TIME DEPENDENT PRIORITY QUEUES

Leonard Kleinrock

University of California, Los Angeles, California

and

Roy P. Finkelstein

Stanford Research Institute, Menlo Park, California

Considered is a class of rth order delay-dependent priority queuing disciplines in which a customer from the pth priority group, who arrives at time T, has a priority $q_p(t)$ at time t given by $q_p(t) = b_p(t-T)^r$. The main result states that the expected wait on queue for p-type customers in an rth order system with parameter set $\{b_p\}$ is identical to the wait in any other such system, say one of order r' with parameter set $\{b_p'\}$ if these parameters are chosen in the proper manner. From this, using the results due to KLEINROCK for the first order systems, we obtain the expected wait on queue, conditioned on the priority groups, for any rth order system. This class of queuing disciplines ranges from COBHAM's fixed priority system (for $r \to 0$) to the first-come–first-served system (for $r \to \infty$). For the case of two priority groups the set $\{b_p\}$ is chosen so as to minimize a class of delay-dependent cost functions. Results from a computer simulation are given to display the behavior of the waiting time variance.

AMONG those priority queuing systems considered in the past is the delay-dependent queue discipline studied by KLEINROCK.[1] In this discipline, the priority of a customer is a linearly increasing function of the amount of time that he has spent in the system. The rate at which priority increases is given by a parameter assigned to a customer's priority class. This set of parameters provides the system designer with a number of degrees of freedom with which to manipulate a customer's average waiting time.

A natural extension to the delay-dependent discipline is one in which a customer's priority increases in proportion to some arbitrary power of his elapsed time, rather than the first power as in reference 1. It is to this generalized delay-dependent priority system that we address ourselves in this paper.

THE MODEL

As IN reference 1, we consider a total of P different priority groups. Units from group $p(p = 1, 2, \cdots, P)$ arrive in a Poisson stream at rate λ_p units/

sec to a single-server system; each unit from priority class p has a required service time selected from an exponential distribution with mean $1/\mu_p$. We define*

$$\lambda = \sum_{p=1}^{p=P} \lambda_p, \tag{1}$$

$$1/\mu = \sum_{p=1}^{p=P} \lambda_p/(\lambda\mu_p), \tag{2}$$

$$\rho_p = \lambda_p/\mu_p, \tag{3}$$

$$\rho = \lambda/\mu = \sum_{p=1}^{p=P} \rho_p, \tag{4}$$

$$W_0 = \sum_{p=1}^{p=P} \rho_p/\mu_p. \tag{5}$$

When a customer from the pth priority group enters the queue at time T (say), he is assigned a number b_p, where †

$$0 \leq b_1 \leq b_2 \leq \cdots \leq b_P. \tag{6}$$

We define, for any nonnegative number r, an rth *order delay dependent priority discipline* as one which calculates the priority $q_p^{(r)}(t)$ at time t associated with a customer arriving at time T as follows

$$q_p^{(r)}(t) = (t-T)^r b_p, \tag{7}$$

where t ranges from T until the time at which this unit's service is completed. Whenever the service facility is ready for a new unit, that unit with the highest instantaneous priority $q_p^{(r)}(t)$ is then taken into service. Whenever a tie for the highest priority occurs, the tie is broken by a first-come–first-served rule. Contrary to the usual convention, a unit with priority $q(t)$ is given preferential treatment over a unit with priority $q'(t)$ where $q(t) > q'(t)$. We note that units from higher priority groups gain priority at a faster rate (b_p) than those from lower priority groups.

We further define

$W_p^{(r)}$ = Expected value of the time spent in the queue of an rth order system for a unit from group p (steady-state waiting time).

The coupling between units (customers) of different priority classes is illustrated in Fig. 1. A unit from priority group p_1 arrives at time T_1 and a unit from priority group p_2 arrives at time T_2. Both units gain priority proportional to the rth power of the time they spend in the system; however, the unit from group p_2, being from a higher priority class $(p_2 > p_1)$, has a larger proportionality constant, $b_{p_2} > b_{p_1}$. In this example, if the service facility becomes free between T_1 and T_0, the unit from group p_1

---

* Note that W_0 is the expected time to complete service on the unit found in the service facility (*see* COBHAM[2]).

† The equality between adjacent b_p is allowed for completeness. However it is clear that any priority groups with identical values of b_p can be grouped together into a single group.

would be taken into service before the unit from group p_2; but if the facility does not become free until after T_0, the reverse is true.   Thus we see that it is possible for units to change their relative positions in the queue. It should be noted that there can be at most only one interaction between any two units.

<div align="center">EXPECTED WAITING TIME</div>

CONSIDER two delay-dependent priority systems, one of order r with a set of parameters $\{b_p\}$ and the other of order r' with parameter set $\{b_p'\}$. In the Appendix, we prove the following theorem.
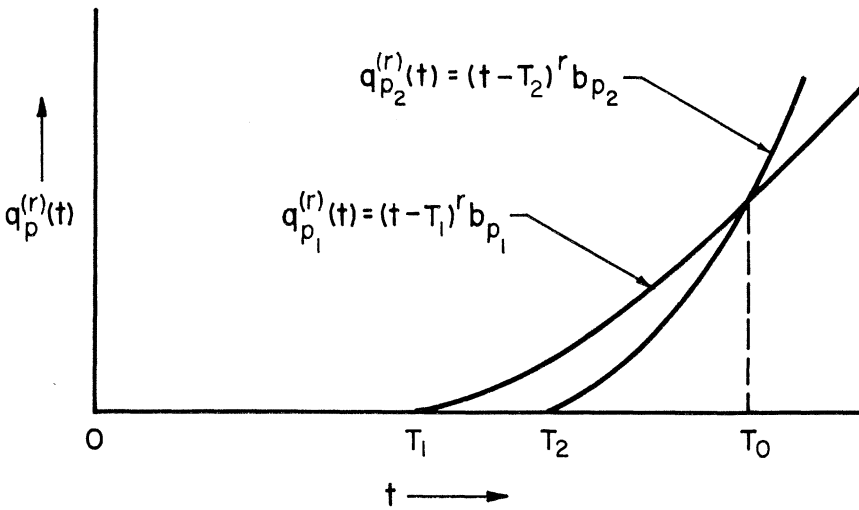
Fig. 1.   Coupling of different priority units.

THEOREM 1.   *If we choose*

$$(b_p/b_{p+1})^{1/r} = (b_p'/b_{p+1}')^{1/r'}, \qquad (p=1, 2, \cdots, P-1) \quad (8)$$

then
$$W_p^{(r)} = W_p^{(r')}. \tag{9}$$

This main theorem states that *all* rth order systems may be characterized (with respect to average waiting times) by an r_0th order system (for any $r_0 > 0$) through a suitable change of parameters as given by equation (8).   The case for $r_0 = 1$ has already been treated by Kleinrock[1] and so, in order to obtain $W_p^{(r)}$ we appeal to his results and obtain the following two theorems.

THEOREM 2.   *For an rth order delay dependent priority system without pre-emption, and $0 \leq \rho < 1$,*

$$W_p^{(r)} = \{[W_0/(1-\rho)] - \sum_{i=1}^{p-1} \rho_i W_i [1-(b_i/b_p)^{1/r}]\}/$$
$$\{1 - \sum_{i=p+1}^{i=P} \rho_i [1-(b_p/b_i)^{1/r}]\}. \quad (p=1,2,\cdots,P) \quad (10)$$

A solution to this type of recursive equation may be found in reference 1.

THEOREM 3. *For an rth order delay-dependent priority system with pre-emption (of the preemptive resume type), and for $0 \leqq \rho < 1$,*

$$W_p^{(r)} = \{[W_0/(1-\rho)] + \sum_{i=p+1}^{i=P} (\rho_i/\mu_p)[1-(b_p/b_i)^{1/r}]$$
$$- \sum_{i=1}^{p-1} \rho_i(W_i+1/\mu_i)[1-(b_i/b_p)^{1/r}]\}/ \quad (11)$$
$$\{1 - \sum_{i=p+1}^{i=P} \rho_i [1-(b_p/b_i)^{1/r}]\}.$$
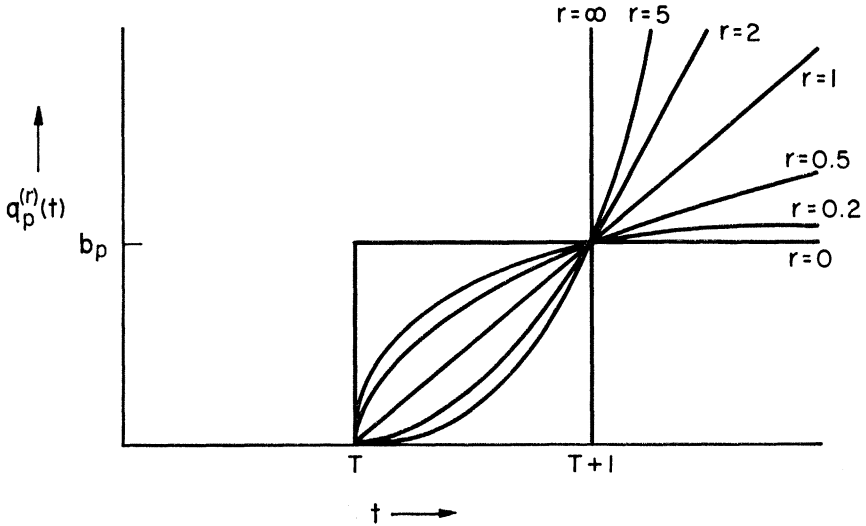
Fig. 2. $q_p^{(r)}(t)$ for several r.

For both cases we have (from LITTLE[3]) that the expected number $n_p^{(r)}$ of type p customers present in the queue is given by

$$n_p^{(r)} = \lambda_p W_p^{(r)}. \quad (12)$$

Curves of $W_p^{(1)}$ may be found in reference 1.

VARIATION OF $W_p^{(r)}$ WITH r; LIMITING CASES ($r\to0$, $r\to\infty$)

IN FIG. 2 we show $q_p^{(r)}(t)$ for a customer arriving at time T_1, with r as a parameter. We see that

$$\lim_{r\to0} q_p^{(r)}(t) = b_p\, u(t-T),$$

where $u(t - T)$ is the unit step function occurring at time T. Thus, an entering customer from group p is assigned a fixed value of priority equal to b_p. This is the fixed priority system studied by Cobham.[2] Moreover, as $r \to \infty$, $q_p^{(r)}(t)$ becomes a step function of infinite height at time $T+1$. Thus, units that have been in the system for more than one second have infinite priority and those that have been in the system for less than one second have zero priority. Remembering that a first-come–first-served criterion is used to break a tie, the limit as r approaches infinity is a strict first-come–first-served system.

These two limiting cases can also be obtained by taking the limit of $W_p^{(r)}$. We demonstrate this for the nonpreemptive case only. From equation (10), for $b_p < b_{p+1} (p=1, 2, \cdots, P-1)$,

$$\lim_{r \to 0} W_p^{(r)} = \lim_{(b_i/b_{i+1})^{1/r} \to 0} W_p^{(r)}$$

$$= [W_0/(1-\rho) - \sum_{i=1}^{p-1} \rho_i W_i]/(1 - \sum_{i=p+1}^{i=P} \rho_i).$$

Solving this last set of recursive equations yields

$$\lim_{r \to 0} W_p^{(r)} = W_0/(1 - \sum_{i=p}^{i=P} \rho_i)(1 - \sum_{i=p+1}^{i=P} \rho_i), \qquad (13)$$

which is the same result obtained by Cobham.[2] Equation (11) reduces to the result of WHITE AND CHRISTIE[4] for fixed priorities with preemption. Also,

$$\lim_{r \to 0} W_p^{(r)} = \lim_{(b_i/b_{i+1})^{1/r} \to 1} W_p^{(r)}.$$

Both Equations (10) and (11) then give

$$\lim_{r \to \infty} W_p^{(r)} = W_0/(1-\rho), \qquad (14)$$

which is the well-known result for a strict first-come–first-served discipline.

We now consider $\{b_p\}$ to be fixed and display the dependence of $W_p^{(r)}$ on r. As discussed above, as $r \to 0$ we obtain a fixed priority system[4] and as $r \to \infty$ we obtain the first-come–first-served system. For $r=1$ we have the first order delay dependent system.[1] In Fig. 3 we illustrate the general behavior of the expected wait on queue as we vary our priority discipline over the class of rth order systems. We show only the nonpreemptive case (Theorem 2) with $P=5$, $b_p/b_{p+1}=\frac{1}{2}$, $\rho_p=\rho/5$, for $p=1, 2, \cdots, 5$, $\rho=0.95$, $W_0=1$, and $\mu_p=\mu$.

As shown by Kleinrock,[5] a conservation law exists for a wide class priority queuing disciplines. This law states that all disciplines in this class conserve the following weighting of the average waiting times:

$$\sum_{p=1}^{p=P} (\rho_p/\rho) W_p = V_1/(1-\rho), \qquad (15)$$

where

W_p = expected wait in queue for member from the pth group,

$$V_1 = \tfrac{1}{2} \sum_{p=1}^{p=P} \lambda_p \, E(t_p^{\,2}),$$

$E(t_p^{\,2}) =$ second moment of the (arbitrary) service time distribution
(whose mean is $1/\mu_p$) for group p.

Equation (15), for exponential service times is merely the expression given in equation (14), namely, the average wait for a first-come–first-served system. This function is shown as the dashed line in Fig. 3, and demonstrates what is meant by the conservation law for this particular class of rth order delay-dependent priority systems. The wide dispersion of $W_p^{(r)}$
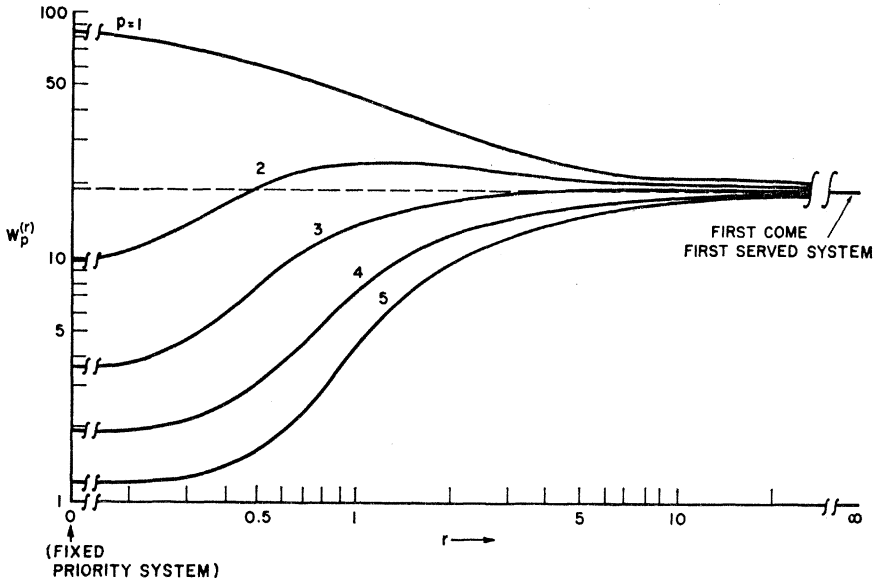

Fig. 3. $W_p^{(r)}$ vs r $(b_p/b_{p+1}=0.5,\ \rho=0.95)$.

among the priority groups shown in Fig. 3 is due to the large value of $\rho(\,=0.95)$, which causes considerable interaction among conflicting arrivals. For smaller values of ρ, the dispersion is not nearly as great. However, as the essence of Theorem 1 shows, the relative waiting times can be adjusted by varying r for a given $\{b_p\}$; moreover for a given r, variation of the $\{b_p\}$ accomplishes the same adjustment of relative waiting times [*see* equation (8)].

It is interesting to note that the class of rth order delay-dependent priority systems covers the spectrum from that queuing discipline that separates priority groups to the greatest possible extent (i.e., the fixed priority system) to the discipline that does not separate them at all (i.e., the first-come–first-served system).

COST FUNCTIONS

In this section we discuss the manner in which the parameter set $\{b_p\}$ may be chosen so as to minimize an appropriately defined cost.

We consider the following cost function

$$C = \sum_{p=1}^{p=P} C_p \, (\lambda_p \, W_p^{(r)})[W_p^{(r)}]^m, \tag{16}$$

where C is the average cost rate to the system and C_p is an appropriate cost rate for the pth group (the dimensions of C_p change as m changes).

Below we consider only $P = 2$ for simplicity. We then have

$$C = C_1 \, \lambda_1 [W_1^{(r)}]^{m+1} + C_2 \, \lambda_2 [W_2^{(r)}]^{m+1}. \tag{17}$$

For the nonpreemptive case, we have, from equation (10),

$$W_1^{(r)} = W_0/(1-\rho)\{1 - \rho_2[1 - (b_1/b_2)^{1/r}]\}, \tag{18}$$

$$W_2^{(r)} = W_0\{1 - \rho[1 - (b_1/b_2)^{1/r}]\}/(1-\rho)\{1 - \rho[1 - (b_1/b_2)^{1/r}]\}. \tag{19}$$

We may then write equation (17) as

$$C = (u + s[1 - \rho x]^{m+1})/[1 - \rho_2 x]^{m+1}, \tag{20}$$

where the following substitutions have been made

$$u = C_1 \, \lambda_1 [W_0/(1-\rho)]^{m+1},$$

$$s = C_2 \, \lambda_2 [W_\bullet/(1-\rho)]^{m+1},$$

$$x = 1 - (b_1/b_2)^{1/r}.$$

Note that only the ratio b_1/b_2 affects the optimization. From $dC/dx = 0$, we get

$$u\rho_2 - s\rho_1(1 - \rho x)^m = 0;$$

the solution to this equation is then

$$x = (1/\rho)[1 - (u\rho_2/s\rho_1)^{1/m}].$$

Thus

$$(b_1/b_2)^{1/r} = (1/\rho)[(\mu_1 C_1/\mu_2 C_2)^{1/m} - (1-\rho)], \tag{21}$$

provided

$$0 \leqq (b_1/b_2)^{1/r} < 1. \tag{22}$$

Equation (22) results in the two constraints

$$\mu_1 \, C_1 < \mu_2 \, C_2,$$

and

$$\rho \geqq \rho_0 = 1 - (\mu_1 C_1/\mu_2 C_2)^{1/m}.$$

Summarizing, we find for $P = 2$ and $\mu_1 \, C_1 < \mu_2 \, C_2$, choosing

$$(b_1/b_2)^{1/r} = \begin{cases} 0 & \rho < \rho_0 \\ (1/\rho) \cdot [(\mu_1 \, C_1/\mu_2 \, C_2)^{1/m} - (1-\rho)] & \rho > \rho_0 \end{cases} \tag{23}$$

minimizes the cost of the system where the cost is given by equation (17). Figure 4 shows a family of curves of $(b_1/b_2)^{1/r}$ vs. ρ with m as a parameter. For this figure $\mu_1 C_1/\mu_2 C_2$ was arbitrarily set equal to 0.5.

Note that, for $r > 0$,

$$\lim_{m \to 0} (b_1/b_2)^{1/r} = 0 \qquad \text{for all} \quad \rho,$$

which is the fixed priority system. This checks with a result of FIFE,[6] which states for a cost proportional to the number of units in the system the best possible priority scheme is a fixed one where the priority classes are
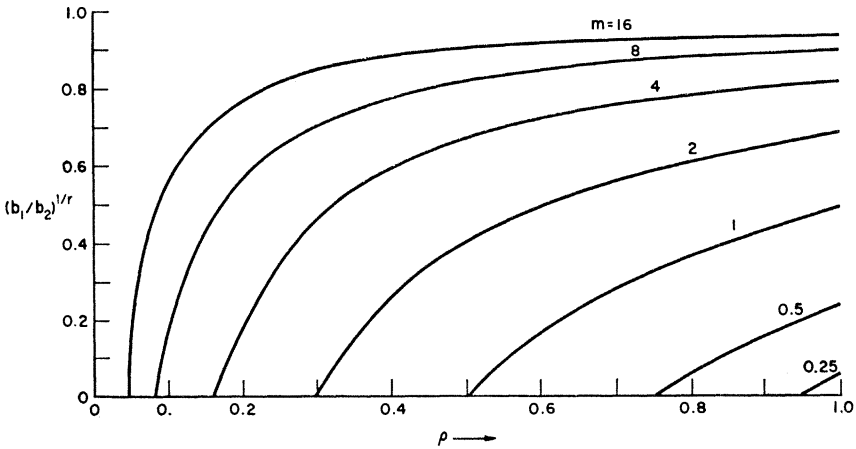

Fig. 4. Optimal values for $(b_1/b_2)^{1/r}$ as a function of ρ for different m.

ordered by the value of their $\mu_p C_p$ product. Furthermore

$$\lim_{m \to \infty} (b_1/b_2)^{1/r} = 1,$$

which is a first-come–first-served system.

For the case of finite m, there is some value $\rho = \rho_0$, $(0 < \rho_0 < 1)$ where there is a transition from the fixed priority system to one in which the priorities are time dependent (*see*, for example, Fig. 4). It is for $\rho_0 < \rho < 1$ that the rth order delay-dependent scheme is shown to be better than the fixed or first-come–first-served priority systems although it has not been proven that this is the best possible way of assigning priorities.

SIMULATION

WHILE expressions for expected waiting times in the first order delay-dependent, and thus any rth-order system have been obtained, it is interesting to study the distribution, particularly the variance, in more detail. To this end, a simulation was run for various values of ρ and r and $P = 5$, $b_p/b_{p+1} = 0.5$

and $\mu_p = \mu$. Figure 5 shows a sample plot of the waiting-time distribution
for the case of $\rho = 0.9$ and $r = 4$. It can be seen that the higher priority
classes tend to have a larger number of units with small waiting times and
the lower classes have more with larger waiting times. The crossover points
tend to occur at a time less than the average waiting time of the classes.
It should also be noted that the distribution tends to look exponential
although it is extremely difficult to say anything quantitative in this regard.

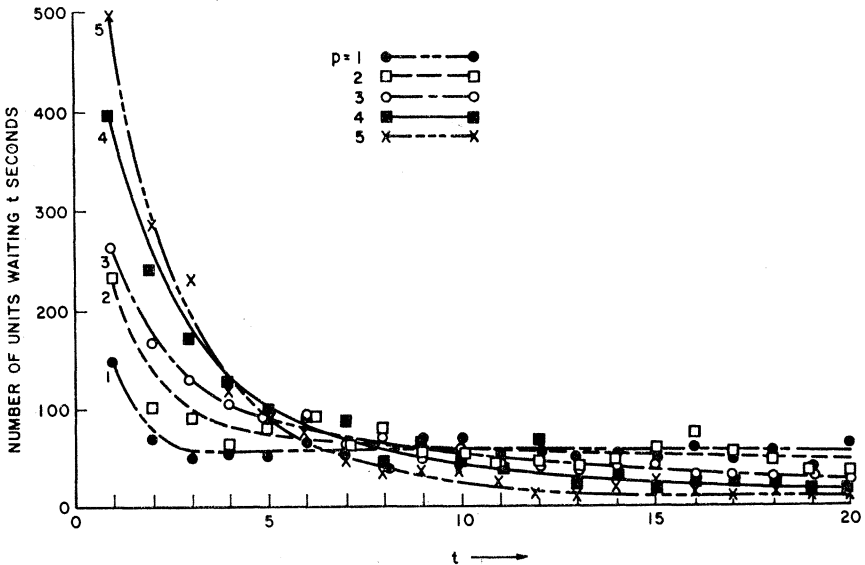Figure 6 shows the simulation results for the variance of the waiting

Fig. 5. Simulation results for the distribution of waiting times.
$$(b_p/b_{p+1})^{1/r} = 0.5, \qquad \rho = 0.9.$$

time in an $r = 0.25$ and an $r = 4.0$ order system. We note that the difference
in variance among the possible groups tends to disappear as r increases.
This is to be expected since the queue discipline is approaching the first-
come–first-served in which *no* distinction among priority groups is recog-
nized.

CONCLUSION

IN THIS paper, we have shown rth order delay-dependent priority schemes
can be reduced to the simpler delay-dependent system of order one by suit-
ably choosing the parameter set $\{b_p\}$. It was also demonstrated that this
scheme can give an entire spectrum of behavior ranging from the fixed
priority to the first-come–first-served schemes. The attempt to extend this
to cover negative values of b_p did not yield any definite results, the difficulty

here being that it is possible for two competing customers to alternate in their relative priority position *twice* instead of only once.

For a $P=2$ system, the optimum selection of $\{b_p\}$ was found for an interesting class of cost functions, which depend only on some power of the average waiting time.

Recently, Kleinrock[7] has shown that it is possible to choose the parameter set $\{b_p\}$ in a first order delay-dependent system in such a way that it displays characteristics of both delay-dependent and fixed-priority systems.
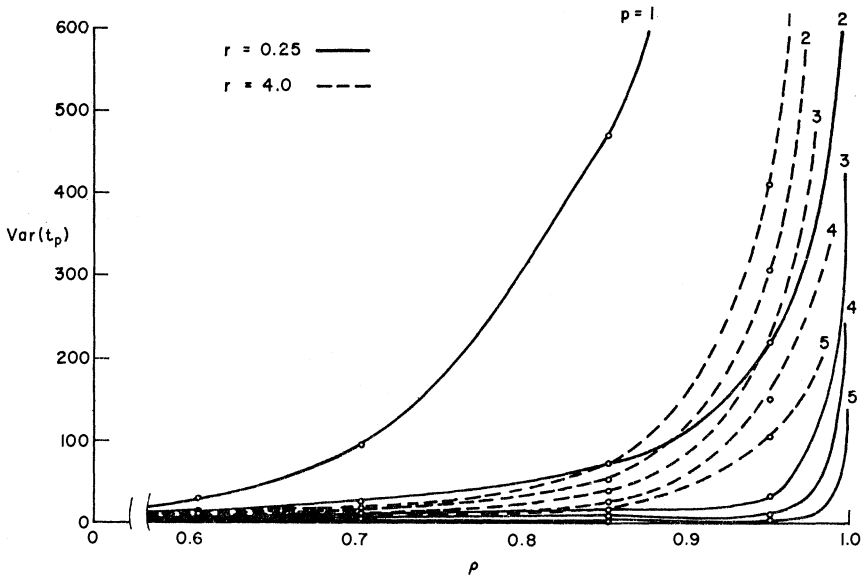

Fig. 6. Variance of t_p vs ρ $(b_p/b_{p+1}=0.5, r=0.25, 4.0)$.

APPENDIX

PROOF OF THEOREM 1. WE REPEAT the two equations of

THEOREM 1. *If we choose*

$$(b_p/b_{p+1})^{1/r} = (b_p{}'/b'_{p+1})^{1/r'} \qquad (p=1, 2, \cdots, P-1) \quad (8)$$

then
$$W_p^{(r)} = W_p^{(r')}. \tag{9}$$

Proof. Two priority schemes are said to be equivalent if the units are taken into service in exactly the same order and at the same times in both systems. The absolute values of the units' priority are unimportant, only the relative positions of the units in the queue must be the same. The positions of units in the queue are the same if and only if the same units interact with each other at the same point in time in both systems.

Without any loss of generality, r' is set equal to unity. In order to avoid any

confusion, the parameters b_p' are relabelled a_p so that equation (8) becomes

$$(b_p/b_{p+1})^{1/r} = a_p/a_{p+1}. \qquad (p=1, \cdots, P-1) \quad (24)$$

In Fig. 7 we see the priorities of three units from priority classes 1, 2, and 3 interacting with each other.* It is easy to see that each unit can exchange places in the queue with any other unit only once at the most. Thus, in general, if there are M units in the system, there can be at most $M(M-1)/2$ distinct intersection times. Thus the total number of times, including arrival times, that must be considered is $M(M+1)/2$ for the case in which all units interact with each other.

We proceed by establishing the proof of the theorem for $M=3$ and then show how to extend this proof in a trivial way for all M.
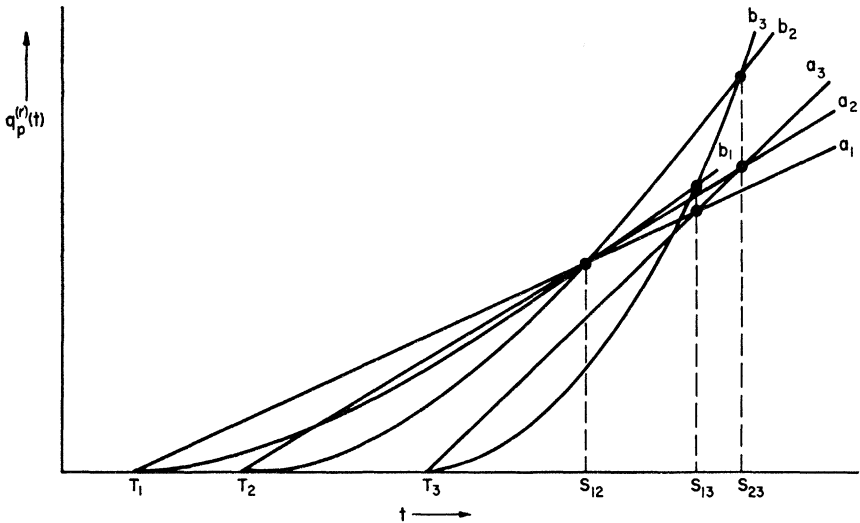

Fig. 7. Equivalence of two priority schemes.

For the case of $M=3$ (*see* Fig. 7), there are at most three intersection times, s_{12}, s_{13}, and s_{23} and three arrival times T_1, T_2, and T_3. The intersection times can be computed from the equations of the priorities as follows for $i=1, 2; j=2, 3;$ $i \neq j$:

$$b_i(s_{ij} - T_i)^r = b_j(s_{ij} - T_j)^r. \qquad (25)$$

Solving these three equations we obtain

$$s_{ij} = [T_j - T_i(b_i/b_j)^{1/r}]/[1 - (b_i/b_j)^{1/r}]. \qquad (26)$$

It will now be demonstrated that the identical interactions result at the same times in a first order system where the parameters are given by equation (24).

_____

* We assume that all these units are from distinct priority groups with no loss of generality, since units from the same group can never have intersecting priorities.

Let us construct a straight line representing the priority of unit 1 in this first order system

$$q_1^{(1)}(t) = a_1(t - T_1),$$

such that it has the same priority at s_{12} as it would have in the rth order system. Therefore

$$b_1(s_{12} - T_1)^r = a_1(s_{12} - T_1).$$

Solving these last two for a_1 and using equation (26) yields,

$$a_1 = b_1 \{ (T_2 - T_1)/[1 - (b_1/b_2)^{1/r}] \}^{r-1}. \tag{27}$$

The line representing unit 2's priority $q_2^{(1)}(t)$ as given by

$$q_2^{(1)}(t) = a_2(t - T_2)$$

is now completely determined since it must intersect unit 1's priority curve at s_{12}. Thus we have,

$$a_2(s_{12} - T_2) = a_1(s_{12} - T_1).$$

Again making use of equation (26) we have

$$a_2 = b_2[(b_1/b_2)^{1/r}(T_2 - T_1)/(1 - (b_1/b_2)^{1/r})]^{r-1}. \tag{28}$$

The straight line representing unit 3's priority must now pass through three points. It must start at T_3, intersect unit 1's priority at s_{13} and intersect unit 2's priority at s_{23}. Letting unit 3's priority be represented by

$$q_3^{(1)}(t) = a_3(t - T_3),$$

and using the first two constraints, we have

$$a_3(s_{13} - T_3) = a_1(s_{13} - T_1).$$

Making use of equation (26) and solving for a_3, we have

$$a_3 = [b_1/(b_1/b_3)^{1/r}] \cdot [(T_2 - T_1)/1 - (b_1/b_2)^{1/r}]^{r-1}. \tag{29}$$

It can be verified that equation (29) also satisfies the third constraint on unit 3, which can be expressed as

$$a_3(s_{23} - T_3) = a_2(s_{23} - T_2).$$

It is clear that we could have scaled all the a_p by the same factor and still obtained the same results. The way in which we chose the value of a_1 was entirely arbitrary. Thus it is not the values of a_p that are important but only their ratios. It is easily verified that the values of a_p in equations (27), (28), and (29) satisfy equation (24).

We have thus established the theorem for $M \leq 3$. For $M = 4$ we consider unit 4 in conjunction with units 1 and 2 and then with units 1 and 3, in each case carrying out solutions as above for $M = 3$. The straight line curve representing the priority of this fourth unit must be the same in both cases since they have two points in common, namely the intersection with the time axis (i.e., the time of arrival of the fourth unit) and the intersection with the priority curve of unit 1.

This argument can be extended to an arbitrary number of units where they are considered in groups of three.

The above argument considered the worst possible case; that for which the largest number of restrictions were placed on the curves by the points of intersection. If all units do not intersect each other, the results are still valid. This concludes the proof.

REFERENCES

1. L. KLEINROCK, "A Delay Dependent Queue Discipline," *Naval Res. Log. Quart.* **11,** 59–73 (1964).
2. A. COBHAM, "Priority Assignments in Waiting Line Problems," *Opns. Res.* **2,** 70–76 (1954).
3. J. D. C. LITTLE, "A Proof for the Queuing Formula $L = \lambda W$," *Opns. Res.* **9,** 383–387 (1961).
4. H. WHITE AND L. S. CHRISTIE, "Queuing with Preemptive Priorities or with Breakdown," *Opns. Res.* **6,** 79–95 (1958).
5. L. KLEINROCK, "A Conservation Law for a Wide Class of Queuing Disciplines," *Naval Res. Log. Quart.* **12,** 181–192 (1965).
6. D. W. FIFE, "Scheduling with Random Arrivals and Linear Loss Functions," *Management Sci.* **11,** 429–437 (1965).
7. L. KLEINROCK, "Queuing with Strict and Lag Priority Mixtures," *Proc. 1966 IFORS Conf.*