A DELAY DEPENDENT QUEUE DISCIPLINE*

L. Kleinrock

Department of Engineering
University of California
Los Angeles, California

ABSTRACT

Queue disciplines studied in the past have not given the system designer sufficient freedom with which to alter the relative waiting times of the various priority groups. In this paper, results are derived for a delay dependent priority system in which a unit's priority is increased, from zero, linearly with time in proportion to a rate assigned to the unit's priority group. The utility of this new priority structure is that it provides a number of degrees of freedom with which to manipulate the relative waiting times for each priority group.

INTRODUCTION

A number of queue disciplines have been studied in the past (see Saaty [5] for a summary of such studies). Whereas these investigations provide a careful and useful analysis, many of the queue disciplines themselves suffer from the lack of a set of adjustable parameters. Specifically, once the arrival and service rates for all priority groups are specified,[†] then the set of average waiting times are determined exactly, and the system designer has no degrees of freedom left with which to adjust the system's behavior. The delay dependent priority system described in this paper provides a set of variable parameters, b_p, which are at the disposal of the designer and which allow him to adjust the relative waiting times of each priority group to a large degree.

THE MODEL

We consider a total of P different priority groups. Units from group p (p = 1, 2, . . ., P) arrive in a Poisson stream at rate λ_p units/sec; each unit from priority class p has a required service time selected from an exponential distribution with mean $1/\mu_p$. We define[‡]

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

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

*This work was done while the author was employed at Lincoln Laboratory (operated with support from the U.S. Army, Navy, and Air Force), Massachusetts Institute of Technology, Cambridge, Massachusetts.

[†] These quantities are usually specified by the user and not by the designer of the system.

[‡] Note that W_0 is the expected time to complete service on the unit found in the service facility (see Cobham [1]).

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

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

and

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

We further define

W_p = Expected value of the time spent in the queue for a unit from group p (steady state waiting time).

The delay dependent queue discipline is such that when a unit from the p^{th} priority group enters the queue at time T (say), it is assigned a number b_p, where

$$0 \le b_1 \le b_2 \le . . . \le b_P . \tag{6}$$

The priority $q_p(t)$ at time t associated with that unit is calculated from

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

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 $q_p(t)$ 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 higher priority units gain priority at a faster rate (b_p) than lower priority units.

Figure 1 shows an example of the manner in which this priority structure allows interaction between the priority functions for two units. Specifically, at time T, a unit from priority group p_1 arrives, and attains priority at a rate equal to $(t - T)\, b_{p_1}$. At time T', a different

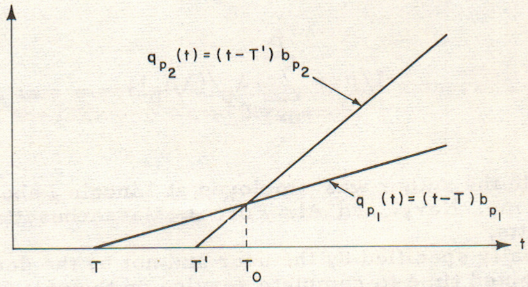

Figure 1 - Interaction between priority functions for the delay dependent priority system

unit enters from a higher priority group p_2; that is, $p_2 > p_1$. When the service facility becomes free, it next chooses that unit in the queue with the highest instantaneous priority. Thus, in our example, the first unit will be chosen in preference to the second unit if the service facility becomes free at any time between T and T_0 (in spite of the fact that the first unit is from a lower priority class); but, for any time after T_0, the second unit will be chosen in preference to the first.

MAIN RESULTS

For the delay dependent priority system, without pre-emption, we give two derived forms for W_p; one is a recursive form in terms of the W_i for the lower priority units, and the other more complicated expression is the solution of the recursive equations.

THEOREM 1:*

For the delay dependent priority system with no pre-emption, and $0 \le \rho < 1$,

(8) $$W_p = \frac{[W_0/(1-\rho)] - \sum_{i=1}^{p-1} \rho_i W_i[1-(b_i/b_p)]}{1 - \sum_{i=p+1}^{P} \rho_i [1-(b_p/b_i)]}$$

or

(9) $$W_p = [W_0/(1-\rho)](1/D_p) \left[1 + \sum_{j=1}^{p-1} \sum_{\substack{0<i_1<i_2\\ <\ldots<i_j<p}} F_{i_1}(i_2) F_{i_2}(i_3) \ldots F_{i_j}(p) \right],$$

where

(10) $$D_p = 1 - \sum_{i=p+1}^{P} \rho_i [1-(b_p/b_i)]$$

and

(11) $$F_k(n) = -(\rho_k/D_k)[1-(b_k/b_n)].$$

It is interesting to note the extremely simple dependence that W_p has on the parameters b_i (namely, only on their ratios).

For the case of the delay dependent priority system with pre-emption,* we give a recursive form for W_p in terms of the W_i for the lower priority messages.

THEOREM 2:†

For the delay dependent priority system with pre-emption, and for $0 \le \rho < 1$,

*See Saaty [5] for a definition of pre-emption.
†See the Appendix for proof of this theorem.

$$(12)\quad W_p = \frac{[W_o/(1-\rho)] + \sum_{i=p+1}^{P} (\rho_i/\mu_p)[1-(b_p/b_i)] - \sum_{i=1}^{p-1} (\rho_i/\mu_i)[1-(b_i/b_p)] - \sum_{i=1}^{p-1} \rho_i W_i[1-(b_i/b_p)]}{1 - \sum_{i=p+1}^{P} \rho_i[1-(b_p/b_i)]}$$

It is interesting to note the behavior of W_p as a function of ρ for these two disciplines and to compare this to the head of the line priority system (see Cobham [1]). The curves in Figures 2-5 have been prepared to illustrate this behavior. The assumptions are that $\lambda_p = \lambda/P$, $\mu_p = \mu$, and $b_p = 2^{p-1}$ ($p = 1, 2, \ldots, P$). These special cases do not reveal the entire structure of the W_p, but they do give one an intuitive feeling about their general properties; the obvious reason for choosing these special values is that they are easy to plot. Figures 2 and 3 show μW_p for the head of the line priority system, and Figures 4 and 5 show μW_p for the delay dependent priority system. The curves shown are for $P = 2$ and $P = 5$. In addition, the case $P = 1$ is shown as a dashed curve in all the figures; clearly, for $P = 1$, $\mu W_p(\rho) = \rho/(1-\rho)$ for all* of the disciplines, and so corresponds to the strict first come first served discipline. As such, the $P = 1$ case serves as a basis of comparison for all the curves.

Observe that, in general, the curves for the pre-emptive case are more widely spaced than the corresponding curves for the nonpre-emptive case. Further, one notes that, in

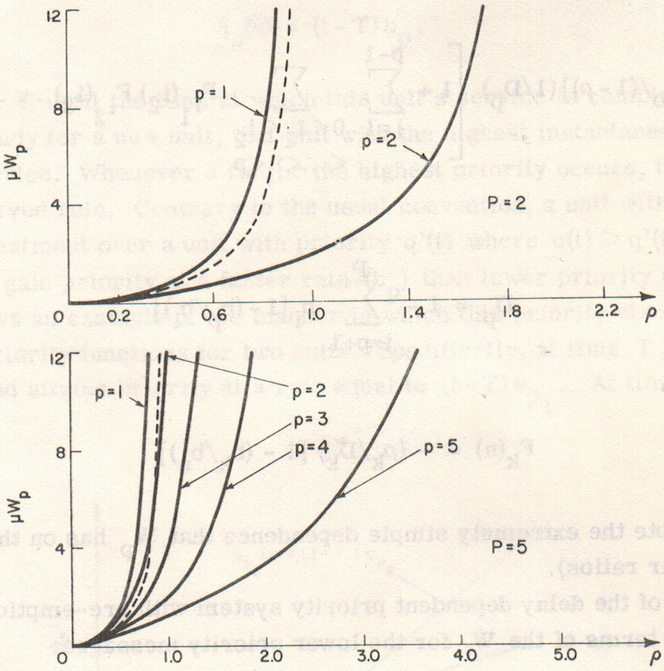

Figure 2 - $\mu W_p(\rho)$ for the head of the line priority system with no pre-emption. a) $P = 2$, b) $P = 5$.

*The Conservation Law (see Kleinrock [2]) shows why $\mu W_p(\rho)$ for the case $P = 1$ must be independent of queue discipline.

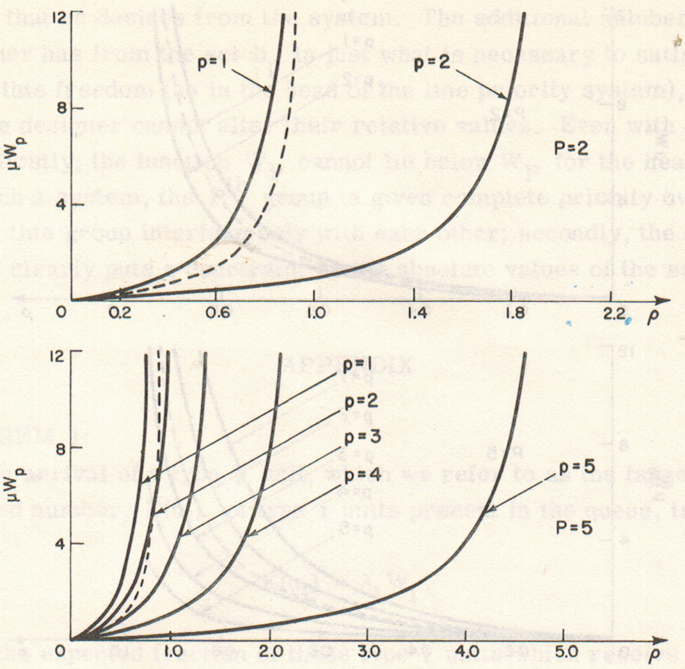

Figure 3 - $\mu W_p(\rho)$ for the head of the line priority system with pre-emption. a) P = 2, b) P = 5.

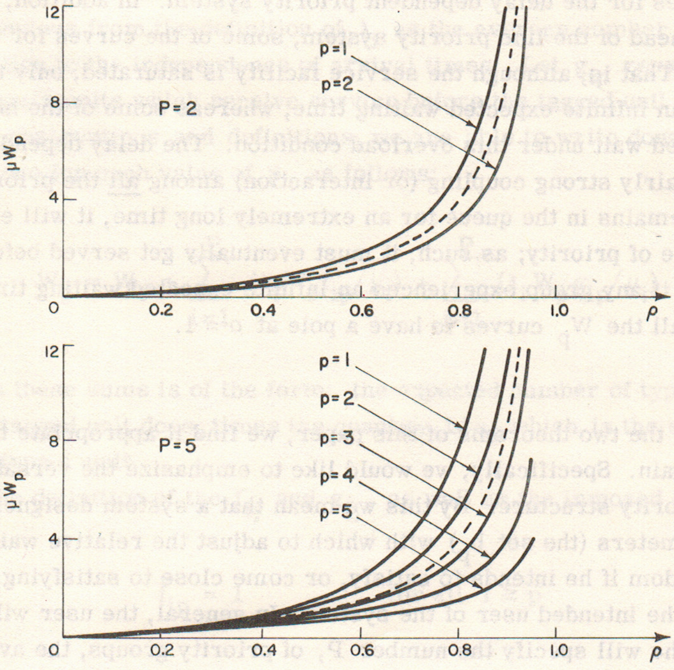

Figure 4 - $\mu W_p(\rho)$ for the delay dependent priority system with no pre-emption. a) P = 2, b) P = 5.

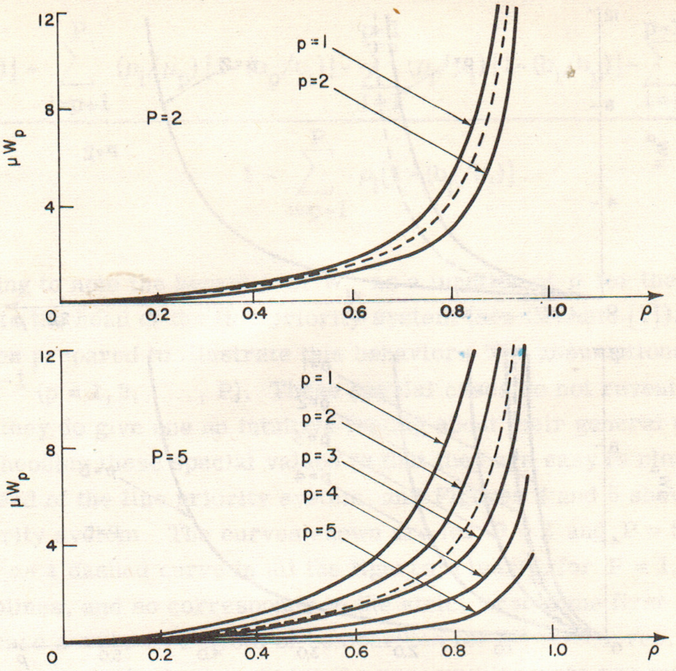

Figure 5 - $\mu W_p(\rho)$ for the delay dependent priority system with pre-emption. a) P = 2, b) P = 5.

general, the curves for the head of the line priority system are more widely spaced than the corresponding curves for the delay dependent priority system. In addition, because of the rigid nature of the head of the line priority system, some of the curves for W_p extend beyond the value of $\rho = 1$. That is, although the service facility is saturated, only the lower priority groups experience an infinite expected waiting time, whereas some of the higher priority groups have a finite expected wait under this overload condition. The delay dependent priority system, however, forces a fairly strong coupling (or interaction) among all the priority groups. Specifically, if any unit remains in the queue for an extremely long time, it will eventually attain an extremely high value of priority; as such, it must eventually get served before any newly entering units. Thus, if any group experiences an infinite expected waiting time, then they all do. This effect causes all the W_p curves to have a pole at $\rho = 1$.

CONCLUSION

In reviewing the two theorems of this paper, we find it appropriate to state the important conclusions once again. Specifically, we would like to emphasize the versatility inherent in a delay dependent priority structure. By this we mean that a system designer has at his disposal, a whole set of parameters (the set b_p) with which to adjust the relative waiting times, W_p. He must have this freedom if he intends to satisfy, or come close to satisfying, a set of specifications given him by the intended user of the system. In general, the user will specify the traffic to be handled; i.e., he will specify the number P, of priority groups, the average arrival rate, λ_p, and average service time $1/\mu_p$ for each of these groups.* Then the user will specify a

*Note that after λ_p and $1/\mu_p$ are specified, then $\rho = \sum_{p=1}^{P} \lambda_p/\mu_p$ is also specified.

set of relative W_p that he desires from the system. The additional number of degrees of freedom that the designer has from the set b_p is just what is necessary to satisfy the user's demands. Without this freedom (as in the head of the line priority system), the set W_p is fully determined, and the designer cannot alter their relative values. Even with the b_p, certain limitations exist: firstly, the function W_p cannot lie below W_p for the head of the line priority system since in such a system, the P^{th} group is given complete priority over all other groups, and members from this group interfere only with each other; secondly, the Conservation Law (see Kleinrock [2]) clearly puts a constraint on the absolute values of the set W_p.

APPENDIX

PROOF OF THEOREM 1:

Consider the arrival of a type p unit, which we refer to as the tagged unit. Upon its arrival, the expected number, $E(n_i)$, of type i units present in the queue, is (see Little [3]),

$$E(n_i) = \lambda_i W_i .$$

Let f_{ip} represent the expected fraction of these type i units which receive service before the tagged unit does. As usual, W_p will represent the expected value of the time that the tagged unit spends in the queue. We know, by assumption, that the expected number, $E(m_i)$, or type i units which arrive during the time interval W_p, is

$$E(m_i) = \lambda_i W_p .$$

That this is so is obvious from the definition of λ_i as the average number of type i arrivals per second, in addition to the independence of arrival times. Let g_{ip} represent the expected fraction of these type i units which receive service before the tagged unit does.

With these observations and definitions, we are able to write down a set of P simultaneous equations, one for each value of p, as follows:

$$W_p = W_o + \sum_{i=1}^{P} (\lambda_i W_i f_{ip}/\mu_i) + \sum_{i=1}^{P} (\lambda_i W_p g_{ip}/\mu_i) . \tag{A1}$$

The typical term in these sums is of the form: the expected number of type i units which get service before the tagged unit does, times the quantity $1/\mu_i$ which is the expected value of the service time for a type i unit.

Now from the definition of the f_{ij} and g_{ij}, as well as the imposed queue discipline, we note that

$$f_{ip} = 1 \qquad \text{for all } i \geq p$$

and

$$g_{ip} = 0 \qquad \text{for all } i \leq p .$$

Using this information and solving for W_p in Eq. (A1), we obtain

$$W_p = \frac{W_0 + \sum_{i=p}^{P} \rho_i W_i + \sum_{i=1}^{p-1} \rho_i W_i f_{ip}}{1 - \sum_{i=p+1}^{P} \rho_i g_{ip}} . \tag{A2}$$

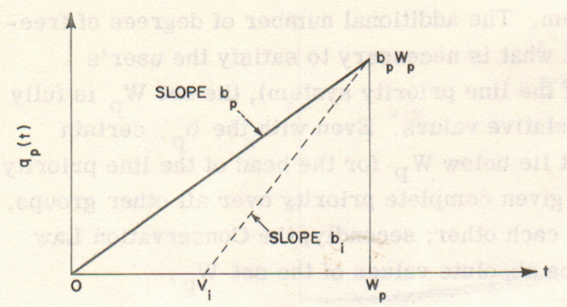

Figure A1 - Diagram of priority, $q_p(t)$, for obtaining g_{ip}

. Let us now derive an expression for g_{ip}. Once again consider the arrival of a p type unit, the tagged unit, at time 0. Since W_p is its expected waiting time, the expected value of its attained priority at the expected time it is accepted for service is $b_p W_p$, as shown in Figure A1. In looking for g_{ip}, we must calculate how many i type units arrive on the average, after time 0 and reach a priority of at least $b_p W_p$ before time W_p. It is obvious from the figure that type i units which arrive in the time interval $(0, V_i)$ will satisfy these conditions. Thus, let us calculate the value of V_i. Clearly,

$$b_p W_p = b_i (W_p - V_i)$$

and so

$$V_i = W_p [1 - (b_p/b_i)] .$$

Therefore, with an input rate of λ_i for the type units, we find that

$$g_{ip} E(m_i) = \lambda_i V_i$$

and so

$$g_{ip} \lambda_i W_p = \lambda_i W_p [1 - (b_p/b_i)]$$

giving

$$g_{ip} = 1 - (b_p/b_i) \qquad \text{for all } i > p .$$

We now prove that $f_{ip} = b_i/b_p$ for $i \leq p$. Consider that a type p unit, the tagged unit, arrives at time $t = 0$, and spends a total time t_p in the queue. Its attained priority at the time of its acceptance into the service facility will be $b_p t_p$, as shown in Figure A2.

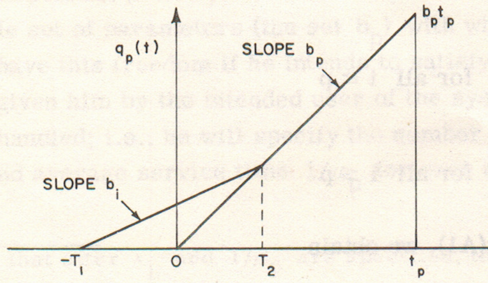

Figure A2 - Diagram of priority, $q_p(t)$, for obtaining f_{ip}

Upon its arrival, the tagged unit finds n_i type i units already in the queue. Let us consider one such type i unit, as shown in the figure, which arrived at $t = -T_1$. In looking for f_{ip}, we must calculate how many type i units arrive before $t = 0$, and obtain service before the tagged unit does. It is obvious from the figure that a type i unit which arrives at time $-T_1(T_1 > 0)$ and which waits in the queue a time $w_i(T_1)$ such that $T_1 \le w_i(T_1) \le T_1 + T_2$ will satisfy these conditions. Obviously, the reason that $w_i(T_1)$ must not exceed $T_1 + T_2$ is that for $w_i(T_1) > T_1 + T_2$ that i type unit will be of lower priority than the tagged unit, and will therefore fail to meet the conditions stipulated above. Note that T_2 may exceed t_p, but this does not violate our conditions since in that case the i type unit must surely be serviced before the tagged unit is serviced.

Therefore, let us first solve for T_2. Clearly,

$$b_p T_2 = b_i(T_1 + T_2)$$

and so

$$T_2 = [b_i/(b_p - b_i)]\, T_1$$

or

$$T_1 + T_2 = [b_p/(b_p - b_i)]\, T_1 .$$

It is clear that the expected number, $E(n_i)\, f_{ip}$, of i type units which are in the queue at $t = 0$ and which also obtain service before the tagged unit does, can be expressed as

$$E(n_i) f_{ip} = \int_0^\infty \lambda_i P_r \left\{ t \le w_i(t) \le [b_p/(b_p - b_i)]t \right\} dt , \tag{A3}$$

where $\lambda_i\, dt$ is the expected number of i type units that arrived during the time interval $(-t - dt, -t)$ and where $P_r \left\{ t \le w_i(t) \le [b_p/(b_p - b_i)]t \right\}$ is the probability that a unit which arrived in that interval spends at least t and at most $[b_p/(b_p - b_i)]t$ seconds in the queue. Equation (A3) can be written as

$$E(n_i) f_{ip} = \lambda_i \int_0^\infty [1 - P_r(w_i \le t)]\, dt - \lambda_i \int_0^\infty \left[1 - P_r \left\{ w_i \le [b_p/(b_p - b_i)]t \right\} \right] dt$$

$$= \lambda_i \int_0^\infty [1 - P_r(w_i \le t)]\, dt - \lambda_i [1 - (b_i/b_p)] \int_0^\infty [1 - P_r(w_i \le \sigma)]\, d\sigma ,$$

where we have set

$$\sigma = [b_p/(b_p - b_i)]\, t .$$

Now, as is well-known* (for w_i a nonnegative random variable),

$$E(w_i) = \int_0^\infty [1 - P_r(w_i \le x)]\, dx$$

and since, in our notation $W_i = E(w_i)$, we obtain

$$E(n_i) f_{ip} = \lambda_i W_i - \lambda_i [1 - (b_i/b_p)] W_i$$

or

$$f_{ip} = [\lambda_i W_i / E(n_i)](b_i/b_p) .$$

But we know that

$$E(n_i) = \lambda_i W_i ,$$

and therefore

$$f_{ip} = b_i/b_p \quad \text{for all } i \le p .$$

Having derived expressions for f_{ip} and g_{ip}, we may now substitute for these quantities in Eq. (A2), and obtain,

$$W_p = \frac{W_o + \sum_{i=p}^{P} \rho_i W_i + \sum_{i=1}^{p-1} \rho_i W_i (b_i/b_p)}{1 - \sum_{i=p+1}^{P} \rho_i [1 - (b_p/b_i)]} .$$

If we now make use of the Conservation Law (see Kleinrock [2]) we can rewrite the above equation as

$$W_p = \frac{[W_o/(1 - \rho)] - \sum_{i=1}^{p-1} \rho_i W_i [1 - (b_i/b_p)]}{1 - \sum_{i=p+1}^{P} \rho_i [1 - (b_p/b_i)]} ,$$

which establishes Eq. (8) of Theorem 1.

Let us now show that Eq. (9) is indeed the solution to the set of recursively defined W_p, as expressed in Eq. (8). We proceed to show this by an inductive proof.

*See, for example, Morse [4], p. 9.

First, for $p = 1$, we get, from Eq. (9)

$$W_1 = [W_o/(1-\rho)]\,(1/D_1) = [W_o/(1-\rho)] \left[\frac{1}{1 - \sum\limits_{i=2}^{P} \rho_i\,[1-(b_1/b_i)]}\right],$$

which checks with the value of W_1 obtained from Eq. (8).

For $p = 2$, we get from Eq. (9)

$$W_2 = [W_o/(1-\rho)]\,(1/D_2)\,[1+F_1(2)] = [W_o/(1-\rho)]\; \frac{1 - \left\{\rho_1[1-(b_1/b_2)] \middle/ \left[1 - \sum\limits_{i=2}^{P} \rho_i\,[1-(b_1/b_i)]\right]\right\}}{1 - \sum\limits_{i=3}^{P} \rho_i\,[1-(b_2/b_i)]}$$

which checks with the value of W_2 obtained from Eq. (8).

Now, as is usual in an inductive proof, we assume that the solution holds for all $p \le k$, and we show that this implies that the solution is correct for $p = k+1$. Let us therefore write down the expression W_{k+1} from Eq. (8), using the fact that $W_k, W_{k-1}, \ldots, W_1$ may be evaluated from Eq. (9):

$$W_{k+1} = [W_o/(1-\rho)]\,(1/D_{k+1}) \left\{1 - \sum_{i=1}^{k} \rho_i\,[1-(b_i/b_{k+1})]\,(1/D_i) \left[1 + \sum_{j=1}^{i-1} \sum_{\substack{0<i_1<\ldots\\ <i_j<i}} F_{i_1}(i_2)\ldots F_{i_j}(i)\right]\right\}$$

$$= [W_o/(1-\rho)]\,(1/D_{k+1}) \left\{1 + \sum_{i=1}^{k} F_i(k+1) \left[1 + \sum_{j=1}^{i-1} \sum_{\substack{0<i_1<\ldots\\ <i_j<i}} F_{i_1}(i_2)\ldots F_{i_j}(i)\right]\right\},$$

where we have taken the liberty of using the notation of Eqs. (10) and (11). Now, comparing this last equation with the expression obtained for W_{k+1} from Eq. (9), we see that the induction proves the result if the following identity exists:

$$\sum_{i=1}^{k} F_i(k+1) \left[1 + \sum_{j=1}^{i-1} \sum_{\substack{0<i_1<\ldots\\ <i_j<i}} F_{i_1}(i_2)\ldots F_{i_j}(i)\right] = \sum_{j=1}^{k} \sum_{\substack{0<i_1<\ldots\\ <i_j<k+1}} F_{i_1}(i_2)\ldots F_{i_j}(k+1).$$

It is clear that both sides of this equation involve n-tuples of the F factors. Therefore, in order to prove the validity of this expression, let us show that the same sets of n-tuples appear on both sides of the equation. First, for $n = 1$, we require that

$$\sum_{i=1}^{k} F_i(k+1) = \sum_{i_1=1}^{k} F_{i_1}(k+1),$$

which is obviously correct. Now for $n > 1$, we require that the n-tuples agree, and so, writing only the n-tuples for each side of the equation, we have

$$\sum_{i=1}^{k} F_i(k+1) \sum_{\substack{0<i_1<\dots\\ <i_{n-1}<i}} F_{i_1}(i_2)\dots F_{i_{n-1}}(i) = \sum_{\substack{0<i_1<\dots\\ <i_n<k+1}} F_{i_1}(i_2)\dots F_{i_n}(k+1).$$

If, on the right hand side of this last equation, we separate out the summation involving i_n, as follows,

$$\sum_{\substack{0<i_1<\dots\\ <i_n<k+1}} F_{i_1}(i_2)\dots F_{i_n}(k+1) = \sum_{i_n=1}^{k} F_{i_n}(k+1) \sum_{\substack{0<i_1<\dots\\ <i_{n-1}<i_n}} F_{i_1}(i_2)\dots F_{i_{n-1}}(i_n),$$

we find that the n-tuples do indeed agree (i.e., let $i_n = i$ in this last expression). Thus, we have proven the validity of Eq. (9) and this completes the proof of THEOREM 1.

PROOF OF THEOREM 2:

Here, we use notation very similar to that used in the proof of Theorem 1 except that all quantities will refer to time spent in the queue plus service facility, instead of just in the queue as was the case in Theorem 1.

Following through with almost identical arguments, we arrive at the following expressions (where $T_i = W_i + (1/\mu_i)$):

$$E(n_i) = \lambda_i T_i,$$

$$E(m_i) = \lambda_i T_p,$$

$$f_{ip} = \begin{cases} b_i/b_p & i \le p, \\ 1 & i \ge p, \end{cases}$$

and

$$g_{ip} = \begin{cases} 0 & i \le p, \\ [1-(b_p/b_i)] & i \ge p, \end{cases}$$

where n_i is now defined as the total number of type i units which were present in the system (queue plus service facility) when the tagged unit arrived, and m_i is defined as the total number of type i units which enter the system while the tagged unit is in the system.

The expression for T_p is therefore

$$T_p = (1/\mu_p) + \sum_{i=1}^{P} (\lambda_i T_i f_{ip}/\mu_i) + \sum_{i=1}^{P} (\lambda_i T_p g_{ip}/\mu_i) .$$

This equation is obtained from reasoning quite similar to that used in forming Eq. (A1). Now, using the expressions for f_{ip} and g_{ip}, and also remembering that $W_i + 1/\mu_i = T_i$ we obtain,

$$W_p = T_p - (1/\mu_p) = \sum_{i=1}^{p} \rho_i [W_i + (1/\mu_i)](b_i/b_p) + [W_o/(1-\rho)] - \sum_{i=1}^{p} \rho_i [W_i + (1/\mu_i)]$$

$$+ \sum_{i=p+1}^{P} \rho_i [W_p + (1/\mu_p)] [1 - (b_p/b_i)] ,$$

where we have also made an application of the Conservation Law in this last expression. Solving for W_p, and collecting terms, we obtain finally,

$$W_p = \frac{[W_o/(1-\rho)] + \sum_{i=p+1}^{P} (\rho_i/\mu_p)[1-(b_p/b_i)] - \sum_{i=1}^{p-1} (\rho_i/\mu_i)[1-(b_i/b_p)] - \sum_{i=1}^{p-1} \rho_i W_i [1-(b_i/b_p)]}{1 - \sum_{i=p+1}^{P} \rho_i [1-(b_p/b_i)]} ,$$

which is the same as Eq. (12) and so proves THEOREM 2. Note that, E(n), the expected number of units in the system, is

$$E(n) = \sum_{p=1}^{P} E(n_p) = \sum_{p=1}^{P} \lambda_p T_p .$$

REFERENCES

[1] A. Cobham, "Priority Assignments in Waiting Line Problems," Operations Research, 2, 70-76 (1954).

[2] L. Kleinrock, "A Conservation Law for a Wide Class of Queue Disciplines" (to be published).

[3] J. D. C. Little, "A Proof for the Queueing Formula L = λW," Operations Research, 9, 383-387 (1961).

[4] P. M. Morse, Queues, Inventories, and Maintenance (John Wiley and Sons, Inc., New York, N.Y., 1958).

[5] T. L. Saaty, Elements of Queueing Theory with Applications (McGraw-Hill Book Co., New York, N.Y., 1961).

* * *