SOME FEEDBACK QUEUEING MODELS FOR TIME-SHARED SYSTEMS†

by Edward G. Coffman
Princeton University, Princeton, New Jersey, U.S.A.
and Leonard Kleinrock
University of California, Los Angeles, California, U.S.A.

Abstract

Time-shared processing systems (e.g., communication or computer systems) are studied by considering priority disciplines operating in a stochastic queueing environment. Results are obtained for the average time spent in the system, conditioned on the length of required service (e.g., message length or number of computations). No charge is made for swap-time, and the results hold only for Markov assumptions for the arrival and service processes.

Two distinct feedback models with a single quantum-controlled service are considered. The first is a Round-Robin (RR) system in which the service facility processes each customer for a maximum of q seconds; if the customer's service is completed during this quantum, he leaves the system, otherwise he returns to the end of the queue to await another quantum of service. The second is a Feedback (FB_N) system with N queues in which a new arrival joins the tail of the first queue. The server gives service to a customer from the n^{th} queue only if all lower-numbered queues are empty. When taken from the n^{th} queue, a customer is given q seconds of service; if this completes his processing requirement he leaves the system, otherwise he joins the tail of the $n+1^{st}$ queue ($n = 1, 2, \ldots, N-1$). The limiting case of $N \to \infty$ is also treated. Both models are therefore quantum-controlled, and involve feedback to the tail of some queue, thus providing rapid service for customers with short service-time requirements. The interesting limiting case in which $q \to 0$ (a "processor-shared" model) is also examined. Comparison is made with first-come-first-served and also shortest-job-first discipline. Finally the FB_∞ system is generalized to include (priority) inputs at each of the queues in the system.

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

I. Introduction

The value of time-shared processing systems as a means of providing a processor to many users concurrently is well established. Examples include the "simultaneous" use of communication channels, and communication networks as well as computers and computer networks. However, it also is clear that the effectiveness of these systems depends in large part on the efficiency with which the resources of the processor are allocated to the individual users. Thus, considerable attention has been focused on the time and space scheduling problems of time-sharing systems giving rise to the description of sophisticated algorithms and, in those cases where it is possible, an analysis of more or less simplified queueing models of these algorithms.

In this paper we are concerned with extending the analyses that have been made for the so-called feedback queueing models of time-shared processor operation. In these models the service received by users (messages, programs, etc.) is made to depend, either implicitly or explicitly, on a user's service time (e.g., transmission time in a communication example or running time in a computer example). However, it is assumed that the service time is not known a priori. In the following we shall discuss informally the queueing models that are subsequently given a precise definition and analyzed under Markov assumptions applied to the service and arrival mechanisms. By Markov assumptions we mean that the interarrival and service times will be assumed to be exponential or geometric random variables depending on whether we are analyzing the model of interest in continuous or discrete time, respectively.

The term "feedback" is a natural one when one considers that in time-sharing disciplines, users are allocated limited time intervals for operation, and if the operation time required exceeds these limits the user is interrupted and "fed back" to the end of the same or some other queue to await its next interval of service. The so-called round-robin algorithm represents what is perhaps the simplest of the feedback (FB) algorithms. With this procedure users are allocated fixed time intervals (quanta) of operation time; if the users terminate within this interval they leave the system and if not, they are placed at the end of the waiting line to await their next quantum of service. It is not difficult to see that users with shorter service requirements receive better treatment in this type of system. (This property will be quantified later.) Indeed, this property characterizes time-sharing disciplines as a whole and will be seen to apply to the other FB models we consider.

The more complicated FB models that we analyze involve multiple queues, each queue corresponding to a priority class of users based on the service requirements of the users. The discipline for selecting which queue to service corresponds to that of conventional priority queues; viz., users at the n^{th} level are not served unless all of the n-1 lower level (higher priority) queues are empty. However, in the FB priority queues the operation time is again allocated on a quantum basis; a user requiring more than the time allocated at a given queue level is moved up (following its quantum of service) to the end of the next higher level (lower priority) queue. Thus, in the multiple level FB system the priority received by a user is made to depend in an explicit way on the amount of service he has already received. Although the dependence of the time-sharing service disciplines on service time is an a posterior one, the general FB model to be studied also includes an initial assignment of users to queue levels based on a priority scheme using a criterion other than service time (e.g., program size). In other words, we shall assume that a new arrival may join any one of the multiple queues according to some fixed probability distribution.

Our principal interest is in the analysis of these algorithms and a study of the results obtained. The basic results will take the form of expressions for expected waiting times conditioned on the amount of service required and, in the most general model, the arrival priority (corresponding to the queue to which the arrival was originally assigned). We shall study these results by considering their variation with changes in the value of such parameters as quantum size and loading factor. Of particular interest in this regard will be the so-called processor-shared models in which the quantum size is allowed to approach zero. As we shall see, these models correspond to systems which divide up their processing capacity among all the users requiring service simultaneously.

II. The Time-Sharing Models

A. The Round-Robin Model

As implied in Figure 1 units arrive to the system from an infinite source. The stochastic input process will be described by an interarrival time distribution which we denote by A(t). The units are assumed to take their place at the end of the queue immediately on arrival. The service requirements of arriving units are subject to a stationary probability distribution $B(\tau)$.

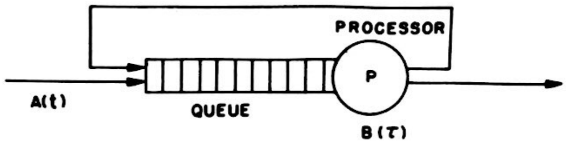

FIG. 1 THE ROUND-ROBIN MODEL

The service discipline is such that units are taken from the queue first-come-first-served and provided with a certain fixed amount of service which we shall denote by q (for quantum). If the unit being served completes within the time q then it is simply ejected from the system. If on the other hand, it requires more time to complete then it is removed from the service facility (processor) and put back to the end of the line. In due course, after the other units in line ahead of this unit have received their quantum of service, the interrupted unit is again served, continuing from the point at which the previous service was interrupted; i.e., we have a "preemptive resume" rule implying that service is not lost because of interruption. The procedure as outlined is continued for all units in the queue, each unit making as many of the "loops" shown in Figure 1 as needed to complete its total service requirement. We shall assume for all of the models described in this paper that no "overhead" or "swap" time is associated with the process of unloading and loading units from the processor. In this respect our results may be viewed as upper bounds on system performance. (See References 4 and 11 for results applying to similar models for which non-zero swap times and a finite source are assumed.)

For the distributions A(t) and $B(\tau)$ we shall present results for the following two sets of (Markov) assumptions.

1. The input process has a discrete time parameter $t = nq$ (n an integer) where the quantum size q is the basic time interval and n is distributed according to the geometric distribution (this describes the so-called Bernoulli arrival process). Thus, we have

$$A(t) = A(nq) = \sum_{k=1}^{n} a(k) \tag{1}$$

where

$$a(k) = (1-\xi)\,\xi^{k-1} \qquad k=1,2,3,\ldots \tag{2}$$
$$0 \le \xi < 1$$

The mean interarrival period is given by

$$q \sum_{k=1}^{\infty} k\, a(k) = \frac{q}{1-\xi} \text{ seconds}$$

from which the mean arrival rate is found to be $(1-\xi)/q$ per second. The above model was first analyzed by Kleinrock.[2] Secondly, the service time is assumed to be a discrete random variable with the same basic time unit of q seconds. In particular, we assume the geometric distribution

$$B(\tau) = B(mq) = \sum_{k=1}^{m} b(k) \tag{3}$$

with

$$b(k) = (1-\zeta)\,\zeta^{k-1} \qquad k=1, 2, 3, \ldots \qquad 0 \le \zeta < 1 \tag{4}$$

The mean servicing time is thus $q/(1-\zeta)$ seconds. For the discrete model an assumption must be made regarding the order in which events take place at the end of a time interval. Consider two types of systems: the first system allows the unit in service to be ejected from the service facility (and then allows it to join the end of the queue, if more service is required for this unit), and instantaneously thereafter a new unit arrives (with probability $1-\xi$). This is referred to as a late arrival system. The second system reverses the order in which these events are allowed to occur, giving rise to the early arrival system. In both systems, a new unit is taken into service at the beginning of a time interval. We shall cite results for both models in the next section.

2. The input process is the Poisson process so that $A(t)$ is given by the exponential distribution

$$A(t) = \begin{cases} 1-\epsilon^{-\lambda t}; & t \ge 0 \\ 0; & t < 0 \end{cases} \qquad \lambda > 0 \tag{5}$$

The mean arrival rate is easily calculated to be λ units per second. The service time τ is assumed to be exponentially distributed as follows

$$B(\tau) = \begin{cases} 1-\epsilon^{-\mu\tau}; & \tau \ge 0 \\ 0; & \tau < 0 \end{cases} \qquad \mu > 0 \tag{6}$$

with a mean (service time) of $1/\mu$ seconds.

B. The Processor-Shared Models

Since we assume swap time to be zero we may consider the case of a round-robin system in which $q \to 0$. For the continuous (Markov) model described above there is no difficulty in taking the limit of the results as $q \to 0$. (See Appendix A.) However, in the discrete model we must be careful in taking this limit since the service and interarrival times also go to zero leaving us with a vacuous system. Thus, we must agree to keep the average service time and average arrival rate constant while letting $q \to 0$. In both the discrete and continuous Markov models the resultant model is the so-called processor-shared model (see Reference 3) of Figure 2 whose interarrival and service times are exponential. As shown by Figure 2, in the processor-shared model all units in the system receive service concurrently and experience no waiting time in queue. However, the rate (e.g., operations/sec)

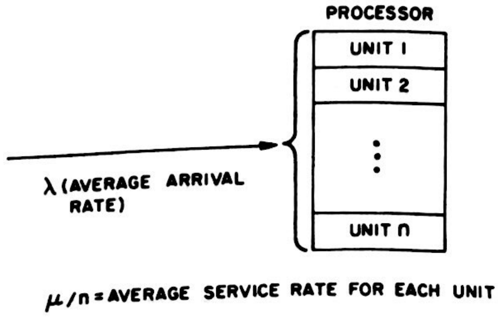

FIG. 2 PROCESSOR-SHARED MODEL WITH n UNITS IN THE SYSTEM

at which the units sharing the processor receive service is inversely proportional to the number of units in the system, which of course varies as new units arrive and old ones leave. Thus, considering a computer program as an example, we see that a program operates at $1/k^{th}$ the speed it would run were it alone in the computer, if we assume there are k-1 other programs in the machine at the same time.

The priority processor-shared model[3] is a generalization of the processor-shared system considered above. With reference to the continuous model we assume here that the input traffic is broken up into P separate priority groups, where the arrivals from the p^{th} group constitute a Poisson process with an average rate of λ_p units per second, and have an exponentially distributed service requirement whose mean is $1/\mu_p$ seconds. For the $q \ne 0$ case, we give a member of the p^{th} priority group $g_p q$ seconds of service each time he cycles around the queue.

For $q \to 0$ this model then reduces to a processor-shared model (see Figure 3) with a priority structure whereby a member from group p receives service at time t at a fractional rate $f_p(1/\mu_p)$, where

$$f_p = \frac{g_p}{\sum_{i=1}^{P} g_i n_i} \tag{7}$$

and where n_i is the number of members from group i present in the system. The non-priority processor-shared model considered earlier is the special case $g_p = 1$ for all p.

C. The Multiple Level FB Model

This model, which we shall denote by FB_N where N is the number of levels, is shown in Figure 4. We shall make the assumptions of exponential interarrival and service times (see Equations (5) and (6)). As pointed out earlier a unit at the service point at any given queue level will not be serviced unless all lower level queues are empty. Thus, immediately after a unit has received service the next unit serviced will be the one at the service point of the lowest level,

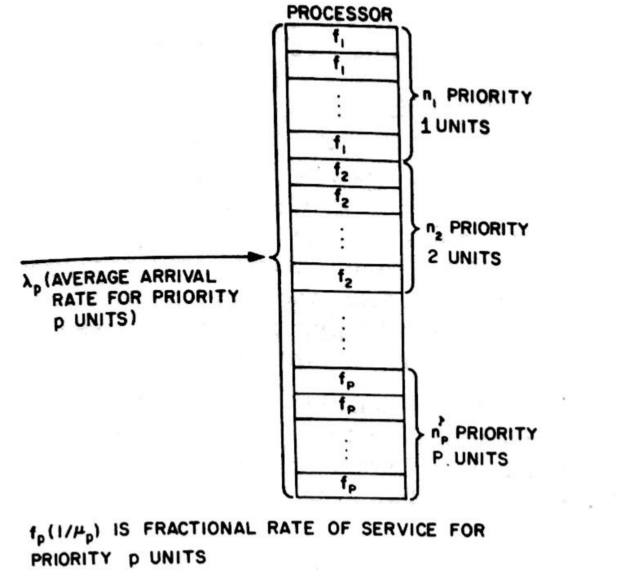

FIG. 3 PRIORITY PROCESSOR-SHARED MODEL

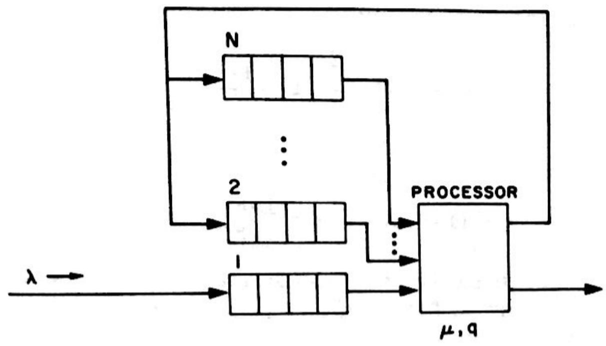

FIG. 4 THE FB_N MODEL

non-empty queue. This unit will be given a quantum (q) of service as in the round-robin model; if more is needed then the unit is subsequently placed at the end of the next higher level queue, otherwise it leaves the system.

If $N < \infty$ the question arises as to what happens at the highest level (the N^{th} level). We shall assume that the N^{th} level queue is a quantum-controlled, first-come-first-served (FCFS) queue. Specifically, units at the N^{th} level are served a quantum at a time until completion (i.e., there is no round-robin in the N^{th} queue but an arrival to a lower level during the servicing of an N^{th} level unit will preempt this unit after it has completed the quantum-service in progress). Note that, with these assumptions, FB_1 denotes the conventional FCFS system.

It is easy to see that the FB_N service discipline shares that property of the RR service discipline according to which the units with shorter service requirements enjoy shorter waiting times at the expense of the waiting times of units with the longer service requirements. However, this property is even more pronounced in the FB_N models, as we shall demonstrate later on.

As pointed out earlier the limiting case in which q goes to zero is of interest. For finite N the FB_N system reduces to a FCFS system. This can be seen by observing that the first N-1 levels of the FB_N system provide an infinitesimal amount of service when q becomes very small, and consequently do not significantly delay the service at the N^{th} level. That is, arrivals can be viewed as being immediately switched to the N^{th} level queue in the limit $q = 0$. At the N^{th} level the units are served to completion in the order of their arrival, receiving an infinite number of infinitesimal quanta, where in the limit we have a FCFS system. This result is verified analytically in the next section.

Of greater interest is the limiting case $q = 0$ when we assume $N = \infty$. By arguments based on very small q sizes it can be seen that the resulting system can be viewed as corresponding to a system in which arrivals always preempt the unit, if any, in service and are allowed service until their service time exceeds that having been received by some other unit in the queue. In short, we have a preemptive-resume queueing discipline in which the unit in service is preempted whenever there exists another unit in the system whose time in the service facility has been less. It is clear that when there exists at least two units having received the same amount of service time then the processor begins switching between them infinitely often. Thus, under these circumstances, we have the processor-sharing case as described earlier for the RR model. The two units together then proceed to share the processor until their received service time reaches that received by some other unit, if any, in the queue. At this time the two units are joined by the third one and all three share the processor. This sort of process continues until units complete (thus reducing the number sharing the processor), or until a new arrival occurs, at which time it receives the whole processor and the procedure above begins once again.

D. The Multiple Level FB Model with Priorities

There exist many ways to increase the number of degrees of freedom for manipulating waiting times in the multiple level queueing model defined above. In the FB_N model we note two degrees of freedom: the quantum size q and the number of levels N. What is perhaps the most obvious way to further control the distribution of waiting times is to assign external priorities to the arriving units.

Figure 5 illustrates this type of extension to the FB_N model for the special case $N = \infty$. In particular, we assume an infinite number of levels (queues) and an independent, Poisson input to each level with average arrival rate λ_p per second. We shall define

$$\lambda = \sum_{p=1}^{\infty} \lambda_p \tag{8}$$

and require that $\lambda < \infty$. The service times for arrivals at every queue or priority level are assumed to be independent, exponential random variables distributed according to Equation (6). As in the FB_N model the lowest level, non-empty

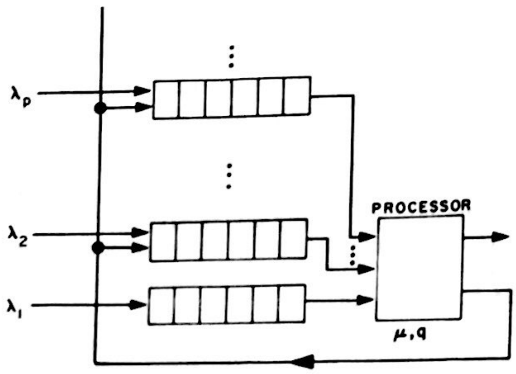

FIG. 5 THE PRIORITY FB_∞ MODEL

queue is chosen for service, and service is allocated q seconds at a time with units requiring more moving up level-by-level as described earlier.

Our description is completed by specifying that the service discipline at each queue level is highest-priority-first. By highest priority we mean the lowest level queue of arrival to the system. That is, in a given queue, the unit to be served next must have entered the system originally at a queue level that is equal to or less than the queue levels of arrival for the remainder of the units in the given queue. Within a priority group in a given queue the discipline will be FCFS.

Further generalizations to the multiple level model that may be considered are those of different quantum sizes for different levels and different mean service times for different priority-level units. To extend the results to include these generalizations is a simple matter conceptually, but introduces more awkward symbology. Although we shall not carry out a complete analysis for these additional degrees of freedom, Reference 4 indicates the basic changes that would be necessary.

Once again, it is of interest to investigate the case when q goes to zero. For this, we proceed in the same manner Phipps[7] employed to extend Cobham's[9] analysis of conventional priority queues to a continuous number of priorities. In our model, as q goes to zero we shall pass from a countable number of priorities to a continuous number of priorities. Following Phipps we introduce λ_τ as the arrival rate for the continuous time-priority τ such that,

$$\lambda = \int_0^\infty \lambda_\tau \, d\tau$$

The present degenerate model differs from the preemptive processor-shared model discussed earlier in only one respect. Arrivals of priority τ are not given their first service unless and until all units of priorities $\xi < \tau$ have been given at least $\tau - \xi$ seconds of service. When this situation eventually does obtain we have the processor-sharing and ascension of levels that was described for the preemptive processor-shared model. Of course, if the above condition exists when the priority τ unit arrives, then preemption of the unit(s) in service occurs immediately.

III. Results for the Time-Sharing Models

In the order of the descriptions in the last section the mean waiting times, conditioned on the amount of service required, are presented below for the FB models. The results are presented in the form of theorems. Some of the results presented are taken from the literature and are referenced accordingly; proofs of the remaining theorems are supplied in Appendices A and B.

First, we consider the discrete RR (round-robin) model. Equations (2) and (4) describe the geometric distributions to be assumed for the interarrival and service times. We have the following theorem.

Theorem 1 (Kleinrock)[2]

(a) In the late arrival system the mean waiting time in system[†] for a unit requiring kq seconds of service is given by

$$W_k = \frac{kq}{1-\rho} - \frac{(1-\xi)q^2}{1-\rho}\left[1 + \frac{(1-\zeta\alpha)(1-\alpha^{k-1})}{(1-\zeta)^2(1-\rho)}\right] \qquad (9)$$

where,

$$\alpha = \zeta + (1-\xi)q, \qquad \rho = \frac{1-\xi}{1-\zeta}\,q \qquad (10)$$

(b) In the early arrival system the mean waiting time in system for a unit requiring k quanta of service is given by

$$W'_k = \frac{kq}{1-\rho} - \rho q - \frac{(1-\xi)\rho q^2}{1-\rho}\left[1 + \frac{(1-\zeta\alpha)(1-\alpha^{k-1})}{(1-\zeta)^2(1-\rho)}\right] \qquad (11)$$

We now consider the continuous RR model in which the exponential distributions defined by Equations (5) and (6) are assumed for the interarrival and service times.

Theorem 2[‡]

Let the "quantum-service" distribution be defined as follows.[§]

$$F_1(\tau) = \begin{cases} 0 \, ; & \tau < 0 \\ 1 - \epsilon^{-\mu\tau} \, ; & 0 \le \tau < q \\ 1 \, ; & \tau \ge q \end{cases} \qquad (12)$$

Then the mean waiting time in the continuous RR system of a unit requiring t seconds of service is

$$W(t) = t + \frac{\rho k q}{1-\rho} + \frac{(\lambda/2)E_1(\tau^2)}{1-\beta}\left[1-\rho\beta^{k-1}\right]$$
$$+ \frac{1}{1-\rho}\left[\frac{\rho^2}{1-\rho}(1/\mu) - \frac{\rho q}{1-\beta}\right]\left[1-\beta^k\right] + \frac{\rho\epsilon^{-\mu q}}{1-\rho}(1/\mu)\left[1-\beta^{k-1}\right] \qquad (13)$$

[†] This will be the sum of the time spent in the queue and the time spent in the service facility.

[‡] The proof appears in Appendix A.

[§] This is simply the distribution of the amount of time taken by a unit to which q seconds of service is allocated.

where

$$\rho = \lambda/\mu \tag{14}$$

$$\beta = \rho + (1-\rho)\,\epsilon^{-\mu q} \tag{15}$$

k is the smallest integer such that $kq > t$, and $E_1(\tau^2)$ is the second moment of the quantum-service distribution in Equation (12). Specifically,

$$E_1(\tau^2) = \int_0^\infty \tau^2\, d\,F_1(\tau) = \frac{1}{\mu^2}\left[1-(2\mu q + \epsilon^{-\mu q})\,\epsilon^{-\mu q}\right] \tag{16}$$

For the limiting case $q \to 0$ we have the following result for the processor-shared model.

Theorem 3 (Kleinrock)[3]

The expected value of the total time spent in the processor-shared system for a unit requiring t seconds of service is

$$W(t) = \frac{t}{1-\rho} \tag{17}$$

where ρ is defined by Equation (14). Although Kleinrock obtains Equation (17) by taking the limit $q \to 0$ for the discrete (either the late or early arrival) system, we shall produce the same result in Appendix A as a limit of the continuous system (Equation (13)). As verified by Kleinrock, the geometric interarrival and service times of the discrete models in the limit $q \to 0$ become exponentially distributed if $\zeta \to 1$ appropriately.

In the conventional FCFS system (i.e., the FB_N system with $q = \infty$) the waiting time in the queue is independent of t and the waiting time in system easily found to be (see Reference 6, for example)

$$W(t) = \frac{\rho(1/\mu)}{1-\rho} + t \tag{18}$$

Comparing Equations (17) and (18) we note immediately that units requiring more than the average amount of service ($1/\mu$ seconds) have longer waiting times in the processor-shared system than for the FCFS system, whereas the opposite is true for units requiring less than the average amount of service.

For the priority processor-shared system in which there are P priority groups each receiving a fractional capacity of the machine determined by Equation (7) we have the following result:

Theorem 4 (Kleinrock)[3]

For the priority processor-shared system the mean waiting time in system of a p^{th} priority unit requiring t seconds of service is

$$W_p(t) = t\left[1 + \sum_{i=1}^{P} \frac{g_i\,\rho_i}{g_p(1-\rho)}\right] \tag{19}$$

where

$$\rho_p = \lambda_p/\mu_p \tag{20}$$

$$\rho = \sum_{p=1}^{P} \rho_p \tag{21}$$

and $g_p > 0$; $p = 1, 2, 3, \ldots, P$.

Turning now the FB_N model let the inter-arrival and service times be independently and exponentially distributed as before. We have the following result.

Theorem 5†

A unit requiring t seconds of service in the FB_N system has an expected waiting time in system of

$$W(t) = \frac{(\lambda/2)\left[E_k(\tau^2) + \gamma_k E_1(\tau^2)\right]}{\left[1-\rho(1-\epsilon^{-\mu kq})\right]\left[1-\rho(1-\epsilon^{-\mu(k-1)q})\right]} + \frac{\rho(1-\epsilon^{-\mu(k-1)q})}{1-\rho(1-\epsilon^{-\mu(k-1)q})}\,(k-1)\,q + t\,; \quad 1 \le k \le N-1 \tag{22a}$$

$$W(t) = \frac{\rho(1/\mu)}{(1-\rho)\left[1-\rho(1-\epsilon^{-\mu(N-1)q})\right]} + \frac{\rho(1-\epsilon^{-\mu(N-1)q})}{1-\rho(1-\epsilon^{-\mu(N-1)q})}\,(k-1)\,q + t\,; \quad k \ge N \tag{22b}$$

where k is the smallest integer such that $kq > t$, where we define $E_k(\tau^2)$ as the second moment of the distribution

$$F_k(\tau) = \begin{cases} 0\,; & \tau < 0 \\ 1-\epsilon^{-\mu\tau}\,; & 0 \le \tau < kq \\ 1\,; & \tau \ge kq \end{cases} \tag{23}$$

with

$$E_k(\tau) = (1/\mu)\left[1-\epsilon^{-\mu kq}\right] \tag{24}$$

$$E_k(\tau^2) = (1/\mu)^2\left[1-(2\mu\, kq + \epsilon^{-\mu kq})\,\epsilon^{-\mu kq}\right] \tag{25}$$

and where

$$\gamma_k = \frac{\epsilon^{-\mu kq}}{1-\epsilon^{-\mu q}} \tag{26}$$

As indicated earlier, Schrage[5] has provided a general analysis of this model in the case $N = \infty$. In particular, the Laplace transform of the waiting time distribution is found under the assumptions of arbitrary quantum sizes for each level. (See also Reference 4 for the generalizations to the priority FB_∞ model.) The methods used in Appendix B are similar to those used by Schrage with a straightforward extension to take care of the boundary condition arising because of a finite N.

For the limiting case in which $q \to 0$ that was discussed earlier we have the following.

† The proof appears in Appendix B.

Corollary 5.1

$$\lim_{q\to 0} W(t) = \begin{cases} \frac{1}{1-\rho}(1/\mu)\,; & N < \infty \quad (27) \\ \dfrac{(\lambda/2)\int_0^t x^2 dF(x)}{\left|1-\rho(1-\epsilon^{-\mu t})\right|^2} + \dfrac{t}{1-\rho(1-\epsilon^{-\mu t})}\,; & N = \infty \quad (28) \end{cases}$$

where,

$$F(x) = \begin{cases} 0; & x < 0 \\ 1-\epsilon^{-\mu x}; & 0 \le x < t \\ 1; & x \ge t \end{cases} \tag{29}$$

As explained in the last section Equation (27) corresponds to the FCFS system while Equation (28) corresponds to a "preemptive" processor-shared system. The result of Equation (28) is easily shown by observing that, from the definition of k, holding t fixed implies

$$\lim_{q\to 0} kq = t$$

Thus, setting $kq = t$ and noting also that $(k-1)q \to kq$ as $q \to 0$ and

$$\lim_{q\to 0} \gamma_k E_1(\tau^2) = 0$$

Equation (22) reduces to Equation (28).

Generalizing to different priority level inputs we now present an expression for the conditional waiting times of the priority FB_∞ model.

Theorem 6†

Let $E_k(\tau)$ and $E_k(\tau^2)$ be defined as in Equations (24) and (25) and let

$$\rho_p = \lambda_p E_1(\tau) \tag{30}$$

denote the utilization factor for the p^{th} level. If we let $W_p(t)$ be the expected waiting time of a p^{th} priority unit (i.e., one entering the system at the p^{th} level) requiring t seconds of service, and let k be the highest numbered level (according to p and t) to which the unit must ascend, then we have

$$W_p(t) = \frac{W_o}{(1-\rho_{pk})\left[1-\rho_{pk}+\rho_p\epsilon^{-\mu(k-p)q}\right]} + \frac{\rho_{pk}-\rho_p\epsilon^{-\mu(k-p)q}}{1-\rho_{pk}+\rho_p\epsilon^{-\mu(k-p)q}}(k-p)q + t \tag{31}$$

where ρ_{pk} is the high priority utilization factor (of an equivalent 2-level model) and is given by

$$\rho_{pk} = \begin{cases} \sum_{r=1}^{p} \lambda_r E_{k-r+1}(\tau) + \sum_{r=p+1}^{k-1} \lambda_r E_{k-r}(\tau); & k > p+1 \\ \sum_{r=1}^{p} \lambda_r E_{k-r+1}(\tau); & k = p \text{ or } p+1 \end{cases} \tag{32}$$

† The proof appears in Appendix C.

and where W_o is the expected time to complete the unit in service at arrival and is given by

$$W_o = \begin{cases} \frac{1}{2}\Lambda_{pk}E_1(\tau^2) + \frac{1}{2}\sum_{r=1}^{p}\lambda_r E_{k-r+1}(\tau^2) + \frac{1}{2}\sum_{r=p+1}^{k-1}\lambda_r E_{k-r}(\tau^2)\,; & k > p+1 \\ \frac{1}{2}\Lambda_{pk}E_1(\tau^2) + \frac{1}{2}\sum_{r=1}^{p}\lambda_r E_{k-r+1}(\tau^2); & k = p \text{ or } p+1 \end{cases} \tag{33}$$

with

$$\Lambda_{pk} = \sum_{r=p+1}^{k-1} \lambda_r \epsilon^{-\mu(k-r)q} + \sum_{j=k+1}^{\infty} \Lambda_j \tag{34}$$

and

$$\Lambda_j = \sum_{r=1}^{j} \lambda_r \epsilon^{-\mu(j-r)q} \tag{35}$$

In the limiting case when $q = 0$ that was described in the last section we have the following result.

Corollary 6.1

Let τ be the continuous time-priority replacing the discrete priority index p when $q \to 0$, and let λ_τ denote the average arrival rate of priority τ units. Then the average waiting time in system $W_\tau(t)$ of a unit entering at priority level τ and requiring t seconds of service is given by

$$W_\tau(t) = \frac{\int_0^{t+\tau} \lambda_\xi E^{(2)}_{t+\tau-\xi}\, d\xi}{2\left[1-\int_0^{t+\tau}\lambda_\xi E^{(1)}_{t+\tau-\xi}\, d\xi\right]^2} + \frac{t}{\left[1-\int_0^{t+\tau}\lambda_\xi E^{(1)}_{t+\tau-\xi}\, d\xi\right]} \tag{36}$$

where

$$E^{(n)}_{t+\tau-\xi} = \int_0^{t+\tau-\xi} \mu x^n \epsilon^{-\mu x}\, dx \tag{37}$$

IV. The Shortest-Job-First Model

The preceding FB models can be characterized by the fact that the type of service received by a unit is made to depend on the total amount required, but with the constraint that this amount is not known a priori. It is desirable to investigate the potential improvement in performance that might exist if this information were available for each unit at arrival time. For this, we shall look at a shortest-job-first (SJF) system which is described as follows. We assume a Poisson input of units with average arrival rate of λ per second. It is assumed that the service time required by a unit is known at the time of arrival, and that it is an exponentially distributed random variable with a mean of $1/\mu$ seconds. Now when the service facility completes the service of a unit it inspects the queue and determines the unit with the shortest service time requirement. It then proceeds to service this unit to completion; that is, there is no preemption by a new arrival with shorter service requirements. The service facility commences immediately the service of a unit that arrives when the facility is idle. Phipps[7] has analyzed this model and derived the following expression for the mean waiting time in queue of a unit whose service requirement is t seconds.

$$W(t) = \frac{\rho(1/\mu)}{1-\lambda/\mu\,[1-\epsilon^{-\mu t}(1+\mu t)]^2} \tag{38}$$

V. Examples and Discussion

The service disciplines discussed in the previous sections offer a variety of techniques by which the waiting times of different classes of units (programs, messages, etc.) can be manipulated or adjusted to meet a set of operational requirements. Of course, for these disciplines to have value it is assumed implicit in the operational requirements of the system that the servicing of certain classes of units is to be favored (in a priority sense) over the servicing of others, based on the service requirements of these classes. An additional, external priority assignment, independent of service times, was also assumed for the generalized multiple level model and for the priority processor-shared model. In this section we shall display, for the FB disciplines of interest, the comparative waiting time performances, how one may manipulate the waiting times by adjusting the basic structural of quantum size, and the effects on performance of variations in loading.

First, let us briefly review the basic nature of the three service disciplines of interest in this section: the RR, FB_N, SJS, and FCFS disciplines. It is clear that each of the RR, FB_N, and SJF disciplines have the common objective of favoring units with short service times. The extent to which this favoritism is shown in each case will be the subject of the following examples. The SJF discipline is distinguished from the FB_N and RR disciplines in that the SJF discipline assumes a priori information on the service time required by new arrivals. Thus, we have:

a) the SJF discipline discriminating on the basis of a known "future" service requirement,

b) the FB_N discipline discriminating explicitly on the basis of past service,

c) the RR discipline making an implicit discrimination on the basis of past service,

d) and the FCFS system making no discrimination at all based on service requirements.

For our first examples we consider the variation of conditional waiting times for the RR and FB_N models with changes in loading. It is more convenient for the FB models in which $q \neq 0$ to display the waiting time in queue. This is quite simply obtained from the expressions for waiting time in the system by subtracting out the time t in the service facility. Thus we shall display

$$W_k = W(t) - t\,; \qquad (k-1)q < t \leqq kq \tag{39}$$

where W(t) is given by Equation (13) and Equation (22) for the RR and FB_N systems, respectively. Note that a broader class of service requirements are now included in Equation (39). Specifically, W_k now represents the waiting time in queue for all units whose service requirements are such that $(k-1)q < t \leqq kq$. Clearly, this is because all units in this class make the same number of "passes" in the RR system or ascend the same number of levels in the FB_N system.

Figure 6 presents curves for various values of k; i.e., the number of RR passes or the number of FB_N levels a unit whose service time is between (k-1)q and kq seconds must experience. The curves come from Equation (39) into which has been substituted Equations (13) and (22) for the RR and FB_N systems, respectively, with the values $\mu = 1.0$/second, $q = 0.5$ seconds, and $N = \infty$. The loading ρ is varied by allowing λ to vary from 0 to 1.0. Also included is the curve for the FCFS model whose waiting time in queue is obtained from Equation (18) by subtracting t.

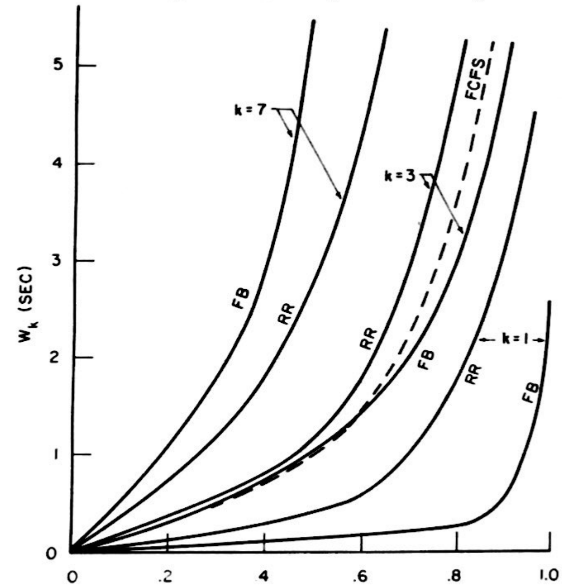

FIG. 6 COMPARISON OF ∞-LEVEL FB AND RR CONDITIONAL WAITING TIMES

The curves clearly show how units with shorter service requirements enjoy shorter average waits in both the RR and FB_∞ systems than in the FCFS system. This effect will be demonstrated further later on. Note also the comparison of the RR and FB_∞ disciplines that is inherent in Figure 6. The fact that the shorter service time units in the FB_∞ model do not have to wait behind the longer ones in the higher queues accounts for the better service they receive in the FB_∞ model. However, it is clear from the figure that this improvement is at the expense of the waiting times for the longer service time units. Thus, the RR system gives better service to the units with longer service requirements. Another way to view this comparison is to observe that the "variance" of the two sets of curves about their cross-over point ($k = 4$) is larger for the FB_∞ model than for the RR model.

We now investigate the variation of conditional waiting times (in queue) with quantum sizes in the RR and FB_N models. For this, we have Figures 7 and 8 from which several interesting observations can be made. The two figures refer to the same two equations mentioned above with the parameter values $\lambda = 0.5$/second, $\mu = 1.0$/second. (Figure 7 refers to the RR system and Figure 8 refers to the FB_∞ system.) In both figures we have plotted curves corresponding to units with service times of 0.5 and 2.0 seconds.

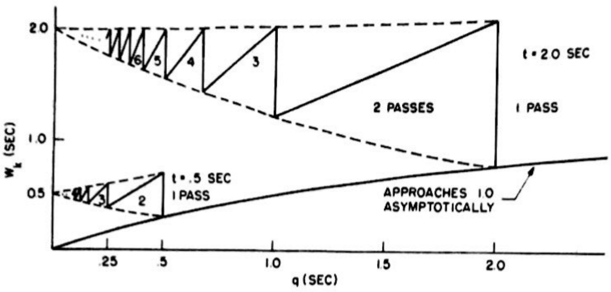

FIG. 7 RR CONDITIONAL WAITING TIMES
$\lambda = .5$/sec. $\mu = 1.0$/sec.

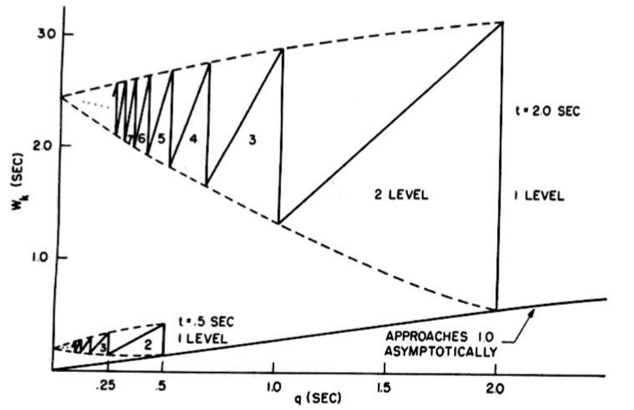

FIG. 8 INFINITE LEVEL FB WAITING TIMES
VS. QUANTUM SIZE
$\lambda = 0.5$/sec. $\mu = 1.0$/sec.

First of all, the jumps or discontinuities, occurring at the same points in both figures, are due to the decrease (looking from left to right) in the number of passes made in the RR system, and to a decrease in the number of levels required in the FB_∞ model. Take, for example, the points in Figures 7 and 8 corresponding to a unit requring 2.0 seconds of service when the quantum size is $2.0 + \epsilon$ where ϵ is very small. We see that the unit makes only one pass in the RR system and waits only in the first level of the FB_∞ system. However, the above remark changes to two passes and two levels when the quantum size is made to be $2.0 - \epsilon$. Since the waiting times are substantially different for one and two passes in the RR system and one and two levels in the FB_∞ system, we have the discontinuity in the limit as ϵ goes to zero. Of course, the above remarks apply to all sub-multiples of 2.0 and 0.5 seconds; i.e., to all q for which there is an integer n such that $nq = 2.0$ for the upper curves of Figures 7 and 8 and $nq = 0.5$ for the lower curves. As q goes to infinity the round-robin reduces to one pass, the FB_∞ system to one level, and both reduce to the FCFS system. Observe that all units, regardless of their service requirements have the same mean wait if they require but one pass in the RR system (or one level in the FB_∞ system); i.e., in the region where $q > t$ in Figures 7 and 8.

We now discuss in an informal way the reasons why the upper envelopes in Figure 7 (for the RR system) increase as q increases. First consider the processor-shared case; i.e., the limit as q goes to zero; we have subtracting t from Equation (17)

$$W_q(t) = \frac{\rho t}{1-\rho}$$

We want to compare this waiting time "in queue" with that of a FCFS system, viz.,

$$W_q(t) = \frac{\rho(1/\mu)}{1-\rho}$$

As noted earlier units requiring greater than average service ($t > 1/\mu$) do worse by sharing the processor than in the FCFS system, whereas for units requiring less than average service the opposite relationship exists. In the processor-shared case new arrivals immediately gain access to the processor and begin service, thus "slowing down" units already in the system. Now in this respect we observe two effects on the waiting time for a finite, non-zero quantum size. First, a given unit does not have to wait for (or be "slowed down" by) new arrivals on the given (tagged) unit's last round-robin pass. This effect causes the tagged unit's waiting time to decrease. Second, the units in the system at arrival of the tagged unit (which now become ahead of the tagged unit in the round-robin cycling) are potentially being allocated more service up to the tagged unit's last pass. For shorter than average service requirements (the 0.5 second example in Figure 7) we see that, on the average, the units ahead of the tagged unit will take greater advantage of this additional time than for units larger than average (the 2.0 second example). As can be observed in the figure the net effect, when considered along with the fact that the last pass leads to essentially zero service, produces an upward slope of upper envelopes which is less pronounced for the longer service time units.

Now consider the reason for the increasing slopes (as q increases) of the envelopes in Figure 8 for the FB_∞ system. For this, consider the example of a 2.0 second unit that requires just over one quantum in some model "A" and just over two quanta in some model "B". That is, model A has a quantum just less than 2.0 seconds and model B has a quantum just less than 1.0 second.

The 2.0 second unit must ascend two levels in model A and three in model B. Now the basic reason why the mean wait is shorter in model B than in model A, even though the number of levels has increased, is because the units ahead of the 2.0 second unit in model A are being allocated two quanta of 2.0 seconds each (4 seconds total), while in model B they are being allocated three quanta of 1.0 second each (3 seconds total). Thus, the units (ahead of the 2.0 second unit) requiring greater than 3 seconds are holding up the 2.0 second unit more in model A than in model B. As for the effects on new arrivals in models A and B we note from the second term of Equation (22) that since $(k-1)q = t$ is constant on each point of the upper envelopes the new arrival processing time is the same in both systems. Thus, the net effect is an increase in W_k. Of course, the fact that the average unit requires but 1.0 second of service explains why the effect is not more marked than it is.

Now consider for both Figures 7 and 8 the downward slope of the lower envelopes. A little reflection shows that the reason for the decrease in the waiting times stems from the necessity of processing new arrivals during the service time of the unit being considered. In other words, if a unit requires n passes (levels) in a given system, then the arrivals during the first (n-1) quanta of its service must be processed. Taking the 2.0 second unit as an example we see that as n increases and q decreases such that $nq = 2.0$ seconds (looking at the points on the lower envelope of the $t = 2.0$ second curve) we see that the product $(n-1)q$ increases. Thus, the increased arrival period implies an increase in the mean number of arrivals, which implies an increase in the minimum, mean waiting times as the number of levels increases (quantum decreases).

Finally, we look at the increase in waiting times as the quantum size varies between the discontinuities; i.e., as the quantum size varies without a change in the number of passes (levels). Although the curves in Figures 7 and 8 are drawn linear, the data showed a very slight downward convexity (dip). When the quantum increases but the priority (number of passes or levels necessary) does not, then it is clear that more time is being allotted to units ahead of the given unit whereas this unit does not need the additional time. Thus, its waiting time clearly increases.

In Figure 9 we have displayed the effect of a finite number of levels in the FB_N system. Specifically, we have plotted versus quantum size the waiting time of a unit requiring 2.0 seconds in a 4-level system (FB_4) with $1/\mu = 1.0$ seconds and $\rho = 0.5$. Clearly, the 2.0 second unit becomes a "background" (4^{th} level) unit just as soon as the quantum size reduces below 2/3 seconds. To the right of the line $q = 2/3$ seconds Figure 9 is identical to the upper curve of Figure 8. To the left of this line we observe the effect of gradually putting all units into the background as the quantum size decreases. The serrations are explained

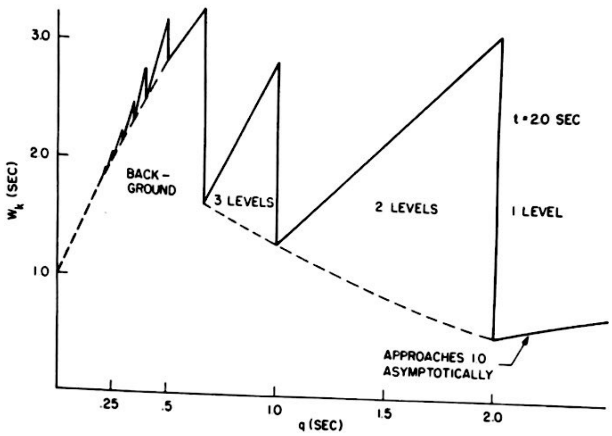

FIG. 9 4-LEVEL FB CONDITIONAL WAITING TIMES VS. QUANTUM SIZE
$\rho = 0.5$ $\mu = 1.0$/sec

as before, and as we explained in Theorem 5 the system becomes a conventional FCFS system in the limit as q goes to zero. It is interesting to observe from Figures 7-9 that there is an optimum RR and FB_N system for every unit with a given service requirement. Clearly, the optimum system is one with a quantum size just over the running time of the given unit. A reduction in this optimum causes an increase in the number of passes or levels, and an increase in this optimum implies giving more service to the units ahead of the unit for which the quantum size is optimum.

We now look at a comparison of the mean waiting times for the processor-shared system (the RR system with $q = 0$), the preemptive processor-shared system (the FB_∞ system with $q = 0$), and the shortest-job-first (SJF) system. In particular, the expressions for the waiting times given in Equations (17), (28), and (38) will be plotted versus loading and versus the service time required. Recall that in the RR^o (processor-shared) system we may view the current units in the system as sharing the processor. If there are n units in the system, then each is serviced at the same time but at $(1/n)^{th}$ the speed they would if they had the processor to themselves. In the FB^o (preemptive processor-shared) system this sharing occurs only between units having the same (highest) priority (i.e., the same amount of past service).

We have plotted the waiting times for all three disciplines versus loading (ρ) in Figure 10, and versus the service requirement t in Figure 11. The number in parentheses following the system designations on the curves represent the corresponding service times. Note in Figure 10 that the RR^o formula reduces to the FCFS formula ($\rho/1-\rho$) for $t = 1.0$ seconds. We observe in Figure 10 that the variance of the curves about the FCFS (or RR^o, $t = 1.0$ second) line is greater for the FB^o system than for the RR^o system. Of particular interest in Figure 11 are the cross-over points for small values of t which give those regions where one discipline improves over another. Note that Equation (17) is linear with respect to t and that Equations (28) and (38) become linear for large t.

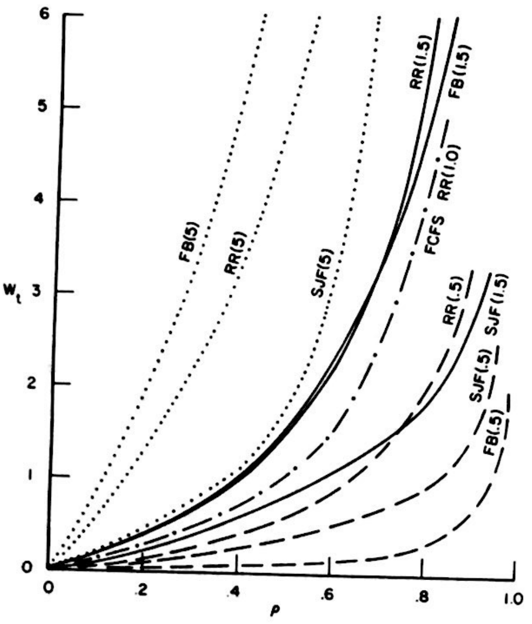

FIG. 10 COMPARISON OF SERVICE DISCIPLINES RR(q = 0), FB(q = 0, N = ∞), SJF, FCFS μ = 1.0/sec.

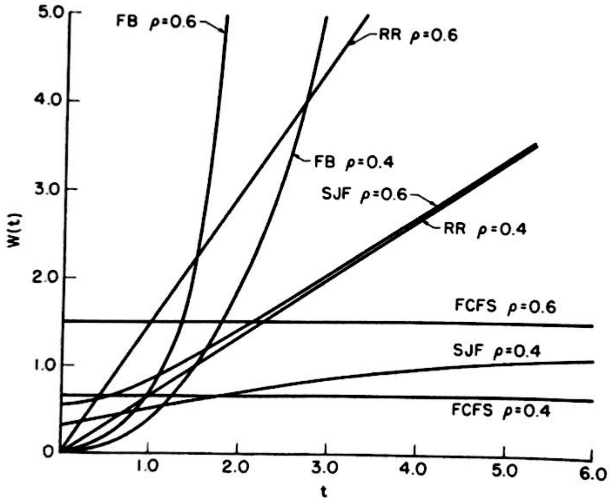

FIG. 11 THE FB(q = 0), RR(q = 0), SJF, and FCFS WAITING TIMES VS. SERVICE TIMES

We comment here that all of the queueing models considered obey the Conservation Law[12] which states that

$$\sum_{p=1}^{P} \int_0^{\infty} \rho_p(t)\, W_p(t)\, dt = \text{constant} \tag{39}$$

where we have broken the input population into P priority groups and where $\rho_p(t)\,dt$ is the fraction of time that the full processor† spends servicing units from priority group p whose total service time requirement lies between t and t+dt. Equation (39) indicates, regardless of the queueing discipline (under some very weak assumptions), that the superior treatment given certain units must result in inferior treatment to some other units. This effect is noticed in Figures 6, 10 and 11.

†Alternatively, we may think of $\rho_p(t)\,dt$ as the fraction of time that the partial processor spends on such units, weighted by the portion of the full processor which is giving service to such units.

VI. Summary

In this paper, we have studied the behavior of the average waiting time (conditioned on required service time and on priority) in a number of feedback queueing models of time-shared systems. The purpose of this study was to analyze certain specific models in order to better understand the way in which they manipulated the various customers' wait in system. All the models considered were quantum controlled, and the analysis was carried out for arbitrary quantum sizes. An especially interesting affect occurs when the quantum approaches zero and these results were elaborated upon.

The basic assumptions made were that the arrival and service processes were Markovian and that swap-time was zero. The effect of the swap-time assumption is to yield results which are ideal in the sense that the waiting times increase in all systems for non-zero swap-time.

This study has been one of analysis – not one of synthesis. Indeed, the general problem of finding optimum algorithms for operating time-shared systems has yet to be formulated, much less solved. We feel, however, that the various models studied here provide the system designer with a number of degrees of freedom with which to synthesize a satisfactory (albeit non-optimum, in some appropriately defined sense) time-shared processing system.

Appendix A

Proof of Theorem 2

We consider a unit (which we call the "tagged" unit) arriving at the RR system in equilibrium and assume a service requirement of t seconds. Defining k as the smallest integer such that $t < kq$ we address the problem of finding the tagged unit's average waiting in queue. To find the mean wait in system we simply add t to the waiting time in queue.

Assume that on arrival of the tagged unit there is one or no unit in service and n in the queue. We decompose the waiting time in queue into two parts, T_1 and T_2. T_1 corresponds to the time required to finish the unit, if any, in service (taking into account the possibility of its returning for more) plus the time required to process (not necessarily to completion) all arrivals during this time. T_2 corresponds to the time required to properly service the n units in the queue at arrival. Of course, both T_1 and T_2 must take into consideration the processing of all arrivals that occur in T_1 and T_2. Evidently, the mean waiting time in queue is

$$W_k = E(T_1) + E(T_2) \qquad (A.1)$$

The resequencing of events implicit in our definitions will clearly not affect the determination of W_k so long as all events are taken into account. This often-used "resequencing" approach is justified by the fact that the input process is time-homogeneous and statistically independent of the state of the system.

Now for $E(T_2)$ we use expected value arguments essentially the same as those used by Kleinrock[2] for the discrete system. Let y_i denote the time spent in queue on the i^{th} pass by the tagged unit. Since the tagged unit must make k passes we may write

$$E(T_2) = E\left\{\sum_{i=1}^{k} y_i\right\} = \sum_{i=1}^{k} E(y_i) \qquad (A.2)$$

Correspondingly, we define N_i as the mean number of units ahead of the tagged unit at the beginning of the i^{th} pass. We shall now develop a general expression for N_i. For $i > 1$, N_i will be composed of the mean number of those units of N_{i-1} whose service requirements exceed q seconds (we call these returning units), and the mean number of new arrivals that occur during the time interval $y_{i-1} + q$. (The q seconds is included because of the tagged unit's service following y_{i-1}.

From the memoryless property[1] of the exponential distribution we may observe that the probability δ with which a unit returns (requires more than q seconds of service) is independent of i and given by

$$\delta = \int_q^{\infty} \mu \epsilon^{-\mu\tau}\, d\tau = \epsilon^{-\mu q} \qquad (A.3)$$

Thus, we have

$$N_i = \delta N_{i-1} + \lambda [E(y_{i-1}) + q] \qquad (A.4)$$

But

$$E(y_{i-1}) = N_{i-1} E_1(\tau)$$

so upon substitution into Equation (A.4) we obtain

$$N_i = N_{i-1} [\delta + \lambda E_1(\tau)] + \lambda q \qquad (A.5)$$

For convenience we define

$$\beta = \delta + \lambda E_1(\tau) \qquad (A.6)$$

so that

$$N_i = \beta N_{i-1} + \lambda q \qquad (A.7)$$

Now solving this equation for N with the condition $N_1 = \bar{n} = E(n)$ yields

$$N_i = \beta^{i-1}\bar{n} + \lambda q \sum_{j=0}^{i-2} \beta^j ; \qquad i > 1 \qquad (A.8)$$

Using induction Equation (A.8) is easily established. From Equation (A.2) we may now write

$$E(T_2) = E_1(\tau) \sum_{i=1}^{k} N_i \qquad (A.9)$$

whereupon substitution of Equation (A.8) into Equation (A.9) yields, after carrying out the summations

$$E(T_2) = \frac{E_1(\tau)}{1-\beta}\left[\lambda k q + \left(\bar{n} - \frac{\lambda q}{1-\beta}\right)(1-\beta^k)\right] \qquad (A.10)$$

where, by evaluating Equation (A.6), we have

$$\beta = \rho + (1-\rho)\,\epsilon^{-\mu q}$$

Now in the RR and FB_N models we have assumed that no losses or "overhead" times exist in system operation, and in both models no advantage is taken of any a priori information concerning the nature of the new arrivals. Thus, it is not difficult to see that the average number of units in the queue for both the RR and FB_N systems is precisely the same as for the exponential FCFS (Erlang's) system. Thus, we may solve for $\bar{n}$ by using the corresponding result for the FCFS system which is given by[6]

$$\bar{n} = \frac{\rho^2(1/\mu)}{1-\rho}$$

Now using $E_1(\tau) = (1/\mu)[1-\epsilon^{-\mu q}]$ from Equation (12) we may render Equation (A.10) as

$$E(T_2) = \frac{(1/\mu)}{1-\rho}\left[\lambda k q + \left(\frac{\rho^2(1/\mu)}{1-\rho} - \frac{\lambda q}{1-\beta}\right)(1-\beta^k)\right] \qquad (A.11)$$

Turning now to $E(T_1)$ let W_o be the mean amount of time required to complete the quantum-service in progress at the time of arrival. Then $E(T_1)$ is equal to W_o plus the expected time to process the mean number of arrivals in W_o plus the time it takes to process the unit in service if it returns for more service. Here again, the processing referred to includes the processing of subsequent arrivals as for $E(T_2)$. The mean number of arrivals in W_o is given by λW_o. If we call σ the probability that the unit in service at arrival returns for more service we have

$$\bar{n}' = \sigma + \lambda W_o \qquad (A.12)$$

as the mean number of units (excluding $\bar{n}$) to service following W_o. The time to process the $\bar{n}'$ units can be calculated as for $E(T_2)$. We note, however, that these units are all "behind" the tagged unit and therefore will be provided with a maximum of only (k-1) quanta of service before the tagged unit receives its last quantum. Thus, we can proceed as before and form the sum

$$N_1 E_1(\tau) + [\delta N_1 + \lambda N_1 E_1(\tau)] E_1(\tau) + \cdots$$
$$+ [\delta N_{k-2} + \lambda N_{k-2} E_1(\tau)] E_1(\tau)$$

from which it is easy to establish by induction and by using $N_1 = \bar{n}'$

$$\sum_{i=1}^{k-1} N_i = \frac{1-\beta^{k-1}}{1-\beta}\,\bar{n}' \qquad (A.13)$$

Finally, therefore, we have

$$E(T_1) = W_o + [\sigma + \lambda W_o]\left[\frac{1-\beta^{k-1}}{1-\beta}\right] E_1(\tau) \qquad (A.14)$$

Using $E_1(\tau) = (1/\mu)[1-\epsilon^{-\mu q}]$ from Equation (12) this may be put into the form

$$E(T_1) = \frac{W_o}{1-\rho}[1-\rho\beta^{k-1}] + \frac{\sigma(1/\mu)}{1-\rho}[1-\beta^{k-1}] \qquad (A.15)$$

It remains to derive expressions for W_o and σ. To find W_o we shall follow Cobham[9] and observe the following. Given that a quantum-service of duration t is in progress at the time of the tagged unit's arrival, then from the point of view of the unit being served the expected time of arrival is simply $(t/2)$. We must now determine the probability $dC(t)$ of arriving when a quantum-service of duration t is in progress. For this Cobham writes

$$dC(t) = \lambda_q t\, dF_1(t) \qquad (A.16)$$

where $F_1(t)$ is the quantum-service distribution given by Equation (12) and λ_q represents the average arrival rate of quantum services. Now Equation (A.16) is based on a Poisson arrival mechanism of quantum-services; in our case unit arrivals are Poisson which gives rise to Poisson "bulk" arrivals of quantum-services. However, Equation (A.16) still applies since for our purposes only the randomness or Poisson nature of the arrival times is necessary for Equation (A.16). Since a unit requires a k^{th} pass (quantum-service) with probability $\epsilon^{-\mu(k-1)q}$ we see that

$$\lambda_q = \lambda \sum_{k=1}^{\infty} \epsilon^{-\mu(k-1)q} = \frac{\lambda}{1-\epsilon^{-\mu q}} \qquad (A.17)$$

Therefore, we obtain with Cobham

$$W_o = \int_o^{\infty} (t/2)\, dC(t) = \frac{\lambda/2}{1-\epsilon^{-\mu q}} E_1(t^2) \qquad (A.18)$$

To determine σ we find the probability that the tagged unit arrives and finds a program being served whose original service requirement was greater than q seconds. Now suppose that the tagged unit arrives when the service facility is busy and that the elapsed time of the program in service is τ; i.e., we know that $t > \tau$ where t is the original service requirement of the unit being served at the time of arrival. From the memoryless property of the exponential distribution we have

$$Pr\{t>q \mid t>\tau\} = \epsilon^{-\mu(q-\tau)}; \quad 0 \leq \tau \leq q \qquad (A.19)$$

But

$$Pr\{t > \tau\} = \epsilon^{-\mu\tau} \qquad (A.20)$$

so that given a unit in service at arrival

$$Pr\{t>q, t>\tau\} = Pr\{t>q \mid t>\tau\} Pr\{t>\tau\} = \epsilon^{-\mu q} \qquad (A.21)$$

which is independent of τ. Thus, $Pr\{t>q\} = \epsilon^{-\mu q}$. Since the probability that the service facility is busy is given by $\rho = \lambda/\mu$ we have

$$\sigma = \rho\, \epsilon^{-\mu q} \qquad (A.22)$$

Inserting Equations (A.18) and (A.22) into Equation (A.15) we get

$$E(T_1) = \frac{\frac{\lambda/2}{1-\epsilon^{-\mu q}} E_1(\tau^2)}{1-\rho}[1-\rho\beta^{k-1}] + \frac{\rho\epsilon^{-\mu q}}{1-\rho}[1-\beta^{k-1}]\,(1/\mu) \qquad (A.23)$$

where

$$\beta = \rho + (1-\rho)\, \epsilon^{-\mu q} \qquad (A.24)$$

Substituting Equations (A.23) and (A.11) into Equation (A.1) now yields

$$W(t) = \frac{\rho kq}{1-\rho} + \frac{(\lambda/2) E_1(\tau^2)}{1-\beta}[1-\rho\beta^{k-1}] + \frac{1}{1-\rho}\left[\frac{\rho^2}{1-\rho}(1/\mu) - \frac{\rho q}{1-\beta}\right][1-\beta^k] + \frac{\rho\epsilon^{-\mu q}}{1-\rho}(1/\mu)[1-\beta^{k-1}] \qquad (A.25)$$

which constitutes the result of Theorem 2 when the service time t is added. <u>Q.E.D.</u>

We may now produce the result for the processor-shared model of Theorem 3 by taking the limit of Equation (A.25) as q goes to zero. Since the waiting time is conditioned on the service required, we want to hold kq constant while allowing q to go to zero in Equation (A.25). Calling $kq = t$ let us first calculate

$$\lim_{q\to 0} \beta^k = \lim_{q\to 0} [\rho + (1-\rho)\delta]^k; \qquad \delta = \epsilon^{-\mu q}$$

With rearrangement we have

$$\beta^k = \sum_{i=0}^{k} \binom{k}{i} \delta^k [\rho(1-\delta)]^{k-i} = \delta^k + k\rho(1-\delta)\delta^{k-1} + \frac{k(k-1)}{2}\rho^2 (1-\delta)^2 \delta^{k-2} \ldots$$

Now $kq = t$ implies $\delta^k = \epsilon^{-\mu t}$ and approximating $(1-\epsilon^{-\mu q})$ by μq for $0<q\ll 1$ we have

$$\lim_{q\to 0} \beta^k = \epsilon^{-\mu t}\left[1 + \rho\mu t + \frac{(\rho\mu t)^2}{2!} + \ldots\right] = \epsilon^{-\mu t(1-\rho)}$$

With the same approximation it is easy to establish

$$\lim_{q\to 0} \frac{\lambda q}{1-\delta} = \rho\,; \qquad \lim_{q\to 0} \frac{E_1(\tau^2)}{1-\delta} = 0$$

so that on substitution of the above limits into Equation (A.25) we get

$$\lim_{q\to 0} W(t) = \frac{1}{1-\rho}\{\rho t - (\rho/\mu)[1-\epsilon^{-\mu t(1-\rho)}] + (\rho/\mu)[1-\epsilon^{-\mu t(1-\rho)}]\} = \frac{\rho t}{1-\rho}$$

which establishes Theorem 3 after adding the service requirement (t).

Appendix B

Proof of Theorem 5

For the proof of Theorem 5 we shall again resequence the events that must occur during the waiting time of an arriving unit so as to simplify the arguments necessary in determining this waiting time. We consider a unit (the tagged unit) arriving at the FB_N system in equilibrium, assume that its service requirement is t seconds, and define k as the smallest integer such that $kq > t$. We break up the waiting time in queue into two parts so that we may write

$$W_k = E(T_1) + E(T_2) \tag{B.1}$$

where T_1 is the time to complete the unit in service plus the time required to process the units which were in the first k queues at the time of arrival, and T_2 is the time to process all new arrivals that occur during the tagged unit's waiting time.

We shall approach the problem of determining W_k for $k < N$ by looking at a special two-level model which is equivalent in the sense of the waiting time we seek. Figure B.1 shows this equivalent two-level model. Note from the figure that arrivals requiring j quanta of service are (artificially) separated into j corresponding parts. The first k parts (or j parts if $k > j$) are combined into a single arrival unit to the high priority (lower level) queue. The remaining parts, if any, each constitute a unit arrival to the low priority queue. In this special model "feedback" is no longer explicit. Indeed, the quantum-at-a-time processing is no longer carried out by the server, but is implicit in the arrival processing mechanism instead. However, for the waiting times of high priority arrivals (requiring kq seconds or less) for which feedback does not exist anyway, it is clear that this artificial arrival mechanism has not changed anything. As can be observed, arrivals to the high priority queue are Poisson while arrivals to the low priority queue are Poisson in "bulk".

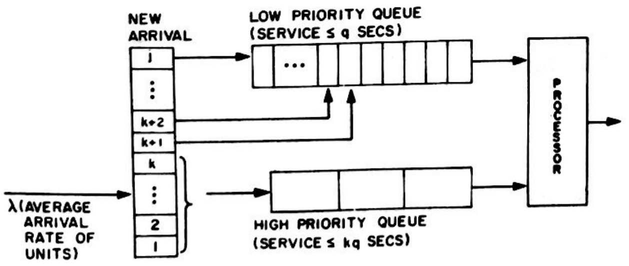

FIG. B.1 EQUIVALENT TWO-LEVEL MODEL WITHOUT FEEDBACK

From the above remarks we now make the simplifying observation that the time (T_1) to process the first k queues in the FB_N model $(N > k)$ and the unit in service at arrival is the same as the waiting time in the high priority queue of the special two-level model in Figure B.1. In both cases the tagged unit must wait through the processing of units being allocated kq seconds of service. It remains, therefore, to determine the high priority waiting time of the special two-level model. But for this statistic we may identify our special two-level model with the corresponding single-channel, head-of-the-line (two-level) priority model of Cobham.[9] The only difference we make between these two two-level models is that in the latter the arrival process to the low priority queue is assumed to be Poisson instead of Poisson in bulk. But for the average waiting time in the high priority queue it is unimportant whether or not the arrival process to the low priority queue is Poisson. Indeed, it can be shown that the high priority waiting time distribution depends on the low priority arrival process only through its average rate (see Reference 10, for example). Thus, using Cobham's result for the high priority average waiting time we have

$$E(T_1) = \frac{W_o}{1-\rho_1} \tag{B.2}$$

where ρ_1 is the utilization factor for the high priority queue and W_o is the average amount of time required to finish the unit being served at the time of arrival. In our case

$$\rho_1 = 1 - \lambda E_k(\tau) = 1-\rho(1-\epsilon^{-\mu kq}) \tag{B.3}$$

where $E_k(\tau)$ is given by Equation (24), and

$$W_o = \frac{\lambda_k}{2}\int_0^\infty \tau^2 \, dF_k(\tau) + \frac{\lambda_1}{2}\int_0^\infty \tau^2 \, dF_1(\tau) \tag{B.4}$$

where λ_k, λ_1 and $F_k(\tau)$, $F_1(\tau)$ represent, respectively, the average arrival rates and service time distributions for the high priority and low priority queues. The distributions are defined by Equation (23). Now since an arrival requires service at the low priority queue only if it requires in excess of kq seconds of service we have

$$\lambda_1 = \lambda, \quad \lambda_k = \lambda \sum_{j=k}^{\infty} \epsilon^{-\mu jq} = \frac{\lambda \epsilon^{-\mu kq}}{1-\epsilon^{-\mu q}} = \gamma_k \lambda \tag{B.5}$$

Thus,

$$E(T_1) = \frac{(\lambda/2)[E_k(\tau^2)+\gamma_k E_1(\tau^2)]}{1-\rho(1-\epsilon^{-\mu kq})} \tag{B.6}$$

To calculate $E(T_2)$ we now return to the original FB_N model. We observe that the average number of arrivals in W_k must be based on $W_k + (k-1)q$ since the tagged unit received $(k-1)q$ seconds of service before reaching the k^{th} ($< N^{th}$) queue. Clearly, each of the new arrivals must be allocated (k-1) quanta of service of which $E_{k-1}(\tau)$ is the average amount taken. Thus,

$$E(T_2) = \lambda[W_k+(k-1)q] \, E_{k-1}(\tau) \tag{B.7}$$

Finally, therefore,

$$W_k = \lambda[W_k + (k-1)q]\,E_{k-1}(\tau) + \frac{(\lambda/2)[E_k(\tau^2) + \gamma_k E_1(\tau^2)]}{1-\rho(1-\epsilon^{-\mu kq})};$$

$$1 \le k \le N-1 \qquad \text{(B.8)}$$

Solving for W_k and substituting for $E_{k-1}(\tau)$ we get

$$W_k = \frac{(\lambda/2)[E_k(\tau^2) + \gamma_k E_1(\tau^2)]}{[1-\rho(1-\epsilon^{-\mu kq})][1-\rho(1-\epsilon^{-\mu(k-1)q})]} +$$

$$\frac{\rho(1-\epsilon^{-\mu(k-1)q})}{1-\rho(1-\epsilon^{-\mu(k-1)q})}(k-1)q\ ;\ 1 \le k \le N-1 \qquad \text{(B.9)}$$

Adding t to Equation (B.9) now produces Equation (22a) of Theorem 5.

Finally, for $k > N-1$ we may simplify matters by observing that all units in the system at the time of arrival must be served to completion before the tagged unit comes to the service point for the N^{th} time. Thus for $E(T_1)$ we may use the result for the waiting time in queue for the FCFS system. In particular, from Equation (18) we have (subtracting the time t in the server)

$$E(T_1) = \frac{\rho(1/\mu)}{1-\rho} \qquad \text{(B.10)}$$

Now the period during which we must allow for new arrivals is again $W_k + (k-1)q$. Because of the nature of the N^{th} queue each of these new arrivals will be allocated $(N-1)q$ seconds of service. Thus

$$E(T_2) = \lambda[W_k + (k-1)q]\,E_{N-1}(\tau) \qquad \text{(B.11)}$$

Adding Equations (B.10) and (B.11) and solving for W_k now yields

$$W_k = \frac{\rho(1/\mu)}{(1-\rho)[1-\rho(1-\epsilon^{-\mu(N-1)q})]} +$$

$$\frac{\rho(1-\epsilon^{-\mu(N-1)q})}{1-\rho(1-\epsilon^{-\mu(N-1)q})}(k-1)q\ ;\ \ k \ge N \qquad \text{(B.12)}$$

Adding t to Equation (B.12) now establishes Equation (22b) of Theorem 5 and completes the proof of Theorem 5. Q.E.D.

Appendix C

Proof of Theorem 6

To find conditional waiting times for the priority FB_∞ model we shall employ a method that is basically similar to that used in the proof of Theorem 5. We consider the mean waiting time in queue W_p^k of a unit entering the system at the p^{th} level and requiring service up to the k^{th} level $(p \le k)$.

First, we shall indicate which units, in the system at arrival, must precede the tagged unit's quantum-service at the k^{th} level and how much service they are entitled to. For the present we shall assume $k > p+1$. From the description of the priority FB_∞ service discipline we see that all units at the $j^{th} \le p^{th}$ level queues will be allocated service, as required, up to and including the k^{th} level, and all units at the j^{th} $(p < j \le k-1)$ level will be allocated service up to and including the $(k-1)^{st}$ level. Now the processing of new arrivals during W_p^k will be as follows. New arrivals at the $j^{th} \le (p-1)^{st}$ level will be given service up to and including the k^{th} level and new arrivals at levels p through $(k-1)$ will be given service up to and including the $(k-1)^{st}$ queue.

As in the proof of Theorem 5 we now construct a modified, two-level model which is equivalent to the original one in terms of the waiting time of a p^{th} priority unit requiring service up to the k^{th} level. The high priority queue of the two-level model will consist of priority r units, where $1 \le r \le p$, being allocated $k-r+1$ quanta of service, and units of priorities $(p+1)$ through $(k-1)$ being allocated $(k-r)$ quanta of service. The low priority queue of the two-level model will consist of all priority r units, with $1 \le r \le p$, that required in excess of $(k-r+1)$ quanta of service, all priority r units, with $p+1 \le r \le k-1$, that required in excess of $(k-r)$ quanta of service, and all units which arrive at level k or above. Now the probability that a unit requires greater than kq seconds of service is simply $\epsilon^{-\mu kq}$. Thus, the total arrival rate of units to the $j^{th} > k^{th}$ level queue is given by,

$$\Lambda_j = \sum_{r=1}^{j} \lambda_r\, \epsilon^{-\mu(j-r)q} \qquad \text{(C.1)}$$

We see that $\sum_{j=k+1}^{\infty} \Lambda_j$ represents the contribution, based on arrivals initially to all levels, to the low priority queue from all levels beyond the k^{th}. However, for the total low priority arrival rate we must also take into account those units of priority r $(p < r < k)$ that require greater than $(k-r)$ quanta of service; these units will be behind the tagged unit when the latter receives its last quantum of service in the k^{th} queue. This contribution (at the k^{th} level) to the low priority queue of the modified model is given by

$$\Lambda_{pk}^{*} = \sum_{r=p+1}^{k-1} \lambda_r\, \epsilon^{-\mu(k-r)q} \qquad \text{(C.2)}$$

We shall define $\Lambda_{pk}^{*} = 0$ for $k=p$ or $p+1$. Finally, therefore,

$$\Lambda_{pk} = \Lambda_{pk}^{*} + \sum_{j=k+1}^{\infty} \Lambda_j \qquad \text{(C.3)}$$

Recall that we need to consider only one low priority queue because all units arriving to the low priority queue receive but one quantum of service at a time. Clearly, the total arrival rate Λ_H to the high priority queue will be simply,

$$\Lambda_H = \sum_{r=1}^{k-1} \lambda_r \qquad \text{(C.4)}$$

In comparing Equations (C.3) and (C.4) note particularly that arrivals to the high priority queue are units taking up to and including k or $(k-1)$ quanta, but that arrivals to the low priority queue are units (irrespective of their original level of entrance) that take up to and including only q seconds of service (see Figure B.1).

We are now in position to calculate W_p^k. Let us first assume that $k > p+1$. Now considering, as in the proof of Theorem 5, the high priority queue of the modified (two-level) model as the higher priority in a two-level conventional priority model, we may again apply Cobham's analysis. Accordingly, we divide the waiting time W_p^k into two intervals T_1 and T_2. T_1 is the time to process the high priority units in the system at the time of arrival and T_2 is the time required to service the new arrivals occurring in $W_p^k + (k-p)q$. Now for the expected value of T_1 we use Cobham's result as given below.

$$E(T_1) = \frac{W_o}{1-\rho_{pk}} \tag{C.5}$$

where W_o is the expected time to complete the unit in service at arrival and ρ_{pk} is the utilization factor for the high priority queue. To find ρ_{pk} we first write the mean service time $E_{pk}(\tau)$ of a unit in the high priority queue of the two-level model. From earlier definitions we have

$$E_{pk}(\tau) = \frac{1}{\Lambda_H}\left[\sum_{r=1}^{p} \lambda_r E_{k-r+1}(\tau) + \sum_{r=p+1}^{k-1} \lambda_r E_{k-r}(\tau)\right] \tag{C.6}$$

From the above it is clear that

$$\rho_{pk} = \Lambda_H E_{pk}(\tau) = \sum_{r=1}^{p} \lambda_r E_{k-r+1}(\tau) + \sum_{r=p+1}^{k-1} \lambda_r E_{k-r}(\tau); \qquad k > p+1 \tag{C.7}$$

Since it is clear that the second term must be omitted for $k = p$ or $p+1$ we have established Equation (32). For W_o we take one-half the weighted sum of the second moments of the high and low priority service time distributions according to the two-level model. Thus,

$$W_o = \frac{1}{2}\left[\sum_{r=1}^{p} \lambda_r E_{k-r+1}(\tau^2) + \sum_{r=p+1}^{k-1} \lambda_r E_{k-r}(\tau^2) + \Lambda_{pk} E_1(\tau^2)\right] \tag{C.8}$$

Here again, the second term must be omitted for $k = p$ or $p+1$, so that in conjunction with Equation (C.3) we have established Equation (33). Thus, Equation (C.5) is determined. Now for $E(T_2)$ we reason as before to obtain, according to the present model

$$E(T_2) = \left[W_p^k + (k-p)q\right]\left[\sum_{r=1}^{p-1} \lambda_r E_{k-r+1}(\tau) + \sum_{r=p}^{k-1} \lambda_r E_{k-r}(\tau)\right] \tag{C.9}$$

Substituting Equations (C.5) and (C.9) into the relation

$$W_p^k = E(T_1) + E(T_2) \tag{C.10}$$

we get

$$W_p^k = \frac{W_o}{(1-\rho_{pk})\left[1 - \sum_{r=1}^{p-1} \lambda_r E_{k-r+1}(\tau) - \sum_{r=p}^{k-1} \lambda_r E_{k-r}(\tau)\right]} + \frac{\sum_{r=1}^{p-1} \lambda_r E_{k-r+1}(\tau) + \sum_{r=p}^{k-1} \lambda_r E_{k-r}(\tau)}{1 - \sum_{r=1}^{p-1} \lambda_r E_{k-r+1}(\tau) - \sum_{r=p}^{k-1} \lambda_r E_{k-r}(\tau)}\,(k-p)q$$

from which Equation (31) follows when we observe

$$\sum_{r=1}^{p-1} \lambda_r E_{k-r+1}(\tau) + \sum_{r=p}^{k-1} \lambda_r E_{k-r}(\tau) = \rho_{pk} - \rho_p \epsilon^{-\mu(k-p)q}$$

where

$$\rho_p = \lambda_p E_1(\tau)$$

Q.E.D.

References

1. Feller, ~~William~~, An Introduction to Probability Theory and its Applications, Wiley, New York, 1957.
2. Kleinrock, L., "Analysis of a Time-Shared Processor," Naval Research Logistics Quarterly, Vol. 11, No. 10, pp. 59-73, March 1964.
3. Kleinrock, L., "Time-Shared Systems; A Theoretical Treatment," to be published.
4. Coffman, E.G., "Stochastic Models of Multiple and Time-Shared Computer Operation," University of California at Los Angeles, Dept. of Engineering Report No. 66-38, June 1966.
5. Schrage, L.E., "The Queue M/G/1 with Feedback to Lower Priority Queues," Management Science, to be published.
6. Takacs, Lajos, Introduction to the Theory of Queues, Oxford University Press, New York, 1962.
7. Phipps, T.E., "Machine Repair as a Priority Waiting-Line Problem," Operations Research, Vol. 9, pp. 732-742, September-October 1961.
8. Scherr, A.L., An Analysis of Time-Shared Computer Systems, Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, Mass., June 1965.
9. Cobham, A., "Priority Assignment in Waiting-Line Problems," Operations Research, Vol. 2, pp. 70-76, February 1954.
10. Miller, L.W., and L.E. Schrage, "The Queue M/G/1 with the Shortest Remaining Processing Time Discipline," Rand Corp., Report P-3263, November 1965.

References (Continued)

11. Krishnamoorthi, B. and R.C. Wood, "Time-Shared Computer Operations with Both Interarrival and Service Times Exponential," Journal of the ACM, Vol. 13, pp. 317-338, July 1966.

12. Kleinrock, L., "A Conservation Law for a Wide Class of Queueing Disciplines," Naval Research Logistics Quarterly, Vol. 12, No. 2, June 1965.