JOURNAL OF COMPUTER AND SYSTEM SCIENCES: 1, 287-298 (1967)

Distribution of Attained Service in Time-Shared Systems*

LEONARD KLEINROCK

Department of Engineering, University of California, Los Angeles, California 90024

AND

EDWARD G. COFFMAN

Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08540

Received March 13, 1967

ABSTRACT

A number of time-shared systems have recently been analyzed in the literature with methods from queueing theory. The quantity usually solved for is the average time spent in the system, conditioned on the total service time required (and also conditional on the priority class, if priority distinctions are considered). In this paper we consider a large class of time-shared systems and solve for the distribution of attained service for any member of this class. The attained service for an incompletely serviced customer is the number of seconds that he has so far spent in the service facility. The results are simply expressed in terms of the average conditional waiting time mentioned above. Examples of the application of this general result are also given.

I. INTRODUCTION

The analytic treatment of time-shared facilities is just beginning to appear in the literature with some regularity [*1–3*]. The usual approach taken is to model existing or proposed time-shared service facilities (the application generally being to computer facilities) as queueing systems. In these models, a user typically joins some queue, works his way up to the front of the queue, obtains service in the facility for some small amount of time (called a quantum), and then joins the same or some new queue to wait for more quanta if needed. The methods of queueing theory have been applied to a number of such models to obtain various measures of performance.

In this paper, we consider a large class of such "feedback" queueing systems and obtain, for all of these systems, a result which describes the distribution of attained service in terms of the previously solved performance measures.

The relationship derived herein between the usual measure of performance (i.e., average waiting time) and the distribution of attained service is extremely simple

* Preparation of this paper was sponsored in part by the Office of Naval Research, the U.S. Atomic Energy Commission, and the Advanced Research Project Agency. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

and quite general and therefore is interesting to pursue. Moreover, it provides one with an index of the composition of the system of queues. This index might find application in designing "optimum" time-shared systems when the criterion of optimality (or cost) is related to the expected number of customers in the queue with a given degree of completed service or with a given expectation of additional required service.

II. The Class of Feedback Queueing Systems

We include in our class any feedback queueing system with the following properties.

Consider Fig. 1. We assume that the population of new arrivals to the system are separated into P priority groups, this priority being determined by some external

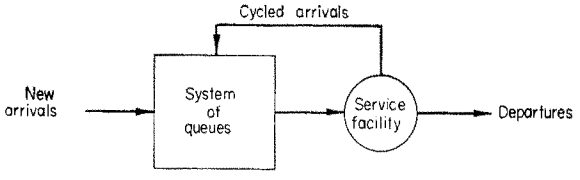

Fig. 1. Feedback queueing systems.

property[1] of the arrival (e.g., status in society, wealth, rank, size, memory space required); the assumption here is that the required time in the service facility (e.g., total computation time) is known only to within a probability distribution. Accordingly, let (for $p = 1, 2, \ldots, P$),

λ_p = average arrival rate of customers to the system from the pth priority group (customers/second)

$B_p(t)$ = Pr[customer from pth priority group requires a total processing time $\leqslant t$].

Upon arrival to the system, a new customer joins some queue in the system of queues. After some appropriate queueing discipline is followed, this customer will then be allowed into the service facility where, if he is from priority group p, he will be allotted a maximum of $g_{p1}Q$ seconds of service. If this quantum of time is greater than or equal to his total required service (t, say) he will then depart from the system as soon as he receives as much time as he needs; if $t > g_{p1}Q$, he will be cycled back to the system of queues where he joins some appropriate queue and waits for another quantum of service. On this nth visit to the service facility, this customer will receive a

[1] Although this priority is assigned from external considerations, the internal structure of these time-shared systems introduces certain implicit priority orderings with respect to progress of computation,

maximum of $g_{pn}Q$ seconds of service; thus if $Q\sum_{i=1}^{n-1} g_{pi} < t \leqslant Q\sum_{i=1}^{n} g_{pi}$ this customer departs from the system during his nth quantum of service. If $t > Q\sum_{i=1}^{n} g_{pi}$ he then recycles and continues.

Whenever the service facility ejects a customer (either for departure or re-cycling), some customer (if any are available) is taken into service. The particular customer chosen depends upon the specific discipline used in arranging customers within the system of queues. No customers are allowed to leave before they receive their total required service (i.e., no defecting).

As can be seen, a large number of time-sharing systems can be modeled by queueing disciplines which fall in this class of feedback queueing systems. For example, the round-robin systems studied in [*1*] and [*2*] are members of this class. The round-robin system consists of a single queue; each time a customer enters the system of queues (only one in this case) he must join the tail of the queue (see Fig. 2).

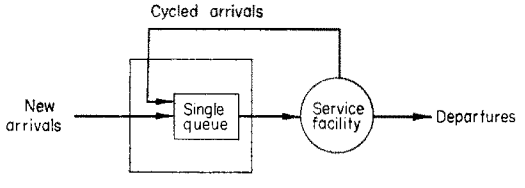

FIG. 2. The round-robin system.

The foreground-background system (also referred to as "feedback to lower priority queues" system [*3*]) is another example of a member of our class. In this system, a new arrival joins the first queue, obtains a quantum of service, and then, if more service is required, joins a second queue, etc., joining the nth queue on his nth visit to the system of queues. The server always gives service to the lowest numbered queue first and proceeds to the nth queue only if the n — 1st, n — 2nd,..., 2nd, 1st queues are all empty (see Fig. 3).

The single most significant performance measure of any queueing system is the average time that a customer spends waiting in queues as he passes through the system. For our feedback queueing systems, we are interested in this average waiting

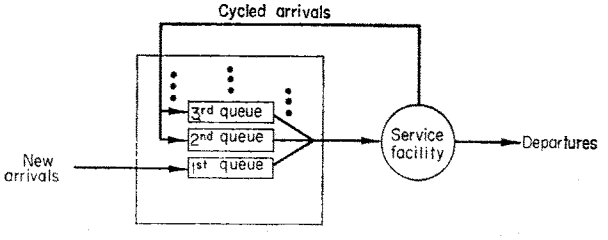

FIG. 3. The foreground-background system.

time, conditioned on a customer's priority group, p, and on his total required service time, t. Thus we define

$$W_p(t) = \text{expected wait in queues for a customer from priority group } p \text{ whose total required service time is } t \text{ seconds.}$$

This quantity $W_p(t)$ is especially interesting since it represents, for the models considered, the *extra* time that a customer must spend in the time-shared system due to the fact that other customers are sharing the system with him.[2] It is clear why we choose to condition the average wait with respect to p, the externally determined priority group. The reason for further conditioning this wait on t, the service time, is that in most time-shared systems, it is desired that customers whose service time is small should have waiting (queueing) times which are correspondingly short. In order to observe this feature we therefore condition our results on service time also.

The set of quanta $\{g_{pn}Q\}$ is an arbitrary set of numbers which may be chosen to model many feedback queueing systems of interest. For example, in [1] a round-robin system is studied in which $P = 1$ (all customers have the same priority) and $g_{pn} = 1$ (all quanta are equal). In [2], a generalized round-robin system is considered in which $g_{pn} = g_p$ (the quanta for all visits to service are the same for a given priority group). In the foreground-background model studied in [3], $g_{pn} = g_n$ (no priority structure, but different returns to service may have different quanta of service).

Note that unless $B_p(t) = 1$ for all $t \geq T$ (for some $T < \infty$) then $\sum_n g_{pn}$ *must diverge*. Otherwise, customers whose service time exceeds $Q \sum_n g_{pn}$ could not possibly ever be completely served. We also require that $g_{pn} < \infty$ for all p and n.

In any attempt to represent real world systems with mathematical models, one is usually forced to compromise precise representation for mathematical tractability; this study is no exception. Consequently, results obtained are precise only relative to the representation and give an indication rather than a prediction as to how the real world systems behave.

III. Results for Attained Service Time Distribution

The function which we are interested in, the distribution of attained service time, is defined as follows:

$$N_p(\tau) = \text{expectation of the number of customers in the system of queues from priority group } p \text{ who have so far received exactly } \tau \text{ seconds of service.}$$

[2] The average total time $T_p(t)$ in system for customers from priority group p whose total service time is t seconds is, of course, $T_p(t) = W_p(t) + t$.

This quantity, $N_p(\tau)$, gives one a description of the composition of the various queues, and a measure of the relative state of partial service received by those customers still in the system.

We begin by considering feedback queueing systems for which $Q > 0$. In such systems, it is clear that all customers from priority group p who have visited the service facility exactly n times must have so far been given exactly $\sum_{i=1}^{n} g_{pi}Q$ seconds of service. Since the times $\sum_{i=1}^{n} g_{pi}Q$ are the only possible lengths of attained service for those customers in queues, we choose to define

$$\tau_n = \sum_{i=1}^{n} g_{pi}Q. \tag{1}$$

Note that τ_n is also a function of the priority group p; however, we choose to suppress this in the notation, since τ_n will usually appear as an argument of a function which already expresses its dependence upon p. The distribution of the expected number of people having received τ_n seconds of service is given in Theorem 1.

THEOREM 1. *For any feedback queueing system in the class defined above, the distribution of attained service (for $Q > 0$) is given by*

$$N_p(\tau_n) = \lambda_p[1 - B_p(\tau_n)][W_p(\tau_{n+1}) - W_p(\tau_n)]. \tag{2}$$

Proof. We begin by observing that for all t lying in the interval $\tau_{n-1} < t \leqslant \tau_n$ we must have $W_p(t) = W_p(\tau_n)$, since a customer who requires a total service time of t seconds in this range must join the queue exactly n times. The only distinction among customers with such service times is that they spend different amounts of time in their final visit to the service facility (but spend the same average time in queues). Moreover, as members from the pth priority group enter the system of queues for the nth time, they are indistinguishable to the queue organizer [recall that the only remaining feature truly distinct among these customers is their total service time, and we assume that this is known only to within the distribution $B_p(t)$]. Consequently, the average queueing time on this nth visit to the queue will be identical for all customers from the pth priority group.[3] As a result, $W_p(\tau_n)$ may also be interpreted as the average time spent waiting on queues, prior to the nth visit to

[3] Observe that we have provided a means for giving different grades of service to the various priority groups by introducing the set of quanta $\{g_{pn}Q\}$. Within our class of feedback queueing systems, some systems will provide for no further distinctive treatment among the priority groups; for these systems, the average queueing time on the nth visit to the queue will be identical for *all* priority groups. However, in some systems, further preferential treatment may be provided to the higher priority groups by allowing them to join the queue (on their nth visit to the system of queues) in front of lower priority groups; in this case different groups will experience different average queueing times.

the service facility, for all p-type customers whose service time requirements *exceed* τ_{n-1} seconds. This also demonstrates the simple fact that the average time spent on the nth queue is merely the difference in expected total queueing time for customers who must join the system of queues n times and those who must join it $n-1$ times, i.e., $W_p(\tau_n) - W_p(\tau_{n-1})$.

We now make use of Little's result [4], which states that, for any ergodic queueing structure, the expected number of units in the structure is equal to the product of the average arrival rate of units to that structure and the expected time such units spend in that structure. We apply this result by focusing attention on those customers (from priority group p) who are currently in their $n+1$st pass through the system of queues (i.e., those who have made exactly n visits to the service facility). Since

$$\lambda_p = \text{average arrival rate of } p\text{th priority units to the system,}$$

and since

$$1 - B_p(\tau_n) = \Pr[\text{a } p\text{th priority unit will visit the service facility more than } n \text{ times}],$$

we conclude that

$$\lambda_p(n) \equiv \lambda_p[1 - B_p(\tau_n)] = \text{arrival rate of } p\text{th priority units to the } n+1\text{st pass through the system of queues.}$$

As discussed above, the average time spent in making the $n+1$st pass through the system of queues is $W_p(\tau_{n+1}) - W_p(\tau_n)$. Thus, by Little's result, the product of these last two gives the expected number of pth priority units making their $n+1$st pass through the system of queues, viz.,

$$\lambda_p(n)[W_p(\tau_{n+1}) - W_p(\tau_n)].$$

But each of these customers has so far received τ_n seconds of service, and so

$$\begin{aligned} N_p(\tau_n) &= \lambda_p(n)[W(\tau_{n+1}) - W_p(\tau_n)] \\ &= \lambda_p[1 - B_p(\tau_n)][W_p(\tau_{n+1}) - W_p(\tau_n)], \end{aligned} \tag{3}$$

which completes the proof of Theorem 1.

We now consider the case in which $Q \to 0$. This corresponds to time-shared systems in which each customer cycles through the system of queues infinitely fast for an infinite number of cycles and spends an infinitesimally small amount of time in the service facility each time he visits it. The service facility in such a case is constantly cycling among different customers in a continuous way. In a real sense, then, all customers present in the system are using a fraction of the service facility's capacity on a full-time basis. Indeed, the fraction of the machine being used by a

customer from the pth priority group at some time T who has an attained service in the interval $(\tau, \tau + d\tau)$ is merely $g_p(\tau)/\sum_{p=1}^{P} \int_0^\infty n_p(s)\, g_p(s)\, ds$, where $g_p(\tau) = \lim_{Q\to 0} g_{pn(Q)}$ [see Eq. (6) for $n(Q)$] and $n_p(\tau)\, d\tau$ is the number of such customers at time T. Such an operating procedure may be referred to as a "processor-shared" system, and a discussion of its behavior may be found in [2]. The usefulness of this limit of processor-sharing lies in its representation of an idealized sharing operation in which "swap-time"[4] is assumed to be zero. This assumption of zero swap-time is an important simplification in these models; the results thus obtained are idealized in the sense that nonzero swap-time can only degrade the performance of such system. Models with nonzero swap-time have been considered in the literature, e.g., see [5]. For this case, we have the natural analogue of Theorem 1,

THEOREM 2. *For any processor-sharing system* $(Q \to 0)$ *in the class defined above, the density*[5] *of attained service is given by*

$$N_p(\tau) = \lambda_p[1 - B_p(\tau)] \frac{dW_p(\tau)}{d\tau}. \tag{4}$$

Proof. For $Q > 0$, we have from Theorem 1,

$$N_p(\tau_n) = \lambda_p[1 - B_p(\tau_n)][W_p(\tau_{n+1}) - W_p(\tau_n)].$$

Dividing both sides by $\tau_{n+1} - \tau_n (= Qg_{p,n+1})$ we have

$$\frac{N_p(\tau_n)}{\tau_{n+1} - \tau_n} = \lambda_p[1 - B_p(\tau_n)] \frac{W_p(\tau_{n+1}) - W_p(\tau_n)}{\tau_{n+1} - \tau_n}. \tag{5}$$

We must concentrate our attention on a particular value of attained service, say τ. If we choose some τ_n with fixed n, then as $Q \to 0$, we have $\tau_n = Q \sum_{i=1}^{n} g_{pi} \to 0$; that is, τ_n goes to zero as Q goes to zero. Therefore, in order to remain at a fixed τ, we must let n increase as Q decreases. Thus letting $n = n(Q)$ we define $n(Q)$ for any value of Q such that

$$\tau_{n(Q)} \leqslant \tau < \tau_{n(Q)+1}. \tag{6}$$

Also, since g_{pn} is bounded, the difference $\tau_{n(Q)+1} - \tau_{n(Q)} = Qg_{p,n(Q)+1}$ must go to zero as Q goes to zero, and so we can make $\tau_{n(Q)}$ as close to τ as we like. Indeed,

$$\lim_{Q\to 0} \tau_{n(Q)} = \tau.$$

[4] Swap-time is the time used in removing one customer from the service facility and bringing a second customer into the facility.

[5] $N_p(\tau)$ is now a density (whose units are customers/second). Thus $\int_{T_1}^{T_2} N_p(\tau)\, d\tau$ gives the expected number of customers from the pth priority group in the system of queues who have so far received between T_1 and T_2 seconds of service.

Choosing $n = n(Q)$ in Eq. (5) and letting $Q \to 0$, we get (using the usual definition of derivative),

$$\lim_{Q \to 0} \frac{N_p(\tau_{n(Q)})}{\tau_{n(Q)+1} - \tau_{n(Q)}} = \lambda_p[1 - B_p(\tau)] \frac{dW_p(\tau)}{d\tau}. \tag{7}$$

The left-hand side of Eq. (7) is the ratio of two quantities, both of which are approaching zero. Viewed in terms of a discrete bar graph, the left-hand side is taking the contribution at $n(Q)$ and distributing it uniformly over the vanishing interval $\tau_{n(Q)} \leqslant \tau < \tau_{n(Q)+1}$. The limit, then, must be a density (customers per second) which we define as

$$N_p(\tau) = \lim_{Q \to 0} \frac{N_p(\tau_{n(Q)})}{\tau_{n(Q)+1} - \tau_{n(Q)}}.$$

The substitution of this last equation into Eq. (7) completes the proof of Theorem 2.

IV. Discussion and Examples

As mentioned in the introduction, the main results of this paper express, for any time-shared system included in the large class considered, the distribution of attained service in terms of known quantities [λ_p and $B_p(t)$] and in terms of the major performance measure of such systems, namely, the average conditional waiting time $W_p(t)$. Indeed, an especially simple expression is developed which involves the difference, Eq. (2), or differential, Eq. (4), of $W_p(t)$.

We comment here that in order to get the distribution of attained service for customers in the total system (queue plus service facility), we need merely replace $W_p(t)$ by $T_p(t)$, the average conditional time spent in the total system.

As examples of the determination of $N_p(\tau)$, we consider two feedback queueing systems studied in the literature, both of the processor-sharing type (i.e., $Q \to 0$). The first is the priority processor-shared system studied in [2]. As stated above, for that system there exists only one queue (see Fig. 2). The $\{g_{pn}\}$ are chosen such that $g_{pn} = g_p$ independent of n, the number of returns to the service facility. The input process is considered to be Poisson, and $B_p(t) = 1 - e^{-\mu_p t}$ (exponential service time distribution). The result for $W_p(\tau)$ in that case has been shown to be

$$W_p(\tau) = \tau \sum_{i=1}^{P} \frac{g_i \rho_i}{g_p(1 - \rho)}, \tag{8}$$

where

$$\rho_i = \lambda_i/\mu_i$$

and

$$\rho = \sum_{i=1}^{P} \rho_i.$$

In this case, the computation of $N_p(\tau)$ from Eqs. (4) and (8) is trivial and yields

$$N_p(\tau) = \lambda_p e^{-\mu_p \tau} \sum_{i=1}^{P} \frac{g_i \rho_i}{g_p(1-\rho)}. \tag{9}$$

Thus the distribution of attained service is exponentially distributed with τ. An example is shown in Fig. 4 in which $g_p = p^2$, $P = 5$, $\lambda_p = 0.15$, $\mu_p = 1$ (thus giving $\rho_p = 0.15$, $\rho = .75$).

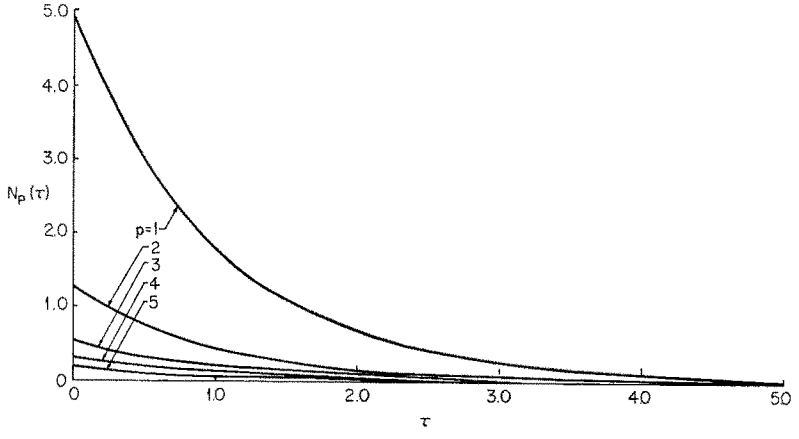

FIG. 4. Distribution of attained service for the priority processor-shared system. $P = 5$, $\lambda_p = 0.15$, $\mu_p = 1.0$, $g_p = p^2$, $\rho = 0.75$.

We note further that the expected total number of customers, N_p, from priority group p (without regard to attained service) is, by definition,

$$N_p = \int_0^\infty N_p(\tau)\, d\tau. \tag{10}$$

By Eq. (9) we get

$$N_p = \lambda_p \sum_{i=1}^{P} \frac{g_i \rho_i}{g_p(1-\rho)} \int_0^\infty e^{-\mu_p \tau} d\tau \tag{11}$$

$$= \frac{\rho_p}{g_p(1-\rho)} \sum_{i=1}^{P} g_i \rho_i, \tag{12}$$

which checks with the result obtained for N_p in [2].[6]

[6] In {2} an expectation E_p is solved for which is related to N_p by $N_p = E_p - \rho_p$.

Also of interest is the expectation $\bar{\tau}_p$ of the attained service for a customer from priority group p. We obtain this as follows

$$\bar{\tau}_p = \frac{\int_0^\infty \tau N_p(\tau)\, d\tau}{\int_0^\infty N_p(\tau)\, d\tau} \tag{13}$$

From Eqs. (4) and (12) we get

$$\bar{\tau}_p = \left(\frac{\lambda_p}{g_p(1-\rho)} \sum_{i=1}^{P} g_i\rho_i \int_0^\infty \tau e^{-\mu_p\tau} d\tau\right)\left(\frac{\rho_p}{g_p(1-\rho)} \sum_{i=1}^{P} g_i\rho_i\right)^{-1}$$

$$= \mu_p \int_0^\infty \tau e^{-\mu_p\tau} d\tau.$$

Thus

$$\bar{\tau}_p = 1/\mu_p\,. \tag{14}$$

Equation (14) states for the priority processor-shared system, the interesting result that the average attained service for customers from priority group p is merely the average of the total service required. Thus, given that a customer is still in the system of queues, he needs on the average as much more service as he needed when he first entered the system.

The second example worth considering is the foreground-background processor-sharing systems ($Q \to 0$) treated in [5, 6]. Here, there is an infinity of queues arranged in a hierarchy as shown in Fig. 3. When a customer is waiting for his nth visit to the service facility, he waits on the nth queue. Here again the arrivals are Poisson, the service is exponential, and $g_{pn} = 1$ (i.e., no priorities and equal quanta for each visit to the service facility). When the service facility can accept a new customer, it takes one from the lowest numbered nonempty queue. The solution obtained for $W_p(\tau)$ [which in this no-priority case may be denoted by $W(\tau)$] is (see [6])

$$W(\tau) = \frac{(\rho/\mu)[1 - e^{-\mu\tau} - \mu\tau e^{-\mu\tau}]}{[1 - \rho(1 - e^{-\mu\tau})]^2} + \frac{\rho(1 - e^{-\mu\tau})\tau}{1 - \rho(1 - e^{-\mu\tau})}\,. \tag{15}$$

Thus $N(\tau)$ from Eq. (4) is

$$N(\tau) = \frac{2\rho\lambda e^{-2\mu\tau}[\rho(1 - e^{-\mu\tau}) + \mu(1 - \rho)\tau]}{[1 - \rho(1 - e^{-\mu\tau})]^3} + \frac{\lambda\rho e^{-\mu\tau}(1 - e^{-\mu\tau})}{1 - \rho + \rho e^{-\mu\tau}}\,. \tag{16}$$

In Fig. 5 we plot $N(\tau)$ as a function of τ, with $\lambda = .75$, $\mu = 1$, and thus $\rho = \lambda/\mu = 0.75$).

Forming the expected number of customers in the system of queues (without regard to attained service) we get, after some computation, from Eq. (16),

$$N = \int_0^\infty N(\tau)\, d\tau = \frac{\rho^2}{1 - \rho}\,. \tag{17}$$

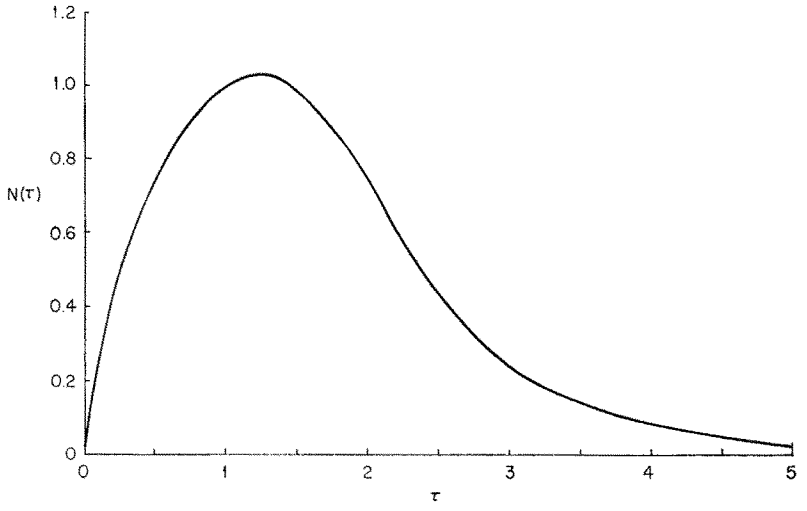

FIG. 5. Distribution of attained service for infinite level foreground-background processor-shared system. $\lambda = 0.75$, $\mu = 1.0$, $\rho = 0.75$.

This last result is especially simple and pleasing since it corresponds to the well-known result for any priority discipline with Poisson arrivals and exponential service. The calculation of the average attained service $\bar{\tau}$ as defined by Eq. (13) without the subscript p is extremely difficult and is not given here.

V. CONCLUSION

The results of this paper give general expressions for the distribution of attained service for any member of a wide class of time-shared systems, including those with priority inputs. The answers are good for finite service quanta (the feedback queueing systems) as well as for service quanta approaching zero (the processor-shared systems). The results are given simply in terms of known functions and in terms of the average conditional waiting time in queue (this last being a useful performance measure for time-shared systems).

The usefulness of obtaining the distribution of attained service lies in the fact that it provides one with an index of the composition of the system of queues. This index may prove useful in choosing time-shared algorithms to minimize cost functions which involve the attained service or the additional required service for customers in the system of queues. The examples included give representative curves for the attained service.

References

1. L. Kleinrock. *Naval Res. Logistics Quart.* **11**, 59-73 (1964).
2. L. Kleinrock. *J. Assoc. Computing Machines* **14**, 242-261 (1967).
3. L. E. Schrage. *Management Sci., Ser. A* **13**, 466-474 (1967).
4. J. D. C. Little. *Operations Res.* **9**, 383-387 (1961).
5. E. G. Coffman. "Stochastic Models of Multiple and Time-Shared Computer Operations," U.C.L.A. Dept. of Engineering Report No. 66-38, June (1966).
6. E. G. Coffman and L. Kleinrock. Some feedback queueing models for time-shared systems, *Proc. of 5th Intern. Teletraffic Congr., New York,* 288–304 (1967).