Journal of the Operations Research Society of America

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Priority Assignment in Waiting Line Problems

Alan Cobham,

PRIORITY ASSIGNMENT IN WAITING LINE PROBLEMS

ALAN COBHAM

Operations Evaluation Group, United States Navy

(Received September 22, 1953)

There are several commonly occurring situations in which the position of a unit or member of a waiting line is determined by a priority assigned to the unit rather than by its time of arrival in the line. An example is the line formed by messages awaiting transmission over a crowded communication channel in which urgent messages may take precedence over routine ones. With the passage of time a given unit may move forward in the line owing to the servicing of units at the front of the line or may move back owing to the arrival of units holding higher priorities. Though it does not provide a complete description of this process, the average elapsed time between the arrival in the line of a unit of a given priority and its admission to the facility for servicing is useful in evaluating the procedure by which priority assignments are made. Expressions for this quantity are derived for two cases—the single-channel system in which the unit servicing times are arbitrarily distributed (Eq. 3) and the multiple-channel system in which the servicing times are exponentially distributed (Eq. 6). In both cases it is assumed that arrivals occur at random.

IN THE ANALYSIS of waiting line problems the assumption is frequently made that the units (customers) arriving in the system are serviced strictly in order of arrival.* There are, however, many situations of practical interest—in the fields of military communications, freight handling, and machine maintenance, for example—in which the order of servicing is determined by a priority system that allows a high-priority unit to displace units of lower priority in the waiting line, the effect being to reduce the waiting period for high-priority units at the cost of increasing the wait for those of lower priority. Quantitative examination of these effects may in some cases lead to a more efficient method of assigning priorities and a consequent reduction of costly delays without any increase in system facilities. In this paper one such assignment procedure is discussed, and expressions for the expected waiting time for units of each priority level are derived for two special cases.

Under the particular priority procedure examined here a unit of one priority level will displace all those of a lower level in the waiting line but may not cut in on any units that have already started servicing. Units of the same level are serviced in order of arrival. The present discussion is limited to the case where the number of priority levels r is finite, priority 1 being taken as the highest, and priority r as the lowest.

* An interesting exception is discussed in a paper by J. Riordan, "Delay Curves for Calls Served at Random," *Bell System Tech. J.* **32,** No. 1, 1953.

Consider the situation arising when a unit U of priority p, $1 \leq p \leq r$, arrives in a system in which such a procedure is in operation. If one or more of the channels (service counters—assumed to be finite in number) of the system is available, then servicing on U may be begun at once; if, on the other hand, all channels are occupied, then U must join the waiting line until one becomes available. Suppose that in this case the number of units of priority p or higher in the waiting line at the time of entry of U is N_0 and the time elapsing between the arrival of U and the discharge of the (N_0+1)st unit from the system is t_0. Then U must wait at least t_0, and if no units of priority higher than p have entered the waiting line during this time then this is all U must wait. However, if some number, say N_1, of units of priority higher than p enter the system during the interval t_0 then, since these units are allowed to get ahead of U in the line, U must wait an additional period t_1 while another N_1 units are discharged. If now t_{k+1} is defined as the time elapsing between the discharge of the $(N_k+\cdots+N_0+1)$st unit and the $(N_{k+1}+\cdots+N_0+1)$st unit and N_{k+1} as the number of units of priority higher than p entering during the interval t_k then it is apparent that U must wait exactly $\sum_0^\infty t_k$ to start servicing.

Obviously now W_p, the expected length of time U will spend in the waiting line, is the expected value of the sum of all t_k, $E\left(\sum_0^\infty t_k\right)$. In order to evaluate this it is necessary to make some assumption concerning the nature of the units being serviced and their rate of arrival. For the present discussion it is assumed that the units of each priority level arrive according to independent Poisson laws, $\lambda_1, \cdots, \lambda_r$ being the respective average arrival rates; as a consequence the arrival of units of all levels combined must also be Poisson with average rate $\lambda = \lambda_1 + \cdots + \lambda_r$. Denoting by $F_p(t)$ the cumulative unit-servicing-time distribution for units of priority p, i.e., the probability that a unit of priority p will have a servicing time not greater than t, then $F(t)$, defined by

$$F(t) = \frac{1}{\lambda} \sum_1^r \lambda_p \, F_p(t),$$

is the combined unit-servicing-time distribution. Discussion will be restricted here to two special cases—the single-channel system with an arbitrary $F_p(t)$ and the multiple-channel system with all $F_p(t)$ having a negative exponential form: $F_p(t) = 1 - e^{-\mu t}$. In both instances consideration is given only to systems in which a state of statistical equilibrium exists.

SINGLE CHANNEL SYSTEM

In the first case, that of a single-channel system with arbitrary unit-servicing-time distributions, $E(\sum t_k)$ may be found by first evaluating $E(t_{k+1}|t_k)$, which represents the expected value of t_{k+1}, given t_k. The expected number of units of priority j entering during a period t_k will be $\lambda_j t_k$, and the expected amount of time consumed in servicing them will be $(\lambda_j/\mu_j)t_k$, where $1/\mu_j$ is the average servicing time for units of the jth priority level $1/\mu_j = \int_0^\infty t\, dF_j(t)$; hence, since all units must be serviced in the same channel, $E(t_{k+1}|t_k)$ may be found immediately as

$$\sum_{j=1}^{p-1} \frac{\lambda_j}{\mu_j} t_k.$$

Now $E\left(\sum_0^{k+1} t_j\right) = E\left[\sum_0^k t_j + E(t_{k+1}|t_k)\right] = E\left(\sum_0^k t_j + \sum_{j=1}^{p-1} \frac{\lambda_j}{\mu_j} t_k\right)$,

and a simple induction can be used to show that

$$E\left(\sum_0^{k+1} t_j\right) = E\left[\left(1 + \sum_1^{p-1} \frac{\lambda_j}{\mu_j} + \cdots + \left[\sum_1^{p-1} \frac{\lambda_j}{\mu_j}\right]^{k+1}\right) t_0\right]$$
$$= \sum_{i=0}^{k+1} \left(\sum_{j=1}^{p-1} \frac{\lambda_j}{\mu_j}\right)^i E(t_0),$$

which on taking the limit becomes

$$W_p = E\left(\sum_0^\infty t_j\right) = E(t_0) \sum_{i=0}^\infty \left(\sum_{j=1}^{p-1} \frac{\lambda_j}{\mu_j}\right)^i = E(t_0) \Big/ \left(1 - \sum_1^{p-1} \frac{\lambda_j}{\mu_j}\right). \qquad (1)$$

$E(t_0)$ may be evaluated by breaking it into two components: that contributed by the delay, if any, between the entry of U into the system and the discharge of the unit occupying the channel when U entered, and that contributed by the delay between the discharge of the 1st and the (N_0+1)st unit from the system. Consider the first of these; if U entered while a unit of servicing time t was occupying the channel its expected point of entry is halfway through, i.e., with $\frac{1}{2}t$ of the servicing to be completed. Since the probability that U did in fact enter under such circumstances is $\lambda t dF(t)$ the contribution from this component must be $\int_0^\infty (\frac{1}{2}t) \times (\lambda t dF(t))$. To find the contribution of the second component it is sufficient to observe that the expected number of units of priority k waiting to be serviced at any time is $\lambda_k W_k$, where W_k is the expected wait for a unit of priority k, so that the expected amount of time that will be consumed in servicing them will be $\lambda_k W_k/\mu_k$. Combining these results leads to the equation for W_p,

$$W_p=\frac{\sum_1^p \lambda_k\, W_k/\mu_k+\tfrac{1}{2}\lambda\int_0^\infty t^2\, dF(t)}{1-\sum_1^{p-1}\lambda_k/\mu_k}, \tag{2}$$

the solution to which is

$$W_p=\frac{\tfrac{1}{2}\lambda\int_0^\infty t^2\, dF(t)}{\left(1-\sum_1^{p-1}\lambda_k/\mu_k\right)\left(1-\sum_1^{p}\lambda_k/\mu_k\right)}, \tag{3}$$

as can easily be shown by induction on p.

MULTIPLE CHANNEL SYSTEM

Derivation of the average waiting time for the multiple-channel system with the unit-servicing-time distribution limited to a negative exponential type follows similar lines. It is assumed that there are a fixed number n of channels and that units of all priority levels are drawn from the same distribution so that $F_p(t)=F(t)=1-e^{-\mu t}$. The average servicing time for a randomly selected unit of any priority is then

$$1/\mu=\int_0^\infty t dF(t)=\int_0^\infty \mu t e^{-\mu t}\, dt.$$

As before, $E(\sum t_k)$ may be found by first determining $E(t_{k+1}|t_k)$. The expected number of units of priority higher than p to enter during t_k is $\sum_{j=1}^{p-1}\lambda_j t_k$ and since, when all channels of the system are full, the discharges are at random with average rate $n\mu$[1], $E(t_{k+1}|t_k)$ is given by $1/n\mu\sum_{j=1}^{p-1}\lambda_j t_k$. Following the same line of argument as previously it is found that

$$W_p=\frac{E(t_0)}{1-\frac{1}{n\mu}\sum_1^{p-1}\lambda_j}. \tag{4}$$

Splitting $E(t_0)$ into the same components as before it is found that the part contributed by the expected wait between the entry of U and the discharge of a unit from the system, provided all channels were occupied when U joined the waiting line, is $\pi/n\mu$ where π is the probability that there were no vacant channels when U entered and $1/n\mu$ represents the expected wait,

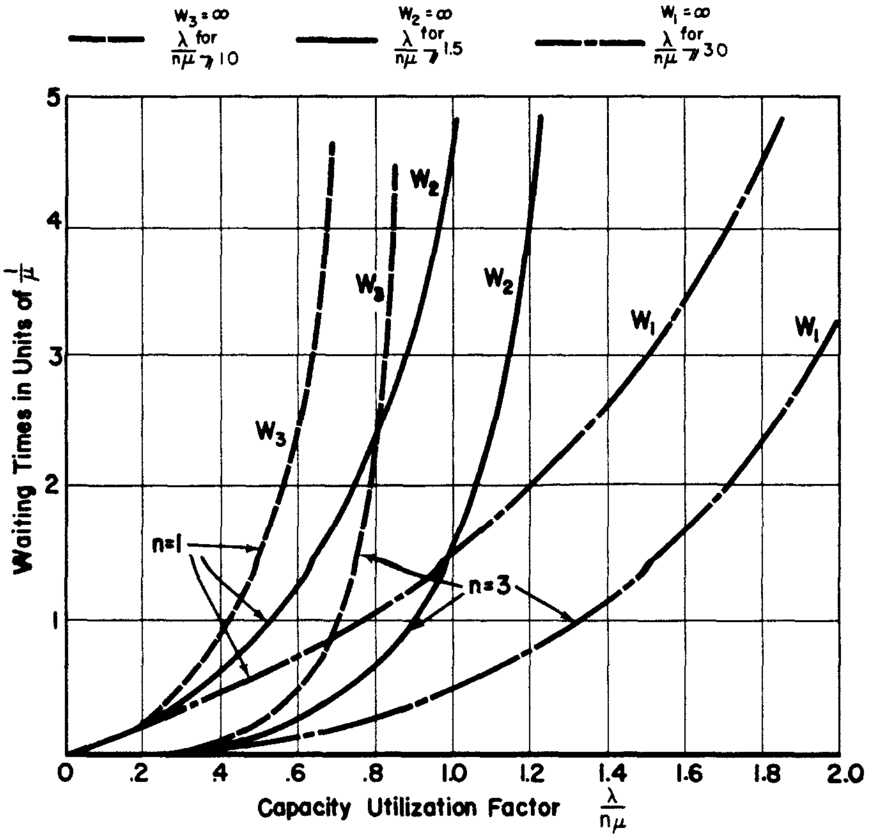

FIG. 1. Waiting time as a function of capacity utilization for an n-channel system ($n=1,3$) with three priority levels. $F(t)=1-e^{-\mu t}$, $\lambda_1=\lambda_2=\lambda_3=\frac{1}{3}\lambda$.

it being given that all channels were occupied. π has been shown[2] to have the value

$$\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!\left(1-\frac{\lambda}{n\mu}\right)\left[\sum_{j=0}^{n-1}\frac{(\lambda/\mu)^j}{j!}+\sum_{j=n}^{\infty}\frac{(\lambda/\mu)^j}{n!n^{j-n}}\right]},$$

when $\lambda/n\mu<1$, and is equal to 1 otherwise. Since the expected number of units of priority k in the waiting line at any time is $\lambda_k W_k$ and since units are discharged at an average rate of $n\mu$ regardless of priority, the second component of $E(t_0)$ is $1/n\mu\sum_{1}^{p}\lambda_k W_k$. When these results are combined, an

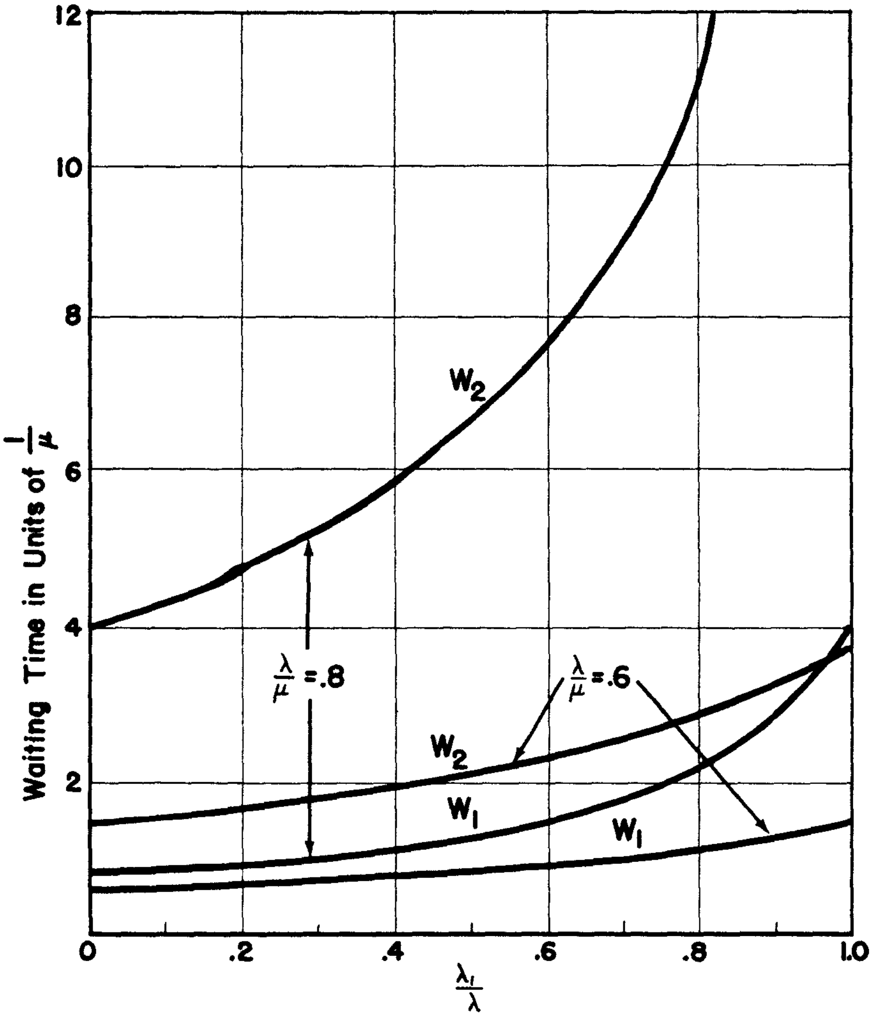

FIG. 2. Waiting time as a function of relative frequency of high-priority units λ_1/λ for a two-priority single-channel system with an exponential servicing-time distribution.

equation very similar to (2) is derived:

$$W_p=\frac{\frac{1}{n\mu}\sum_1^p \lambda_k W_k+\frac{\pi}{n\mu}}{1-\frac{1}{n\mu}\sum_1^{p-1}\lambda_k}, \tag{5}$$

which can by induction be shown to have the solution

$$W_p=\frac{\pi/n\mu}{\left(1-\frac{1}{n\mu}\sum_1^{p-1}\lambda_k\right)\left(1-\frac{1}{n\mu}\sum_1^{p}\lambda_k\right)}. \tag{6}$$

Equations (2) and (5) hold, provided only that $\sum_1^p \lambda_j/\mu_j<1$ in (2) and $1/n\mu \sum_1^p \lambda_j<1$ in (5). It follows that equations (3) and (6) also hold under these conditions even if $\lambda/n\mu$ is itself greater than unity. This may be interpreted as meaning that, even if the capacity of the system is insufficient to handle the full load placed on it, units of priority p will be serviced provided p satisfies the appropriate one of the above inequalities. An example of this effect is shown in Fig. 1.

In a system such as has been described here the difference between the expected waits for high- and low-priority units is small if the system is operating at a point well under capacity. As the load on the system increases, the difference between waiting times for units of the various priority levels becomes more apparent, and it is increasingly more important to make the best use of the priority system. Since, in general, neither the rate of influx λ of units into the system nor the servicing-time distribution $F(t)$ can be controlled, such control as can be exercised must be in the assignment of priorities. In a single-channel system, for exsample, any increase in the relative frequency of priority 1 units increases not only the expected delay for units of that priority level but for units of all other levels as well. In other words it is best to keep the frequency of high-priority units as low as possible if the priority system is to serve its supposed function of permitting the more important units to get through the system as rapidly as possible (Fig. 2). For the same reason it is best where possible to keep the servicing times for high-priority units as short as possible since occurrence of long servicing times in units of high priority tends to increase the expected wait for units of all levels. Similar remarks apply to multiple-channel systems although the analysis given here does not allow for the distributions of servicing times to differ among the various priority levels.

LITERATURE CITED

[1] W. Feller, *Probability Theory and its Applications*, Vol. 1, John Wiley & Sons, Inc., 1950.

[2] E. C. Molina, "Application of the Theory of Probabilities to Telephone Trunking Problems," *Bell System Tech. J.*, **6**, 1927