Copyright © 1983 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 5, May–June 1983
Printed in U.S.A.

An M/G/1 Queue With a Hybrid Discipline

By B. T. DOSHI

(Manuscript received October 27, 1982)

In this paper we analyze the delay in a single-server queue in which the server, when it becomes free, selects for the next service the oldest customer with current delay smaller than T. If no such customer is present, then it selects the youngest customer with the current delay in excess of T. This service discipline is desirable in applications where the success or failure of a service depends on the delay in providing the service. Telephone call processing and steel rolling are two of these applications. We obtain the delay distribution for this service discipline using a combination of level-crossing arguments and renewal theory, and compare this performance with that of the last-in-first-out discipline with respect to the throughput of successfully served customers.

I. INTRODUCTION

The following situation is common in telephone call processing or data-processing systems. When a customer requests service, an entry is made in the queue that is serviced by a processor. The processor serves the queue of entries according to some specified service discipline. When an entry is served, the corresponding customer is notified of the completion of the service. The customer, however, does not wait forever for the completion of the service. At some random time, R, after its arrival, the customer will renege if service is not completed. The associated entry remains in the queue, and the server does not know that the customer has reneged until after it completes the service and attempts to notify the customer. Such a service is wasted. The customer may make the situation worse by reattempts at getting the desired service, thereby increasing the load on the server. It is therefore necessary to keep the proportion of reneging customers as small as possible. This can be done by selecting an appropriate discipline.

In this paper we study a specific model of the above situation. In

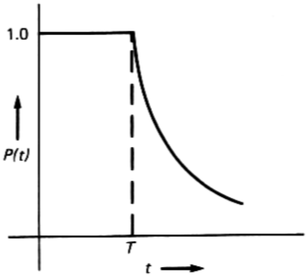

Fig. 1—Customer behavior function 1.

particular, we have a single-server queue with Poisson arrivals at rate $\lambda > 0$.

Let G be the distribution of the service time. Let $\bar{P}(t)$ be the probability that the customer does not renege before time t. We may think of $\bar{P}(t)$ as the expected reward obtained by completing the service of a customer at t time units after its arrival. With this general interpretation $\bar{P}(t)$ need not be restricted to be between 0 and 1. Let

$$P(t) = \int_{0^-}^{\infty} \bar{P}(t + y)dG(y) \tag{1}$$

for $0 \le t < \infty$. Then $P(t)$ is the expected reward from a customer whose *waiting time* (excluding the service time) is t. Let W_π be the distribution of the waiting time under the specified service discipline, π. Then the expected reward from an arbitrary customer is

$$V_\pi = \int_{0^-}^{\infty} P(t)dW_\pi(t). \tag{2}$$

We want to select a service discipline that maximizes V_π. It was shown in Doshi and Lipper[1] that if $P(t)$ is convex (respectively, concave), then the last-in-first-out (LIFO) [respectively, first-in-first-out (FIFO)] discipline is optimal. More realistic functions $P(t)$, however, are of the forms given in Figs. 1 and 2.

For such functions, $P(t)$, an optimal service discipline, is not known. However, our results for concave and convex $P(t)$ indicate that a hybrid discipline may provide better performance than either the FIFO or the LIFO discipline does. In this hybrid discipline the server, when it completes a service, first looks at the customers with the current waiting time less than T and selects the oldest waiting customer for the next service. If no such customer is waiting, then the server looks at the customers with the current waiting time in excess of T and selects the youngest customer. Note that this hybrid discipline includes FIFO ($T = \infty$) and LIFO ($T = 0$) as special cases. Since $P(t)$

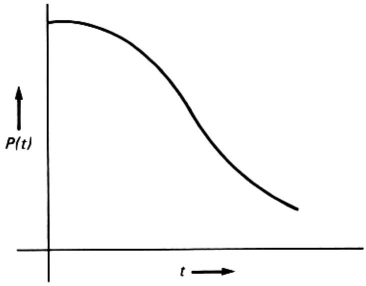

Fig. 2—Customer behavior function 2.

is assumed to be given, we only need $W_\pi(t)$, $0 \leq t < \infty$, for this hybrid discipline π, to calculate V_π. In this paper we obtain W_π for general service time distribution. We do this in three steps. First we describe another queueing system for which the distribution of the waiting time is the same as for the original system. Moreover, for this equivalent queueing system the distribution of the waiting time can easily be expressed in terms of the distribution, $F(x)$, of the work in a subsystem. We then use level-crossing arguments to derive an integral equation satisfied by $f(x) = F'(x)$. Finally, we use some results from renewal theory to solve this integral equation.

Some comments about the model are in order before we proceed to give an outline of the rest of this paper. Models similar to the one studied here can be useful in a variety of other applications. Some of these are the management of steel-rolling operation and the management of blood bank. Also, in many applications the customers do not necessarily renege. They simply take actions (start to dial, become cold, etc.) which make any subsequent service worthless.

This paper is organized as follows: In Section II we formally define the queueing system under consideration. We describe an equivalent queueing system in Section III. There we also show the relationship between the distributions of the waiting time in the original system and of the work in a subsystem of the equivalent queueing system. In Section IV we derive an integral equation for the steady-state density of the work in the subsystem. We give the solution of this integral equation in Section V. There we also derive the steady-state distribution of the waiting time in the original system. Finally, we give some numerical results in Section VI.

II. MODEL

The queueing system and the hybrid service discipline discussed in Section I can be formally described as follows: We have a queueing system with a single server and two queues, Q1 and Q2. Customers in

Q1 have a nonpreemptive priority over those in Q2. The order of service is first-in-first-out (FIFO) in Q1 and last-in-first-out (LIFO) in Q2. Customers arrive according to a Poisson process at rate $\lambda > 0$. On arrival the customer is put in Q1. If its service has not started within T seconds after its arrival, then at that time it is transferred to Q2. The service times of the customers are independent and identically distributed with distribution function G with continuous density g.

Let μ_K denote the Kth moment of the service time and let $\rho = \lambda\mu_1$. Assume that the waiting-time process is in the steady state. Let W denote the distribution function of the waiting time seen by an arbitrary customer. Since G has a continuous density, W is differentiable on $(0, \infty)$. Let

$$w(x) = W'(x) \qquad 0 < x < \infty.$$

We are interested in obtaining an expression for $W(x)$ or, equivalently, for $W(0)$ and $w(x)$, $0 < x < \infty$.

III. AN EQUIVALENT QUEUEING SYSTEM

We now describe a queueing system that is equivalent to the one described in Section II as far as the waiting times of the customers are concerned. However, the number of customers in Q1 and Q2 at a given time may be different in the two systems.

Consider the subsystem consisting of the server and Q1. Let X_t denote the work, at time t, in this subsystem. Thus, X_t is the sum of the remaining service time of the customer, if any, being served and the service times of all the customers in Q1. If a customer arriving at time t finds $X_t \leq T$, then it joins Q1; otherwise it joins Q2. Recall that Q1 has a nonpreemptive priority over Q2 and that Q1 is served FIFO and Q2 is served LIFO. A little reflection shows that the waiting time of a customer is the same in this system as in the one described in Section II.

We now relate the waiting-time distribution, W, to the steady-state distribution, F, of the work, X, in the subsystem consisting of the server and Q1. If an arriving customer sees $X \leq T$, then its waiting time will be X because the service in Q1 is FIFO and because Q1 has a nonpreemptive priority over Q2. Thus,

$$W(x) = F(x) \qquad 0 \leq x \leq T. \tag{3}$$

In particular,

$$W(0) = F(0), \tag{4}$$

and

$$w(x) = f(x) = F'(x) \qquad 0 < x < T. \tag{5}$$

If an arriving customer sees $X > T$, then it joins Q2. Since Q1 has priority over Q2 and since Q2 is served LIFO, this customer has to wait for the duration of a busy period in the M/G/1 queue with initial work, X. Let B_y denote the distribution of the busy period in the M/G/1 queue started by initial work, y. Since G has a continuous density,

$$b_y(x) = B'_y(x) \tag{6}$$

exists for all $x > y$. Then

$$w(x) = \int_T^x f(y) b_y(x) d(y) \qquad x > T. \tag{7}$$

Thus, it is sufficient to find the distribution of X.

IV. INTEGRAL EQUATION FOR $f(x)$

We now use level-crossing arguments (see Ref. 2) to derive an integral equation for f. Figure 3 shows a typical sample function for the process $\{X_t\}$. Assume that at time 0 the queues are empty, the server is idle, and a customer arrives and enters service. If $X_t > 0$, then it decreases at unit rate as in a M/G/1 queue. Symbols ○ represent arrivals which see $X_t \leq T$ and join the subsystem, thus increasing X_t by a service time. A symbol in the shape of a dot (●) represents arrivals that see $X_t > T$, and join Q2 without affecting X_t on their arrival. When X_t reaches zero, two things can happen: Q2 is empty and the server remains idle until the next arrival, or Q2 is nonempty and a customer from Q2 enters service, thus increasing X_t by its service time. Such arrivals from Q2 into the server are denoted by a symbol in the shape of a square (■).

The stochastic process $\{X_t\}$ is not Markovian because what happens

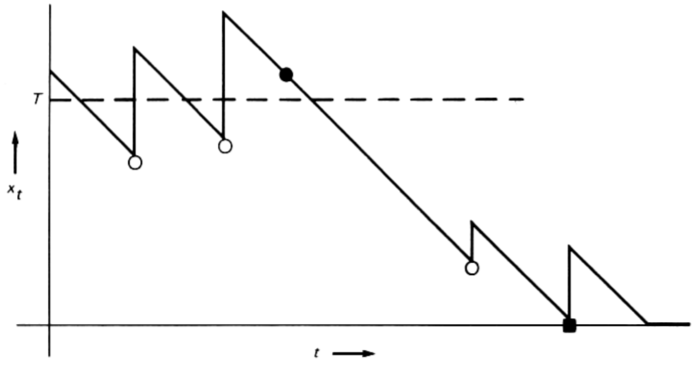

Fig. 3—Typical sample function for the process $\{X_t\}$.

when X_t reaches zero depends on the past. On the other hand, the vector-valued process $\{(X_t, N_t)\}$, where N_t is the number in Q2 at time t, is a Markov process. We can derive the steady-state distribution of (X_t, N_t) using the standard results and use that to obtain the marginal steady-state distribution of X. We, however, use a simpler approach here. Recall that $\mu_2 < \infty$. Consider the following two cases:

(*i*) $\rho < 1$. In this case the process $\{X_t\}$ is regenerative and with the regeneration points corresponding to the external arrivals that make an idle server busy. Denote this event by E. Then E is a positive recurrent, regenerative event.

(*ii*) $\rho \geq 1$. In this case the queue length in Q2 grows without bound and, in the steady state, can be assumed to be ∞. Thus, a customer is removed from Q2 to enter service every time X_t reaches 0. This event, E′, is then positive recurrent.

Standard regenerative arguments now show that $\{X_t\}$ has a steady-state distribution and that

$$X_t \xrightarrow{D} X,$$

where the distribution of X is the steady-state distribution of $\{X_t\}$. Moreover, for any x, $0 < x < \infty$, the steady-state rate, $D(x)$, at which X_t crosses x from above, equals the rate $U(x)$, at which X_t jumps from below x to above x. We now express $D(x)$ and $U(x)$ in terms of $f(x)$ and get the desired integral equation by equating these expressions.

X_t decreases at unit rate until an arrival occurs or until $X_t = 0$. Thus, during every downcrossing of level x, the $\{X_t\}$ process spends dx units of time in the interval $(x - dx, x)$. Hence,

$$D(x) = f(x) \qquad 0 < x < \infty. \tag{8}$$

Before deriving an expression for $U(x)$ we introduce some notation. Let p denote the rate at which X_t jumps from 0 to some positive value. These jumps may be due to either external arrivals coming to an idle system or to customers from Q2 moving to the server. Also, let $\bar{G}$ denote the complementary service time distribution defined by

$$\bar{G}(x) = 1 - G(x) \qquad 0 \leq x < \infty. \tag{9}$$

Assume that $G(0) = 0$, $\bar{G}(0) = 1$. Then, for $x \leq T$

$$U(x) = \lambda \int_0^x f(y)\bar{G}(x - y)dy + p\bar{G}(x). \tag{10}$$

Since an external arrival causes a jump in X_t only when $X_t \leq T$, we have, for $x > T$,

$$U(x) = \lambda \int_0^T f(y)\bar{G}(x - y)dy + p\bar{G}(x). \tag{11}$$

Thus,

$$U(x) = \lambda \int_0^{T\wedge x} f(y)\bar{G}(x-y)dy + p\bar{G}(x) \qquad 0 < x < \infty. \tag{12}$$

Let

$$f(0) = \lim_{x\downarrow 0} f(x). \tag{13}$$

Then $f(0)$ is the rate at which $\{X_t\}$ hits level zero. In the steady state, this must equal the rate at which $\{X_t\}$ jumps from 0 to some positive value. Thus

$$p = f(0). \tag{14}$$

We now have for $0 < \rho < \infty$, $0 < x < \infty$,

$$\begin{aligned} f(x) &= D(x) \\ &= U(x) \\ &= \lambda \int_0^{x\wedge T} f(y)\bar{G}(x-y)dy + f(0)\bar{G}(x). \end{aligned} \tag{15}$$

Let

$$q = \frac{f(0)}{\lambda}.$$

Then

$$f(x) = \lambda \int_0^{x\wedge T} f(y)\bar{G}(x-y)dy + q\lambda\bar{G}(x). \tag{16}$$

This is the desired integral equation for f. The additional conditions needed to solve this completely depend on whether $\rho < 1$ or $\rho \geq 1$.

First consider the case where $\rho < 1$. Let P_2 be the probability that an arriving customer sees $X_t > T$ and joins Q2. Since, for $\rho < 1$, every arriving customer is eventually served, the rate at which customers enter the server from Q2 is λP_2. Also, the rate of arrivals coming to an empty system is $\lambda F(0)$. Thus,

$$q = \frac{f(0)}{\lambda} = \frac{\lambda[F(0) + P_2]}{\lambda} = F(0) + P_2. \tag{17}$$

Also,

$$P_2 = \int_T^\infty f(x)dx, \tag{18}$$

and

$$F(0) + \int_0^\infty f(x)dx = 1. \tag{19}$$

For $\rho \geq 1$, Q2 is always nonempty in steady state. Hence $F(0) = 0$,

$$q = \frac{f(0)}{\lambda}, \tag{20}$$

and

$$\int_0^\infty f(x)dx = 1. \tag{21}$$

Equation (16), together with either conditions [eqs. (17) to (19) or 20], characterizes f completely. We solve this equation in the next section.

V. SOLUTION OF THE INTEGRAL EQUATION

We now solve eq. (15) to obtain an expression for $f(x)$. For $0 < x < \infty$, let

$$h(x) = \lambda\bar{G}(x) \tag{22}$$

and let $m(x)$ be the renewal density function for $h(x)$. Then $m(x)$ satisfies (see Ref. 3):

$$m(x) = h(x) + \int_0^x h(x - y)m(y)dy. \tag{23}$$

Equation (15) can now be rewritten as

$$f(x) = qh(x) + \int_0^x f(x - y)h(y)dy \qquad (0 < x < T), \tag{24}$$

and

$$f(x) = qh(x) + \int_0^T f(y)h(x - y)dy \qquad T \leq x < \infty. \tag{25}$$

Equation (24) is a renewal equation and its solution is given by

$$\begin{aligned} f(x) &= qh(x) + q\int_0^x h(x - y)m(y)dy \\ &= q\left[h(x) + \int_0^x h(x - y)m(y)dy\right] \\ &= qm(x), \end{aligned} \tag{26}$$

where the last equality follows from (23). Note that $0 < h(x) < \infty$, and

$$\int_0^\infty h(x)dx = \lambda\mu_1 = \rho.$$

Thus, h is like a probability density with total mass, ρ. The function $m(x)$ is well defined for any finite x irrespective of the value of ρ, $0 < \rho < \infty$. To obtain an expression for $f(x)$, $x \geq T$, we note that the right-hand side of (25) involves $f(y)$ for y only in the interval $(0, T)$, which we have obtained in (26). Thus, replacing $f(y)$ on the right-hand side of (25) by $qm(y)$, we get

$$\begin{aligned} f(x) &= \left[qh(x) + \int_0^T qm(y)h(x-y)dy \right] \\ &= q \left[h(x) + \int_0^T h(x-y)m(y)dy \right] \qquad x \geq T. \end{aligned} \tag{27}$$

We now use conditions (17) through (19) or (20) through (21) to evaluate q and thus completely characterize f. First, consider the case $\rho < 1$. We have

$$q = F(0) + P_2, \tag{28}$$

and

$$F(0) + q\int_0^T m(x)dx + P_2 = 1. \tag{29}$$

Also, equating the rate of customers coming to the system with the rate of customers leaving the system, we get

$$\lambda = \frac{1}{\mu_1}[1 - F(0)],$$

or

$$F(0) = 1 - \rho. \tag{30}$$

We can now solve (28) through (30) for q and P_2 to get

$$P_2 = \frac{\rho - (1-\rho)\displaystyle\int_0^T m(x)dx}{1 + \displaystyle\int_0^T m(x)dx},$$

and

$$q = F(0) + P_2 = \frac{1}{1 + \int_0^T m(x)dx}.$$

Thus,

$$F(0) = 1 - \rho \tag{31}$$

$$f(x) = \frac{m(x)}{1 + \int_0^T m(x)dx} \qquad x < T \tag{32}$$

$$= \frac{h(x) + \int_{y=0}^T h(x-y)m(y)dy}{1 + \int_0^T m(x)dx} \qquad x \geq T. \tag{33}$$

Next, consider the case where $\rho \geq 1$. Here,

$$q = \frac{f(0)}{\lambda},$$

and

$$\int_0^T f(x)dx + \int_T^\infty f(x)dx = 1.$$

Thus,

$$f(0)\left\{\int_0^T m(x)dx + \rho\left[1 + \int_0^T m(x)dx\right] - \int_0^T m(x)dx\right\} = 1,$$

or

$$f(0) = \frac{1}{\rho\left[1 + \int_0^T m(x)dx\right]}, \tag{34}$$

$$f(x) = \frac{m(x)}{\rho\left[1 + \int_0^T m(x)dx\right]} \qquad x < T, \tag{35}$$

and

$$f(x) = \frac{h(x) + \displaystyle\int_0^T h(x-y)m(y)dy}{\rho\left[1 + \displaystyle\int_0^T m(x)dx\right]} \qquad x \geq T. \tag{36}$$

We now consider a special case where $\bar{G}(x) = \mathrm{e}^{-\mu x}$, $\mu = 1/\mu_1$. Then,

$$h(x) = \lambda e^{-\mu x} \qquad 0 < x < \infty,$$

$$m(x) = \lambda e^{(\lambda-\mu)x} \qquad 0 < x < \infty,$$

$$1 + \int_0^T m(x)dx = \frac{1}{(1-\rho)}[1 - \rho e^{-\mu(1-\rho)T}],$$

and, for $x \geq T$,

$$\begin{aligned} h(x) &+ \int_0^T h(x-y)m(y)dy \\ &= \lambda e^{-\mu x} + \lambda^2 \int_0^T e^{-\mu(x-y)} e^{-(\mu-\lambda)y} dy \\ &= \lambda e^{-\mu x} + \lambda^2 e^{-\mu x} \int_0^T e^{+\lambda y} dy \\ &= \lambda e^{-\mu x + \lambda T}. \end{aligned}$$

Thus, for $\rho < 1$

$$F(0) = 1 - \rho$$

$$f(x) = \frac{1-\rho}{1-\rho e^{-\mu(1-\rho)T}} e^{-\mu(1-\rho)x} \qquad x < T,$$

and

$$f(x) = \frac{1-\rho}{1-\rho e^{-\mu(1-\rho)T}} e^{-\mu x + \lambda T} \qquad x \geq T.$$

For $\rho \geq 1$

$$F(0) = 0,$$

$$f(x) = \frac{1-\rho}{\rho[1-\rho e^{\mu(\rho-1)T}]} e^{\mu(\rho-1)x} \qquad x < T,$$

and

$$f(x) = \frac{1-\rho}{\rho[1-\rho e^{\mu(\rho-1)T}]} e^{-\mu x+\lambda T} \qquad x \geq T.$$

VI. WAITING-TIME DISTRIBUTION

We now use the results of Section V to obtain an expression for the waiting-time distribution and its Laplace Stieltjes Transform. From eqs. (4), (5) and (6) we get

$$W(0) = F(0) = \begin{cases} 1-\rho & \rho < 1 \\ 0 & \rho \geq 1, \end{cases} \tag{37}$$

$$w(x) = f(x) = \begin{cases} \dfrac{m(x)}{\left[1 + \displaystyle\int_0^T m(x)dx\right]} & \rho < 1 \\[2ex] \dfrac{m(x)}{\rho\left[1 + \displaystyle\int_0^T m(x)dx\right]} & \rho \geq 1, \end{cases} \tag{38}$$

and

$$w(x) = \begin{cases} \dfrac{\displaystyle\int_T^x \left[h(y) + \int_0^T h(y-z)m(z)dz\right] b_y(x)dy}{1 + \displaystyle\int_0^T m(x)dx} & \rho < 1 \\[3ex] \dfrac{\displaystyle\int_T^x \left[h(y) + \int_0^y h(y-z)m(z)dz\right] b_y(x)dy}{\rho\left[1 + \displaystyle\int_0^T m(x)dx\right]} & \rho \geq 1. \end{cases} \tag{39}$$

Let

$$W^*(\theta) = E[e^{-\theta W}] = W(0) + \int_0^\infty e^{-\theta x} w(x)dx \qquad \text{Re } \theta > 0. \tag{40}$$

Also, for Re $\theta > 0$ let

$$g^*(\theta) = \int_0^\infty e^{-\theta x} g(x)dx \tag{41}$$

$$h^*(\theta) = \int_0^\infty e^{-\theta x} h(x)dx$$
$$= \frac{\lambda[1 - g^*(\theta)]}{\theta}, \tag{42}$$

and

$$m^*(\theta) = \int_0^\infty e^{-\theta x} m(x)dx$$
$$= \frac{h^*(\theta)}{1 - h^*(\theta)}$$
$$= \frac{\lambda[1 - g^*(\theta)]}{\theta - \lambda[1 - g^*(\theta)]}. \tag{43}$$

Let $B^*(\theta)$ be the Laplace Stieltjes Transform of the ordinary busy period in an M/G/1 queue with the arrival rate $\lambda > 0$ and the service distribution, G. Then B^* satisfies

$$B^*(\theta) = g^*\{\theta + \lambda[1 - B^*(\theta)]\}. \tag{44}$$

Also, let $B_y^*(\theta)$ be the Laplace Stieltjes Transform of B_y. Then

$$B_y^*(\theta) = e^{-y\{\theta+\lambda[1-B^*(\theta)]\}}. \tag{45}$$

We now express $W^*(\theta)$ in terms of m, m^* and B^*. First, consider the case $\rho < 1$, where

$$W^*(\theta) = 1 - \rho + \frac{1}{1 + \int_0^T m(x)dx} \left\{ \int_0^T e^{-\theta x} m(x)dx \right.$$
$$\left. + \int_T^\infty e^{-\theta x} \int_{y=T}^x \left[h(y) + \int_0^T h(y-z)m(z)dz \right] b_y(x)dydx \right\}$$
$$= 1 - \rho + \frac{1}{1 + \int_0^T m(x)dx} \left(\int_0^T e^{-\theta x} m(x)dx \right.$$
$$\left. + \frac{\theta}{\theta + \lambda[1 - B^*(\theta)]} \int_T^\infty e^{-\{\theta+\lambda[1-B^*(\theta)]\}x} m(x)dx \right). \tag{46}$$

For $\rho > 1$,

$$W^*(\theta) = \frac{1}{\rho\left[1 + \int_0^T m(x)dx\right]}\left\{\int_0^T e^{-\theta x}m(x)dx\right.$$

$$\left. + \frac{\theta}{\theta + \lambda[1 - B^*(\theta)]}\int_T^\infty e^{-[\theta+\lambda(1-B^*(\theta))]x}m(x)dx\right\}. \qquad (47)$$

6.1 Mean value of the waiting time

Let $\bar{W}$ denote the mean value of the waiting time. For $\rho < 1$, every customer is eventually served. Hence, $\bar{W}$ is the average over all the customers. For $\rho > 1$, some of the arriving customers do not get served and, in this case, $\bar{W}$ is the average waiting time of those who do get served. In the first case,

$$\bar{W} = -W^{*\prime}(0^+),$$

and for the second case,

$$\bar{W} = \frac{-W^{*\prime}(0^+)}{W^*(0^+)}.$$

For $\rho < 1$, all customers are served and the order of service does not affect the mean waiting time. Thus, in this case the mean waiting time is the same as that for an M/G/1 queue with the FIFO discipline. That is,

$$\bar{W} = \frac{\lambda\mu_2}{2(1-\rho)}. \qquad (48)$$

For $\rho > 1$, the busy period distribution is defective. Let

$$b_o = P\{\text{Busy period} < \infty\}.$$

Then b_o is the unique solution in (0, 1) of

$$B^*(0^+) = b_o = g^*[\lambda(1 - b_o)]. \qquad (49)$$

Also,

$$B^{*\prime}(0^+) = \frac{g^{*\prime}[\lambda(1 - b_o)]}{1 + \lambda g^{*\prime}[\lambda(1 - b_o)]} \qquad (50)$$

and

$$B^{*\prime\prime}(0^+) = \frac{g^{*\prime\prime}[\lambda(1 - b_o)]}{\{1 + \lambda g^{*\prime}[\lambda(1 - b_o)]\}^3}. \qquad (51)$$

Now, from eq. (47), we get

$$W^*(0^+) = \frac{1}{\rho\left[1 + \int_0^T m(x)dx\right]}$$

$$\cdot\left[\int_0^T m(x)dx + \frac{1}{\lambda(1-b_o)}\lim_{\theta\to 0^+}\int_T^\infty \theta e^{-\{\theta+\lambda[1-B^*(\theta)]\}x} m(x)dx\right]$$

$$= \frac{1}{\rho\left[1 + \int_0^T m(x)dx\right]}\left[\int_0^T m(x)dx + 1\right]$$

$$= \frac{1}{\rho} \tag{52}$$

and

$$-W^{*\prime}(0^+) = \frac{1}{\rho\left[1 + \int_0^T m(x)dx\right]}$$

$$\cdot\left\{\int_0^T xm(x)dx + \frac{1}{\lambda(1-b_o)}\left[1 + \int_0^T m(x)dx\right]\right\}$$

$$= \frac{1}{\rho}\left[\frac{\int_0^T xm(x)dx}{1 + \int_0^T m(x)dx} + \frac{1}{\lambda(1-b_o)}\right]. \tag{53}$$

From (52) and (53) we have

$$\bar{W} = \frac{-W^{*\prime}(0^+)}{W^*(0^+)} = \frac{\int_0^T xm(x)dx}{1 + \int_0^T m(x)dx} + \frac{1}{\lambda(1-b_o)}. \tag{54}$$

Equation (54) shows that the mean waiting time of the customers who get served is minimized by the LIFO discipline ($T = 0$).

VII. NUMERICAL RESULTS

In this section we present some numerical results. Instead of calculating the waiting-time distribution, we calculate the quantity of interest, namely

$$V = \int_{0^-}^{\infty} P(t)dW(t)$$

for a specific P. Let

$$P(t) = \begin{cases} 1 & t \leq T \\ e^{-\sigma(t-T)} & t > T. \end{cases} \tag{55}$$

We now evaluate V for the hybrid discipline with parameter T and also for the LIFO discipline. For the hybrid discipline we have

$$\begin{aligned} V_T = q \cdot \Bigg[\int_0^T m(x)dx &+ e^{\sigma T} m^*\{\sigma + \lambda[1 - B^*(\sigma)]\} \\ &\cdot (1 + h^*\{\sigma + \lambda[1 - B^*(\sigma)]\}) \\ &- e^{\sigma T}(1 + h^*\{\sigma + \lambda[1 - B^*(\sigma)]\}) \\ &\cdot \int_0^T m(y)e^{-y(\sigma+\lambda)[1-B^*(\sigma)]}dy \Bigg], \end{aligned} \tag{56}$$

where

$$q = \begin{cases} \dfrac{1}{1 + \displaystyle\int_0^T m(x)dx} & \rho < 1 & (57) \\[2ex] \dfrac{1}{\rho\left[1 + \displaystyle\int_0^T m(x)dx\right]} & \rho > 1. & (58) \end{cases}$$

We need some more notation before writing an expression for V_L. Let $b_{\mathrm{FR}}(\cdot)$ denote the density function for the busy period started by the forward recurrence time of the service time. Thus,

$$b_{\mathrm{FR}}(x) = \int_0^x b_y(x) \frac{1 - G(y)}{\mu_1} dy. \tag{59}$$

Then, for $\rho < 1$,

$$\begin{aligned} V_L &= 1 - \rho + \rho \int_0^T b_{\mathrm{FR}}(x) + \rho e^{\sigma T} \int_T^{\infty} e^{-\sigma x} b_{\mathrm{FR}}(x)dx \\ &= 1 - \rho + \rho \left[\int_0^T b_{\mathrm{FR}}(x)dx - e^{\sigma T} \int_0^T e^{-\sigma x} b_{\mathrm{FR}}(x)dx \right] \\ &\quad + e^{\sigma T} \frac{\lambda[1 - B^*(\sigma)]}{\sigma + \lambda[1 - B^*(\sigma)]}, \end{aligned} \tag{60}$$

and, for $\rho > 1$,

$$V_L = \int_0^T b_{\mathrm{FR}}(x)dx - e^{\sigma T}\int_0^T e^{-\sigma x} b_{\mathrm{FR}}(x)dx + \frac{1 - B^*(\sigma)}{\mu_1\{\sigma + \lambda[1 - B^*(\sigma)]\}}. \tag{61}$$

For numerical calculations we considered all integrals in eqs. (56) through (58) and (60) through (61) as functions of T, obtained their Laplace transforms, and inverted the transform at the specified value of T using the inversion method of D. Jagerman.[3] Thus, let

$$R_1(t) = \int_0^t m(x)dx,$$

$$R_2(t) = \int_0^t m(x)e^{-x\{\sigma+\lambda[1-B^*(\sigma)]\}}dx,$$

$$R_3(t) = \int_0^t b_{\mathrm{FR}}(x)dx,$$

and

$$R_4(t) = \int_0^t e^{-\sigma x} b_{\mathrm{FR}}(x)dx.$$

Also, for $i = 1, 2, 3, 4$, and θ in the appropriate domain, let

$$R_i^*(\theta) = \int_0^\infty e^{-\theta t} R_i(t)dt.$$

Then

$$R_1^*(\theta) = \frac{m^*(\theta)}{\theta},$$

$$R_2^*(\theta) = \frac{m^*\{\theta + \sigma + \lambda[1 - B^*(\sigma)]\}}{\theta},$$

$$R_3^*(\theta) = \frac{1 - B^*(\theta)}{\theta\{\theta + \lambda[1 - B^*(\theta)]\}},$$

and

$$R_4^*(\theta) = \frac{1 - B^*(\theta + \sigma)}{\sigma + \theta + \lambda[1 - B^*(\theta + \sigma)]}.$$

For numerical examples we had the service time distribution gamma with mean 1 and variance $1/K$. We used two different values of K, $K = 1$ (exponential distribution), and $K = 10$. We used two different values of T, 1 and 3. Finally, we used two values of σ, 2.0 and 0.15. These give us eight parameter sets. The values of V_T and V_L as functions of the load $\lambda = \rho$ are given in Figs. 4 through 11. From these figures we observe:

(*i*) For both service disciplines, the throughput of good calls is larger for larger T, smaller σ, and larger K. The behavior of the throughput with respect to T and σ is obvious. Larger K implies smaller variability in the service time, thus reducing the probability of a customer getting served after a long wait. This, in turn, results in a higher throughput.

(*ii*) For the assumed customer behavior, the hybrid discipline always provides higher throughput than the LIFO discipline does. The difference is larger for larger T, larger σ, and larger K.

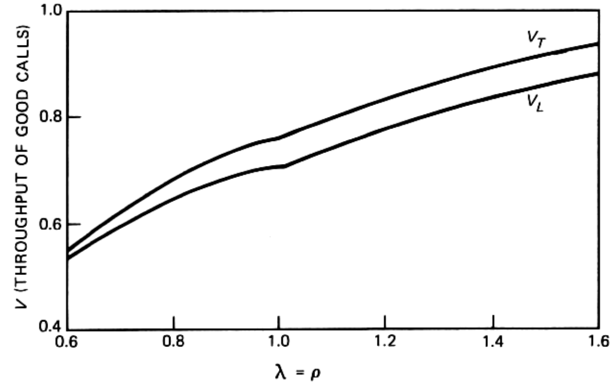

Fig. 4—The values V_T and V_L as functions of the load $\lambda = \rho$ for $T = 3.0$, $\sigma = 2.0$, and $K = 1$.

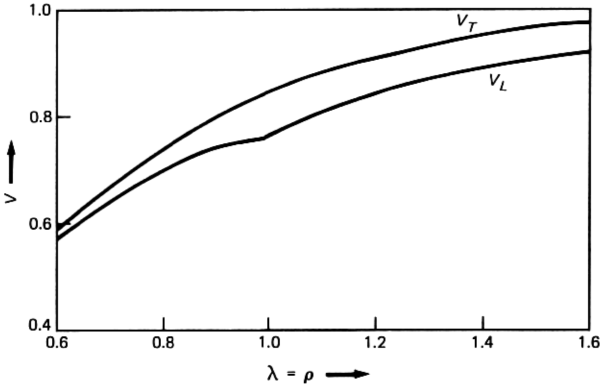

Fig. 5—The values V_T and V_L as functions of the load $\lambda = \rho$ for $T = 3.0$, $\sigma = 2.0$, and $K = 10$.

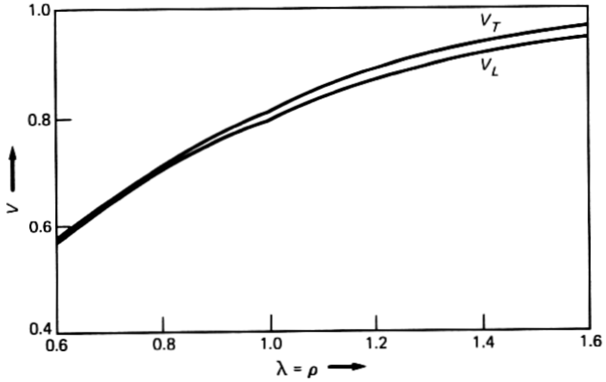

Fig. 6—The values V_T and V_L as functions of the load $\lambda = \rho$ for $T = 3.0$, $\sigma = 0.15$, and $K = 1$.

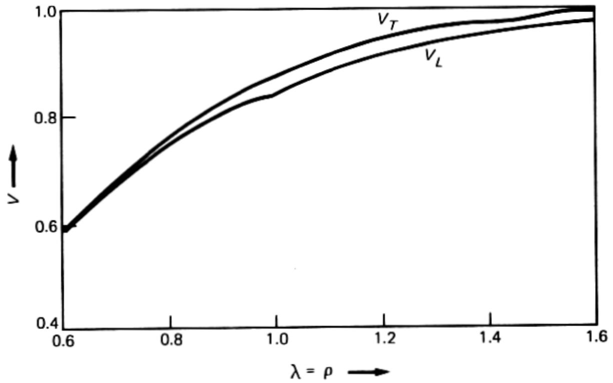

Fig. 7—The values V_T and V_L as functions of the load $\lambda = \rho$ for $T = 3.0$, $\sigma = 0.15$, and $K = 10$.

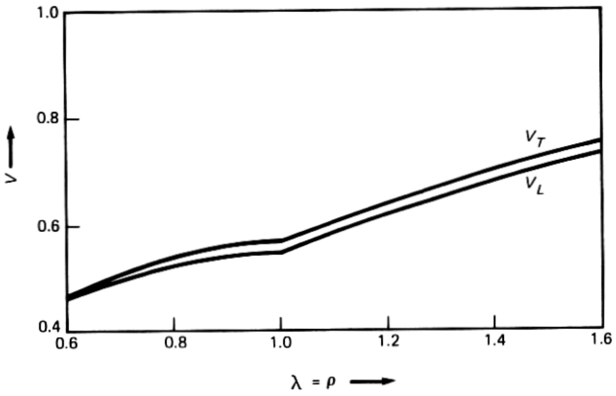

Fig. 8—The values V_T and V_L as functions of the load $\lambda = \rho$ for $T = 1.0$, $\sigma = 2.0$, and $K = 1$.

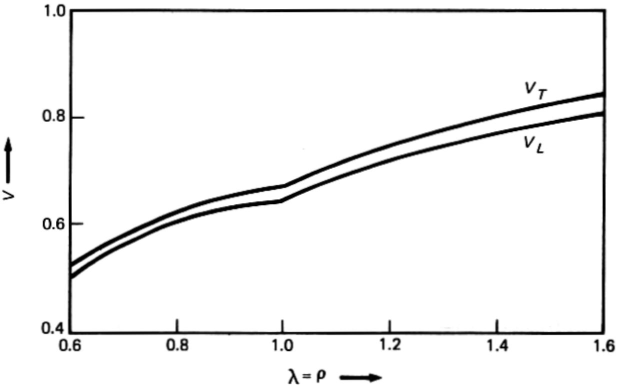

Fig. 9—The values V_T and V_L as functions of the load $\lambda = \rho$ for $T = 1.0$, $\sigma = 2.0$, and $K = 10$.

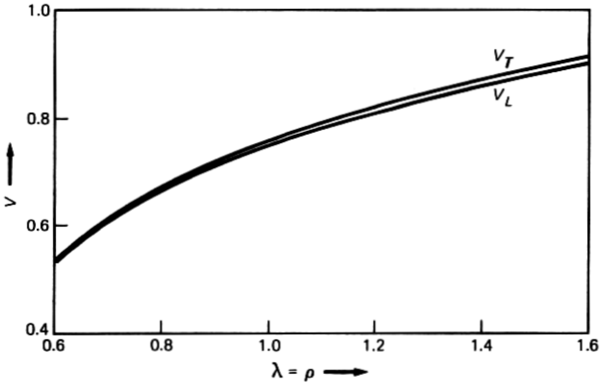

Fig. 10—The values V_T and V_L as functions of the load $\lambda = \rho$ for $T = 1.0$, $\sigma = 0.15$, and $K = 1$.

Of course, our knowledge of the customer behavior may be more or less accurate, depending on the application. An issue of interest then is the sensitivity of the throughput to the assumed customer behavior. This was studied for a special case ($K = 1$) in Ref. 1. The analysis in this paper can be used to answer such issues for more general service time distributions. Qualitatively, however, the conclusions will remain the same: the last-in-first-out (LIFO) discipline is robust with respect to the knowledge of customer behavior. The hybrid discipline, on the other hand, is very sensitive to the customer behavior and should be used only when the customer behavior is adequately known and does not change in time, or when the parameters of the customer behavior can be estimated and used to change the control parameters dynamically.

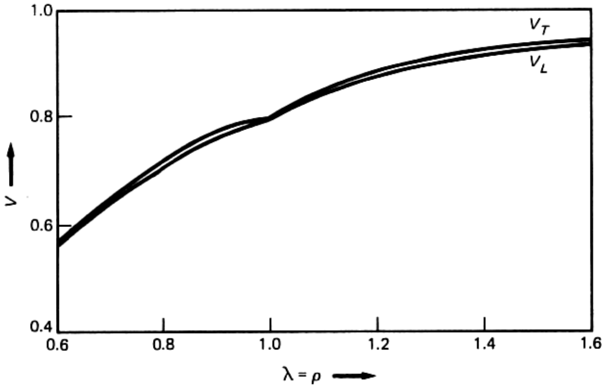

Fig. 11—The values V_T and V_L as functions of the load $\lambda = \rho$ for $T = 1.0$, $\sigma = 0.15$, and $K = 10$.

REFERENCES

1. B. T. Doshi and E. H. Lipper, "Comparisons of Service Disciplines in a Queueing System with Delay Dependent Customer Behavior," Proceedings of "Applied Probability, Computer Science, The Interface" Conf., *II,* Birkhauser, 1982, pp. 269–301.
2. P. H. Brill and M. J. M. Posner, "Level Crossings in Point Processes Applied to Queues: Single Server Case," Opns. Res., *25* (July 1977), pp. 662–74.
3. S. Karlin and H. Taylor, *A First Course in Stochastic Processes,* New York, NY: Academic Press, 1975, p. 183.
4. D. Jagerman, "An Inversion Technique for the Laplace Transform with Applications to Approximation," B.S.T.J., *61* (October 1982), pp. 1995–2002.