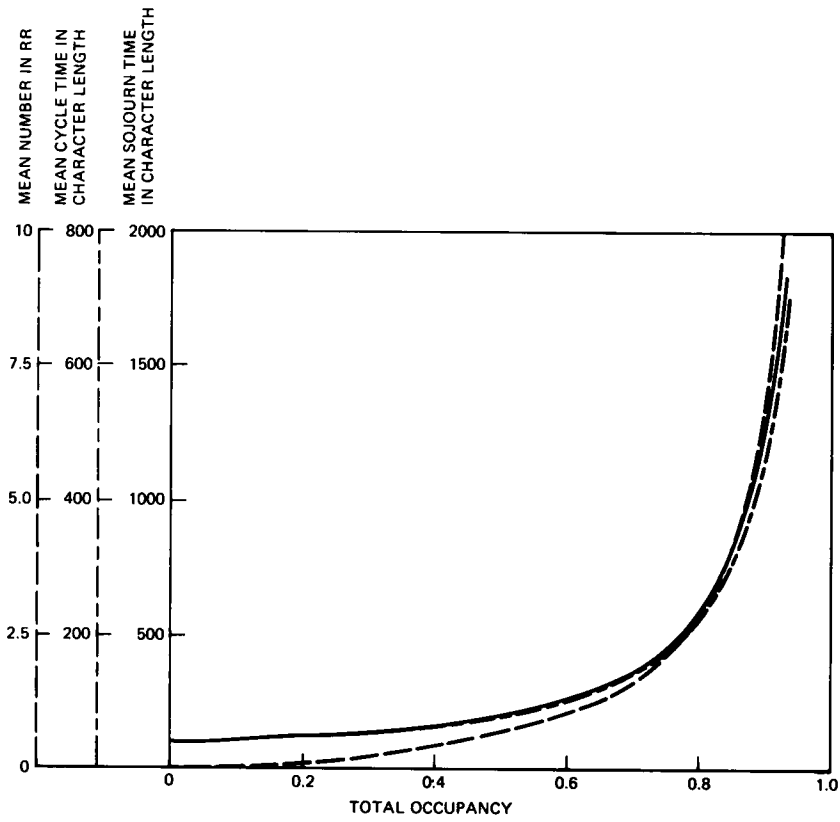

Fig. 5—Some RR performance curves for traffic mix $M4$.

a similar behavior ($\alpha\ \rho/(1 - \rho)$). The assumption of exponentially distributed message lengths for jobs in queue 2 certainly plays an important role in this behavior. However, it is heartening that the mean cycle time should depend upon the server occupancy alone and be insensitive to the mean message length.

VI. REMARK

In data communication systems providing virtual circuit service it is necessary to move virtual circuits rather than individual messages from the FIFO to the RR queue and vice-versa. That is, once Δ_1 characters are removed for a virtual circuit in the FIFO queue, it is moved to the RR queue. On successive turns Δ_2 characters are removed from this virtual circuit until there are no data to be transmitted on the virtual circuit. At that time the virtual circuit is moved back to the FIFO queue so that the first part of the next message is served

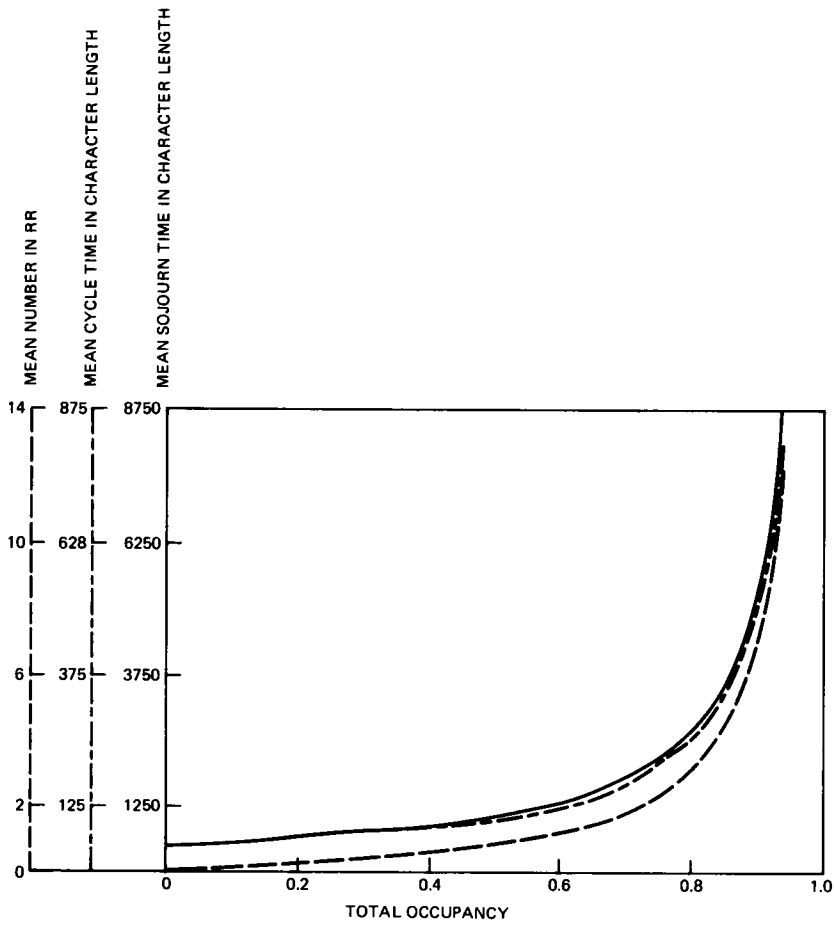

Fig. 6—Some RR performance curves for traffic mix $M5$.

from the FIFO queue. This, of course, implies that a short message immediately following a long message may see considerably longer delay than predicted by our analysis. This is unavoidable but not likely to happen often in practice.

Next, the access lines bringing data to the node that serves the link under consideration may be running slower than the 56-kb/s link. In that case a long message will be seen as a number of short messages by the node using FIFO-RR discipline. This will tend to increase the delay for the short messages. The effects of slower access lines are studied in Refs. 9 and 11. Also, under a heavy load, flow control may force a long message to be transmitted as a number of shorter messages, thus increasing the utilization in the FIFO queue. The effect is

similar to that of the slower access lines. Essentially, this breaking up of long messages allows the same message to reappear in the FIFO queue every so often. This could be discouraged by forcing each virtual circuit to pass through the RR queue for at least one cycle after every service in the FIFO queue. Only when a virtual circuit in the RR queue is found empty, it is moved back to the FIFO queue. Genuinely short messages could be exempted from this requirement by making the decision to move the virtual circuit from the FIFO to the RR queue depend on whether Δ_1, or fewer than Δ_1, characters were transmitted.

VII. ACKNOWLEDGMENT

The authors would like to thank S. Morgan for helpful suggestions on an earlier draft and for providing simulation results in Fig. 4.

REFERENCES

1. R. W. Wolff, "Time Sharing With Priorities," SIAM, J. Appl. Math., *19*, No. 3 (November 1970), pp. 566–74.
2. L. E. Schrage, "The Queue M/G/1 With Feedback to Lower Priority Queues," Manage. Sci., *13* (May–June 1967), pp. 466–74.
3. A. G. Fraser and S. P. Morgan, "Queueing and Framing Disciplines for a Mixture of Data Traffic Types," AT&T Bell. Lab. Tech. J., *63*, No. 6, Part 2 (July–August 1984), pp. 1061–87.
4. V. Ramaswami, "Explicit Matrix Geometric Solutions for a Class of Markov Processes," private communication.
5. P. H. Brill and M. J. M. Posner, "Level Crossings in Point Processes Applied to Queues: Single Server Case," Oper. Res., *25*, No. 4 (July–August 1977), pp. 662–73.
6. B. T. Doshi, "An M/G/1 Queue With a Hybrid Discipline," B.S.T.J., *62*, No. 5 (May–June 1983), pp. 1251–71.
7. L. Kleinrock, *Queueing Systems, Vol. II*, New York: Wiley, 1975.
8. D. L. Jagerman, "An Inversion Technique for the Laplace Transform With Application to Approximation," B.S.T.J., *57*, No. 3 (March 1978), pp. 669–710.
9. R. W. Wolff, "Poisson Arrivals See Time Averages," Oper. Res., *30*, No. 2 (March–April 1982), pp. 223–31.
10. A. G. Fraser and S. P. Morgan, unpublished work.
11. H. Rudin, Jr., "Buffered Packet Switching: A Queue With Clustered Arrivals," Int. Switching Symp. Rec., MIT, 1972, pp. 259–65.

APPENDIX

Analysis of the Busy Periods

We now derive expressions for the joint transforms of $\beta_1(x, k)$ and $\beta_2(x, k)$ and the Laplace-Stieltjes transform of the system-busy period.

Recall that

$$\beta_i(s, z) = E[e^{-sb_i}z^K]$$

$$= \int_{0-}^{\infty} \sum_{k=0}^{\infty} e^{-sx}z^k dB_i(x, k), \qquad i = 1, 2, \tag{61}$$

where b_i denotes the length of an i-busy-period and K the number of jobs moving to the back of the RR queue during this busy period.

Let X denote the total service time of a job, X_1 the portion that gets served in the FIFO queue, and N_1 the number of new arrivals during the time X_1. Then

$$\beta_1(s, z) = E[E[e^{-sb_1}z^K \mid X, N_1]], \tag{62}$$

where

$$E[e^{-sb_1}z^K \mid X, N_1] = \begin{cases} e^{-sX}\beta_1^{N_1}(s, z) & 0 \le X \le \Delta_1 \\ e^{-s\Delta_1}z\beta_1^{N_1}(s, z) & X > \Delta_1. \end{cases} \tag{63}$$

Thus

$$\begin{aligned}\beta_1(s, z) &= p\int_0^{\Delta_1^+} e^{-sx}e^{-\lambda x[1-\beta_1(s,z)]}dH_1(x) \\ &\quad + (1-p)\int_0^{\Delta_1} \mu e^{-\mu x}e^{-sx}e^{-\lambda x[1-\beta_1(s,z)]}dx \\ &\quad + (1-p)\int_{\Delta_1}^{\infty} e^{-s\Delta_1}ze^{-\lambda\Delta_1[1-\beta_1(s,z)]}\mu e^{-\mu x}dx \\ &= p\tilde{h}_1[s + \lambda(1-\beta_1(s, z))] + (1-p) \\ &\quad \cdot\frac{\mu}{\mu + s + \lambda(1-\beta_1(s, z))}\left[1 - e^{-\Delta_1(\mu+s+\lambda(1-\beta_1(s,z)))}\right] \\ &\quad + (1-p)ze^{-(\mu+s+\lambda(1-\beta_1(s,z)))\Delta_1} \\ &= \tilde{f}_1(s + \lambda(1-\beta_1(s, z))) \\ &\quad + (1-p)(z-1)e^{\Delta_1(\mu+s+\lambda(1-\beta_1(s,z)))},\end{aligned} \tag{64}$$

where $\tilde{f}_1$ is as in Section IV.

Similarly, let X_2 denote the length of a typical service in the RR queue. Then

$$\begin{aligned}E[e^{-sb_2}z^K \mid X_2] &= e^{-sX_2}e^{-\lambda X_2(1-\beta_1(s,z))}, \qquad 0 \le X_2 < \Delta_2, \\ &= e^{-sX_2}ze^{-\lambda X_2(1-\beta_1(s,z))}, \qquad X_2 = \Delta_2.\end{aligned} \tag{65}$$

Thus

$$\beta_2(s, z) = E[E[e^{-sb_2}z^K \mid X_2]]$$

$$= \int_0^{\Delta_2} \mu e^{-\mu x} e^{-x(s+\lambda(1-\beta_1(s,z)))} dx$$

$$+ ze^{-\Delta_2(\mu+s+\lambda(1-\beta_1(s,z)))}$$

$$= \frac{\mu}{\mu + s + \lambda(1 - \beta_1(s, z))} (1 - e^{-\Delta_2(\mu+s+\lambda(1-\beta_1(s,z)))})$$

$$+ ze^{-\Delta_2(\mu+s+\lambda(1-\beta_1(s,z)))}$$

$$= \tilde{f}_2(s + \lambda(1 - \beta_1(s, z))) + (z - 1)e^{-\Delta_2(\mu+s+\lambda(1-\beta_1(s,z)))}. \quad (66)$$

Finally, since the system is work conserving, the system-busy period is the same as that in an ordinary FIFO system. Thus, its Laplace-Stieltjes transform β is given by

$$\beta(s) = \tilde{h}[s + \lambda(1 - \beta(s))], \quad (67)$$

where

$$\tilde{h}(s) = p\tilde{h}_1(s) + (1 - p)\frac{\mu}{\mu + s}. \quad (68)$$

In the presence of "chunk overheads," analytic expressions for the busy-period transforms can be obtained by making proper substitutions for $\tilde{h}(\cdot)$, $\tilde{f}_1(\cdot)$, $\tilde{f}_2(\cdot)$, etc., in eqs. (64), (66), and (67).

AUTHORS

Bharat T. Doshi, B. Tech. (Mechanical Engineering), 1970, I.I.T. Bombay; Ph.D. (Operations Research), 1974, Cornell University; AT&T Bell Laboratories, 1979—. Before joining AT&T Bell Laboratories, Mr. Doshi was an assistant Professor at Rutgers University. At AT&T Bell Laboratories his technical work includes modeling and analysis of processor schedules, communication network performance, and overload control. His research interests include queueing and scheduling theory applied to performance analysis of computer, communication, and production systems. Member, IEEE, ORSA; Associate Editor, OR Letters.

Kiran M. Rege, B. Tech. (Electrical Engineering), 1977, I.I.T., Bombay; Ph.D. (Electrical Engineering), 1981, University of Hawaii; AT&T Bell Laboratories, 1982—. Mr. Rege spent 1984 on leave of absence, teaching at I.I.T. Bombay in the Department of Electrical Engineering. His technical work at AT&T Bell Laboratories includes modeling and analysis of switching, computer, and communication systems. His research interests include communication theory, queueing theory, and performance analysis of computer and communication systems.