3

Priority Queueing

Nobody likes to wait in line; however, some of us dislike it more than others. In fact, some of us dislike it so much that we are willing to do something about it. In order to improve one's position in line, one may cheat, bribe, push, or quit. A more cunning action might be to join a class structure that is afforded preferential treatment at the expense of others. Such schemes are referred to as priority queueing systems, and they form the subject of the present chapter.

Immediately when one considers priority queueing systems one naturally thinks in terms of minimizing some cost function with regard to delay for various customers. Surprisingly there exist relatively few works in the literature dealing with optimal queueing disciplines for priority groups such that some well-stated and realistic cost functions are minimized. Rather, what abounds is a vast literature on the construction of mathematical models for ingenious priority systems, followed by an analysis of the performance of such systems.*

Our purpose in this chapter is to discuss a few priority systems of interest (particularly to the author) and we in no way attempt or profess to cover the material in this field in any degree of completion. Rather, we raise questions and illustrate methods of approach that we feel are meaningful and general. These considerations lay the groundwork for some of our computer applications in Chapters 4, 5 and 6. For a much more complete work on the subject, the reader is referred to Jaiswal's book on priority queues [JAIS 68].

3.1. THE MODEL

A queueing discipline is nothing more than a means for choosing which customer in the queue is to be serviced next. This decision may be based

* On the other hand, a literature on the optimal *control* of queueing systems does exist. Here one is concerned with adjusting the service and arrival rates of the system under various cost structures. This material is summarized in [PRAB 73] and [CRAB 73], both of which contain useful bibliographies. Applications to closed queueing networks are given in [TORB 73]. A recent approach is also reported in [REED 74].

on any or all of the following:

1. some measure related to the relative arrival times for those customers in the queue;
2. some measure (exact value, estimate, pdf) of the service time required or the service so far received;
3. or some function of group membership.

The third case is usually referred to as a priority queueing discipline, although in this chapter we use the broader definition to include any of those three discriminators. Examples of queueing disciplines that depend only upon arrival time are first-come–first-serve (FCFS), last-come–first-serve (LCFS), and random order of service. Discrimination on the basis of service time only may take the following forms: shortest-job-first (SJF), longest-job-first (LJF), similar rules based on averages, and so on. Order of service based on an externally imposed priority class structure may take many forms as, for example, the head-of-the-line (HOL) system described below. Mixtures of these disciplines are also common, and we discuss one such mixture in Section 3.7.

We assume that arriving customers belong to one of a set of P different priority classes, indexed by the subscript p ($p = 1, 2, \ldots, P$). We adopt the convention that the *larger* the value of the index associated with the priority group, the *higher* is the so-called priority associated with that group; that is, customers from priority group p are given preferential treatment in one form or another on the average over customers from priority group $p-1$. We consider only equilibrium results here; however, we do encounter systems below in which some groups have no stable behavior while other groups do reach a limiting stable behavior, and it is the stable groups that we consider in such cases.

In general, then, we assume that an arriving customer is assigned a set of parameters (either at random or based on some property of the customer) that determine his relative position in the queue through the decision rule known as the queueing discipline. This position may vary as a function of time owing to the appearance of customers of higher or lower priority in the queue. At time t a customer's priority is calculated as a function of his assigned parameters, his service time and his time in the system. In fact, we associate with a customer from priority group p a numerically valued priority function $q_p(t)$ at time t. The higher the value obtained by this function, the higher is said to be the customer's priority; whenever the decision rule is called upon to select a customer for service, the choice is made in favor of that customer with the largest $q_p(t)$. All ties are broken on an FCFS basis.

We consider a fairly general model based on the system M/G/1 (in

some cases, however, we constrain the system to be of the form M/M/1; at other times, we generalize to G/G/1). Thus we assume that customers from priority group p arrive in a Poisson stream at rate λ_p customers per second; each customer from this group has his service time selected independently from the distribution $B_p(x)$ with mean $\bar{x}_p$ sec. We define the following:

$$\lambda = \sum_{p=1}^{P} \lambda_p \tag{3.1}$$

$$\bar{x} = \sum_{p=1}^{P} \frac{\lambda_p}{\lambda} \bar{x}_p \tag{3.2}$$

$$\rho_p = \lambda_p \bar{x}_p \tag{3.3}$$

$$\rho = \lambda \bar{x} = \sum_{p=1}^{P} \rho_p \tag{3.4}$$

The interpretation of ρ here is, as usual, the fraction of time the server is busy (so long as $\rho < 1$). Moreover, ρ_p is the fraction of time the server is busy with customers from group p (again for $\rho < 1$). If a customer in the process of being served is liable to be ejected from service and returned to the queue whenever a customer with a higher value of priority appears in the queue, then we say that the system is a *preemptive* priority queueing system; if such is not allowed, then the system is said to be *nonpreemptive*. If only one customer is allowed in the (single) service facility at a time, then when there exists a tie between customers, the tie is broken on a first-come–first-serve basis.

3.2. AN APPROACH FOR CALCULATING AVERAGE WAITING TIMES

According to our earlier notation we have reserved the symbols W and T to denote a customer's average waiting time (in queue) and average total time in system (queue plus service), respectively; of course, the two are related through $T = W + \bar{x}$. We make the corresponding definitions for the case of priority classes, namely

$$W_p \triangleq E[\text{waiting time for customers from group } p] \tag{3.5}$$

$$T_p \triangleq E[\text{total time in system for customers from group } p] = W_p + \bar{x}_p \tag{3.6}$$

A customer's waiting time is easily decomposed into three parts: any delay he encounters due to the customer found in service upon his arrival; any delay he experiences due to customers he finds in the queue upon his arrival; and lastly, any delay due to customers who arrive after he does.

This is the basic observation from which we may establish a set of equations that define the quantities W_p or T_p.

We begin by considering the case of *nonpreemptive* systems and establish the equations for the average waiting times W_p. We study the system from the point of view of a newly arriving customer from priority group p (say); we shall refer to this customer as the "tagged" customer. We observe that the first part of the tagged customer's delay is due to the customer he finds in service; this delay will be equal to this other customer's residual life, the distribution of which will depend upon the priority group to which this other customer belongs. Let us denote by W_0 the *average delay to our tagged customer due to the man found in service.* Since ρ_i is the fraction of time that the server is occupied by customers from group i and since we have a Poisson process, then ρ_i is the probability that our tagged customer finds a type-i customer in service. In Section 1.7 we stated that with Poisson arrivals, the mean residual life of a service time as observed by an arrival is equal to the second moment of service divided by twice the first moment; these statements permit us to calculate W_0 as

$$W_0 = \sum_{i=1}^{P} \rho_i \frac{\overline{x_i^2}}{2\bar{x}_i}$$

$$= \sum_{i=1}^{P} \frac{\lambda_i \overline{x_i^2}}{2} \tag{3.7}$$

where $\overline{x_i^2}$ is the second moment of service time for a customer from group i.

Now we consider the second component of delay, namely, the delay due to customers found in the queue by our tagged customer who receive service before he does. We define

$$N_{ip} \triangleq \text{the number of customers from group } i \text{ found in the queue by our tagged customer (from group } p\text{) and who receive service before our tagged customer does} \tag{3.8}$$

where the average of this quantity is $E[N_{ip}] \triangleq \bar{N}_{ip}$. Since the service time for any member from group i is drawn independently from $B_i(x)$, the second component of average delay to our tagged customer is given by

$$\sum_{i=1}^{P} \bar{x}_i \bar{N}_{ip} \tag{3.9}$$

We may make similar statements regarding the third component (the delay to our tagged customer by later arrivals than he). Thus we define

$$M_{ip} \triangleq \text{the number of customers from group } i \text{ who arrive to the system while our tagged customer (from group } p\text{) is in the queue and who receive service before he does} \tag{3.10}$$

with average $\bar{M}_{ip}$. Thus we see that the third component of average delay is similar to that given in Eq. (3.9). Consequently, the total average delay in queue for our tagged customer may finally be written as

$$W_p = W_0 + \sum_{i=1}^{P} \bar{x}_i(\bar{N}_{ip} + \bar{M}_{ip}) \qquad p = 1, 2, \ldots, P \tag{3.11}$$

For any given priority queueing discipline the solution procedure then contains two steps: first, an evaluation of the averages $\bar{N}_{ip}$ and $\bar{M}_{ip}$; and second, a solution of the resulting set of equations (3.11).

In the general case both $\bar{N}_{ip}$ and $\bar{M}_{ip}$ may be expressed in terms of the average waiting times W_i, and therefore (3.11) leads to a set of simultaneous linear equations in the W_i. The simple approach herein described for calculating the average waiting times will be used in later sections of this chapter and is possible since the average of a sum is always equal to the sum of the averages. Higher moments are not so easily obtained and so in the next section we consider an approach for finding the *distribution* of waiting time for various priority groups.

The computation for *preemptive* queueing disciplines is similar to the above, but involves the additional complexity regarding how a customer recovers when he reenters service after having been preempted. Three cases are usually identified here. The first, where a customer picks up from where he left off (with perhaps a cost in time to either the customer or the system), is known as *preemptive resume*. The second and third cases assume that the customer loses credit for all service he has so far received: the second case assumes that a returning customer starts from scratch but with the same total service time requirement as he had upon his earlier visit, and this is known as *preemptive repeat without resampling*; the third case assumes that a *new* service time is chosen for our reentering customer and is referred to as *preemptive repeat with resampling*. (We study some examples of preemptive resume systems below and in Chapter 4.)

3.3. THE DELAY CYCLE, GENERALIZED BUSY PERIODS, AND WAITING TIME DISTRIBUTIONS

In this section we consider the analysis of "delay cycles," which permit us to calculate the Laplace transform for the pdf of "generalized" busy

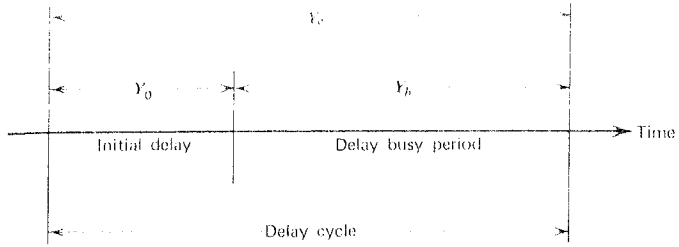

Figure 3.1 The delay cycle.

periods. From this we may obtain the Laplace transform for the waiting time density (as in Section 5.10 of Volume I [KLEI 75]) for a number of queueing disciplines. The concept of delay cycle analysis seems to have originated with Gaver [GAVE 62] (he used a notion known as "completion time"). Similar ideas appeared in Keilson [KEIL 62], who used the "basic server sojourn time," and in Avi-Itzhak and Naor [AVI 63], who used the "residence time." This work was extended by Miller [CONW 67], and it is his nomenclature (delay cycle) that we adopt here.

A delay cycle is similar to a busy period and is shown in Figure 3.1. The delay cycle Y_c consists of two portions: an initial delay of duration Y_0 and a delay busy period of duration Y_b. The initial delay is usually some special interval that may correspond to the completion of some partly completed customer or may correspond to some other "special" task. The delay busy period corresponds to the servicing of "ordinary" customers and may be viewed as a sequence of sub-busy periods*; the delay busy period ends when there are no more ordinary customers to be serviced. The generalization here over that of an ordinary busy period is that in the former we permit an arbitrary distribution for the initial delay, whereas in the latter we require that the initial delay be a service time distributed the same as the individual service times making up the elements of the delay busy period. In all cases, however, we have

$$Y_c = Y_0 + Y_b$$

The ordinary customers arrive according to a Poisson process; for purposes of this section we will assume that such customers arrive at a rate λ. We note that when the initial delay terminates, there may have accumulated during this initial delay a number of ordinary customers awaiting

* A sub-busy period is that interval of time which is required to service an arbitrary customer and all those (his "descendents") who enter the system during his service time or during the service time of any of his descendents. In the system M/G/1, the pdf for the duration of a sub-busy period is the same as that for a busy period. A sub-busy period is said to be "generated" by the customer who initiates the period. See Chapter 5 in [KLEI 75].

service; *each of these will generate his own sub-busy period, which, taken together, form the delay busy period.*

Now for some notation. Previously we had defined $B^*(s)$ and $G^*(s)$ as the transform for the pdf of service time and busy period durations, respectively, and the basic equation relating these two is given as Eq. (1.89). For the random variables Y_0, Y_b, and Y_c, we now define the PDF's $G_0(y)$, $G_b(y)$, and $G_c(y)$ along with the transforms (of the corresponding pdf's) denoted by $G_0^*(s)$, $G_b^*(s)$, and $G_c^*(s)$, respectively.

We assume that we are given $G_0^*(s)$, or if not we usually can calculate it. We are interested in solving for $G_b^*(s)$ and $G_c^*(s)$ in terms of the known functions $B^*(s)$, $G^*(s)$, and $G_0^*(s)$ where the relation given in Eq. (1.89) will be required to evaluate $G^*(s)$ from $B^*(s)$. The derivations we give here are rather abbreviated since they closely parallel the development of Section 5.8 of Volume I [KLEI 75]. We begin with the calculation of $G_b^*(s) = E[e^{-sY_b}]$. Let N_0 be the number of ordinary customer arrivals during the interval Y_0. We condition the transform we are seeking on Y_0 and N_0 to arrive at the simple expression

$$E[e^{-sY_b} \mid Y_0 = y, N_0 = n] = [G^*(s)]^n$$

This last follows from the fact that all n sub-busy periods (one is generated by each of the arrivals during Y_0) are independent and each is distributed exactly the same as a busy period. We proceed to uncondition first on N_0,

$$E[e^{-sY_b} \mid Y_0 = y] = \sum_{n=0}^{\infty} [G^*(s)]^n \frac{(\lambda y)^n}{n!} e^{-\lambda y}$$
$$= e^{-[\lambda - \lambda G^*(s)]y}$$

and finally on Y_0 to give

$$E[e^{-sY_b}] \triangleq G_b^*(s) = \int_{y=0}^{\infty} e^{-[\lambda - \lambda G^*(s)]y} \, dG_0(y)$$

This integral we recognize as the transform of the pdf for Y_0, and we therefore have the final result

$$G_b^*(s) = G_0^*(\lambda - \lambda G^*(s)) \qquad \blacksquare \quad (3.12)$$

Now for $G_c^*(s)$. Proceeding as above we have

$$E[e^{-sY_c} \mid Y_0 = y, N_0 = n] = e^{-\lambda y}[G^*(s)]^n$$

and removing the conditions on N_0 and Y_0 we have

$$G_c^*(s) \triangleq E[e^{-sY_c}] = \int_{y=0}^{\infty} e^{-sy} \sum_{n=0}^{\infty} [G^*(s)]^n \frac{(\lambda y)^n}{n!} e^{-\lambda y} \, dG_0(y)$$

which yields the result

$$G_c^*(s) = G_0^*(s + \lambda - \lambda G^*(s)) \qquad \blacksquare \quad (3.13)$$

Thus Eqs. (3.12) and (3.13) provide the defining equations for our unknowns where, of course, $G^*(s)$ is given in Eq. (1.89), that is,

$$G^*(s) = B^*(s + \lambda - \lambda G^*(s)) \qquad (3.14)$$

We note in the special case when Y_0 is distributed as an ordinary customer's service time that Y_c will merely be a regular busy period and Eq. (3.13) will then reduce to Eq. (3.14).

As we shall soon see, the delay cycle analysis is an extremely powerful method for obtaining results in many queueing systems, especially those with priorities.

3.4. CONSERVATION LAWS

In most physical systems, "you don't get something for nothing." So too in priority queueing systems—preferential treatment given to one class of customers is afforded at the expense of other customers. In a real sense then we "borrow from Peter to pay Paul." In this section we investigate such invariances or conservations within priority queueing systems.

Our conservation relations are based upon the fact that the unfinished work $U(t)$ during any busy period is independent of the order of service so long as the system is "conservative." By conservative we mean that no work (service requirement) is created or destroyed within the system; for example, destruction of work would occur if a customer were to leave the system before completing his service and the creation of work might correspond to a server standing idle in the face of a nonempty queue. Thus we consider only work-conserving systems in this section. The simplest case to consider is the FCFS system about which we know so much already. Most priority queueing systems are compared to the FCFS system and we see below that its performance enters our conservation relationships in a very natural way.

We begin by observing that the *distribution of waiting time* will indeed depend upon the order in which service is given. However, we now show that *so long as the queueing discipline selects customers in a way that is independent of their service time (or any measure of their service time) then the distribution of the number in the system will be invariant to the order of service; the same will also be shown to be true for the average waiting time of customers.* Let us consider the M/G/1 queue. For this system we have the basic relation given in Eq. (1.81). The definition for q_n was given as the number of customers left behind by the departure of C_n. Let us change

our point of view now and redefine this quantity to refer to the number of customers left behind by the departure of the nth *departing* customer (thereby allowing arbitrary order of service). Similarly v_n is to be interpreted as the number of customers arriving during the service of the nth customer to be served. It is clear that the relationship (1.81) now holds even for these more general queueing disciplines (and reduces to our former interpretation for the FCFS system). The identical steps (see [KLEI 75]) that take us from this relationship to an expression for the z-transform $[Q(z)]$ of the number of customers in an FCFS system will now take us from that defining equation to $Q(z)$ for a system with arbitrary order of service. Thus we can state for any queueing discipline whose decision rules are independent of a customer's service time that we must have the following as the z-transform for the number of customers in system:

$$Q(z) = B^*(\lambda - \lambda z)\frac{(1-\rho)(1-z)}{B^*(\lambda-\lambda z)-z} \tag{3.15}$$

where the notation here is the same as in Section 1.7. Therefore we immediately have complete information about the number in system. Bear in mind that this independence of order of service for number in system has only been shown to hold when the decision rule is itself independent of any aspect of service time of the customers.

Let us now conserve the unfinished work $U(t)$. From its definition, $U(t)$ is a function which (a) decreases at a rate of 1 sec/sec whenever $U(t) > 0$, (b) remains saturated at zero when it hits the horizontal axis, and (c) takes vertical jumps at the arrival instants in amounts equal to the service requirements brought in by the arrivals. Thus *it is clear that regardless of the order of service (service-dependent or not) $U(t)$ will not change; this is true for G/G/1.* For M/G/1 the following conservation law was first stated and proven in [KLEI 64a, 65]:

The M/G/1 Conservation Law. For any M/G/1 system and any non-preemptive work-conserving queueing discipline it must be that

$$\sum_{p=1}^{P} \rho_p W_p = \begin{cases} \dfrac{\rho W_0}{1-\rho} & \rho < 1 \\ \infty & \rho \geq 1 \end{cases} \tag{3.16}$$

[Recall from Section 1.7 and Eq. (3.7) that W_0 represents the residual life of the customer found in service upon an arrival's entry.] Thus this weighted sum of the waiting times W_p *can never change* no matter how sophisticated or elaborate the queueing discipline may be. Let us prove the validity of this conservation law. If at time t there are $N_p(t)$ customers

from group p in the queue and if the ith of these $[i = 1, 2, \ldots, N_p(t)]$ is to have a service time x_{ip} and if x_0 represents the work yet to be done on the man in service at time t (that is, his residual service time) then we may say

$$U(t) = x_0 + \sum_{p=1}^{P} \sum_{i=1}^{N_p(t)} x_{ip}$$

regardless of the order of service. Then taking expectations on both sides we have*

$$E[U(t)] = W_0 + \sum_{p=1}^{P} \sum_{n_p=0}^{\infty} P[N_p(t) = n_p] \sum_{i=1}^{n_p} E[x_{ip}]$$

We observe that $E[x_{ip}] = \bar{x}_p$ independent of the index i. With t taken at random (and large) we may write $\bar{U} \triangleq \lim_{t\to\infty} E[U(t)]$, which will be the limiting average of the unfinished work. Thus we may write

$$\begin{aligned} \bar{U} &= W_0 + \lim_{t\to\infty} \sum_{p=1}^{P} \sum_{n_p=0}^{\infty} n_p P[N_p(t) = n_p]\bar{x}_p \\ &= W_0 + \sum_{p=1}^{P} \bar{x}_p E[N_p] \end{aligned}$$

However, Little's result [Eq. (1.31)] tells us that $E[N_p] = \lambda_p W_p$ since this result is valid for individual priorities as well. Thus we conclude that

$$\bar{U} = W_0 + \sum_{p=1}^{P} \rho_p W_p \tag{3.17}$$

Now since $\bar{U}$ is independent of the order of service we may as well use our FCFS result, which states that for Poisson arrivals the average unfinished work (the average virtual waiting time) must equal the average waiting time for customers, which we denote by W. This quantity is given in Eq. (1.82); here the second moment of service time is easily expressed in terms of the second moment associated with each group's service time, namely

$$\overline{x^2} = \sum_{p=1}^{P} \frac{\lambda_p}{\lambda} \overline{x_p^2} = \frac{2W_0}{\lambda}$$

and so we may write

$$\bar{U} = W = \frac{W_0}{1-\rho} \tag{3.18}$$

* Here we take $E[x_0] = W_0$, where W_0 is defined in Eq. (3.7). This value for $E[x_0]$, which is the average unfinished work for the customer in service, is correct even for G/G/1 since we are *not* averaging over customer arrival instants, but are averaging uniformly over all time; as we know, Poisson arrivals also observe the system uniformly over all time and for this reason our result is the same as the mean residual service time seen by Poisson arrivals.

If we use this value for $\bar{U}$ in Eq. (3.17) we have the conservation law given in Eq. (3.16) (where for $\rho \geq 1$ the value of ∞ is obvious). Q.E.D.

Thus the conservation law puts a linear equality constraint on the set of average waiting times W_p. We see that any attempt to modify the queueing discipline so as to reduce one of the W_p will force an increase in some of the other W_p; however, this need not be an "even trade" since the weighting factors for the W_p are generally distinct. Now in the special case where $\bar{x}_p = \bar{x}$ for all p then the conservation law gives (for $\rho < 1$)

$$\sum_{p=1}^{P} \lambda_p W_p = \frac{\lambda W_0}{1-\rho} \qquad \bar{x}_p = \bar{x} \tag{3.19}$$

However, Little's result gives us $\lambda_p W_p = E[N_p]$ again and so

$$\sum_{p=1}^{P} E[N_p] = \frac{\lambda W_0}{1-\rho}$$

But this sum is merely the average total number in queue for which we have the notation $\bar{N}_q = E[\text{number in queue}]$ giving

$$\bar{N}_q = \frac{\lambda W_0}{1-\rho} = \text{constant}\Big|_{\substack{\text{queue} \\ \text{discipline}}} \qquad \bar{x}_p = \bar{x} \tag{3.20}$$

and of course from Little's result we further have

$$W = \frac{W_0}{1-\rho} = \text{constant}\Big|_{\substack{\text{queue} \\ \text{discipline}}} \qquad \bar{x}_p = \bar{x} \tag{3.21}$$

Thus in the special case where $\bar{x}_p = \bar{x}$, the average number in queue and the average waiting time in queue are independent of the queue discipline. Note the correspondence between this statement and our statement above regarding the invariance of the distribution of number of customers in the system (when the order of service was independent of service time). When $\bar{x}_p = \bar{x}$, then, at least regarding first moments, all customers behave the same with regard to service time and therefore order of service does not depend on average service time: this apparently leads to the invariance properties mentioned in Eqs. (3.20) and (3.21). If the average service times are not equal then it is not true in general that the average queue size and average waiting time are independent of queue discipline (which clearly depends upon (average) service time). These same statements, of course, apply to the average number in system (since we know that $\bar{N}_q + \rho = \bar{N}$) and for the average time in system (since $W + \bar{x} = T$). Note further that we are not claiming $W_p = W$ but merely that the sum in Eq. (3.19) is constant.

This conservation law has been extended [SCHR 70] to the case G/G/1 where not only are we dropping the Poisson arrival assumption but also no assumption regarding independence is required; what is required is that equilibrium distributions exist. In the above proof for M/G/1, we first used the Poisson arrival assumption following Eq. (3.17); however, that equation itself is good for G/G/1, and this gives us the generalized version of the conservation, namely

The G/G/1 Conservation Law

$$\sum_{p=1}^{P} \rho_p W_p = \bar{U} - W_0 \tag{3.22}$$

Of course for each problem $\bar{U}$ must be evaluated since this quantity is in general unknown for the system G/G/1! However, the notion of conservation stands out: Given a specific work-conserving G/G/1 queueing system with a nonpreemptive priority queueing discipline then the linear equality constraint given in Eq. (3.22) must be satisfied regardless of that queueing discipline. We note of course that Eq. (3.22) reduces to Eq. (3.16) in the case M/G/1.

It is not hard to conceive of priority queueing disciplines in which some groups experience finite waiting times while other groups find themselves in the abominable situation of experiencing infinite average waiting times; in fact the system considered below in Section 3.6 is an example of this. One inquires as to whether there exists a form of conservation law for those groups that do experience finite waits even in this unstable situation. We find in fact there do exist appropriate conservation laws that tell us more than does Eq. (3.16) in the case $\rho \geq 1$. This material may be found in [KLEI 65] and is elaborated on in Exercises 3.5 and 3.6 at the end of this chapter.

Equation (3.22) certainly applies for nonpriority G/G/1 systems as well. In this case from Eq. (3.7) we see that $W_0 = \lambda \overline{x^2}/2$ and so the conservation law for the nonpriority G/G/1 system becomes

$$\bar{U} = \rho W + \frac{\overline{x^2}}{2\bar{t}} \tag{3.23}$$

where we have written $\bar{t} = 1/\lambda$ to correspond with our more usual notation for G/G/1. Brumelle [BRUM 69] also derives this as a special case of his general class of formulas of the form, "time average = $\lambda \cdot$ customer average." The most famous example of this form of equation is of course Little's formula; so too is Eq. (3.23), which is also rather important and for some reason seems to be somewhat obscure in the queueing literature. Brumelle permits the case where there is dependence

among the basic processes, and in this case the term ρW is replaced by $E[\tilde{x}\tilde{w}]/\bar{T}$.

3.5. THE LAST-COME-FIRST-SERVE QUEUEING DISCIPLINE

Let us return to the M/G/1 queue again and consider the case in which service is given to the most recent arrival on a nonpreemptive basis. Here we have $P = 1$ (no externally assigned priorities). This order of service is not uncommon, strange as it may appear; for example, any push-down stack operates in this fashion. Since the decision rule is independent of service time, we see immediately that the average queue size and the average waiting time must be the same as for FCFS [see Eqs. (3.20) and (3.21)]. Moreover, we know that Eq. (3.15) gives the distribution of number in the system. However, we suspect that the waiting time distribution differs from the FCFS case, and it is this which we solve for below. Our intuiton correctly suggests that this rule will give a large variance of waiting time even though the average is the same as FCFS.

This queueing discipline lends itself especially well to analysis. We observe that a new arrival is in no way affected by the queue size he finds upon his entry to the system; only the customer found in service can make him wait and the balance of his delay is due to arrivals that enter the system after he does but prior to his initiation of service. This is a perfect set-up for the delay cycle analysis of Section 3.3 where the initial delay is the residual life of the customer found in service and the delay busy period is the interval required to empty the system of all those arrivals who follow him prior to his entry into service, at which point his service commences. The Laplace transform for the residual life pdf is given in the footnote on page 16; we rewrite this transform using our notation for delay cycle analysis as

$$G_0^*(s) = \frac{1 - B^*(s)}{s\bar{x}}$$

Moreover, $G_c^*(s)$, the Laplace transform for the delay cycle pdf, is given in terms of $G_0^*(s)$ and $G^*(s)$ in Eqs. (3.13) and (3.14). The delay cycle here corresponds to the waiting time for our customer in this LCFS system, and so using these transform relations we may write down the conditional transform for waiting time as

$$\begin{aligned} E[e^{-sw} \mid \text{system busy upon arrival}] &= G_c^*(s) \\ &= G_0^*(s + \lambda - \lambda G^*(s)) \\ &= \frac{1 - B^*(s + \lambda - \lambda G^*(s))}{[s + \lambda - \lambda G^*(s)]\bar{x}} \end{aligned}$$

Then from Eq. (3.14) we may simplify the numerator to give

$$E[e^{-sw} \mid \text{system busy upon arrival}] = \frac{1 - G^*(s)}{[s + \lambda - \lambda G^*(s)]\bar{x}}$$

If we now uncondition this expression, we find that with probability $1 - \rho$ our customer has a waiting time of zero and with probability ρ he has a waiting time whose transform is given in this last equation. Thus

$$W^*(s) = E[e^{-sw}] = 1 - \rho + \frac{\lambda[1 - G^*(s)]}{s + \lambda - \lambda G^*(s)} \qquad (3.24)$$

and this is the result we were seeking. We note that it differs significantly from the Pollaczek–Khinchin transform equation for FCFS given in Eq. (1.85). In Exercise 3.4 we compare the mean and variance of waiting time for these two systems. We find that the first moments are of course the same (as we stated earlier) but that the variance for LCFS is larger than for FCFS.

Let us now consider the case where an *external* priority structure is imposed.

3.6. HEAD-OF-THE-LINE PRIORITIES

Among the queueing disciplines that impose an *external* priority structure on the arriving customers, the head-of-the-line (HOL) priority queueing system is perhaps the most common and most natural. This system, first studied in 1954 by Cobham [COBH 54] is known also by the name of strict priority queueing or fixed priority queueing. The system structure is given in Figure 3.2. In this system customers queue according to priority groups and are strictly separated on the basis of the group to which they belong. Thus an arrival from group p joins the "torso" of the queue behind all customers from group p (and higher) and in front of all customers from group $p - 1$ (and lower). The value of one's priority in this case remains constant in time and so we may take the priority

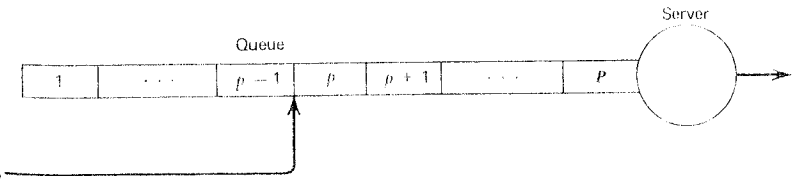

Figure 3.2 Head-of-the-line priority queue.

function to be

$$q_p(t) = p \tag{3.25}$$

Let us now use the method of Section 3.2 to derive the average waiting time W_p for members of the pth priority group in the case of a nonpreemptive HOL system. Equation (3.11) is our point of departure and we must evaluate the two functions $\bar{N}_{ip}$ and $\bar{M}_{ip}$ that represent the average number of customers from priority group i who get served before our tagged customer (from group p) and who are present in the queue upon his arrival ($\bar{N}_{ip}$) or who arrive while he is in the queue ($\bar{M}_{ip}$). Because of the strict order of queueing and under the assumption that customers within the same priority group get served according to an FCFS rule, it is clear that

$$\bar{N}_{ip} = 0 \qquad i = 1, 2, \ldots, p-1$$
$$\bar{M}_{ip} = 0 \qquad i = 1, 2, \ldots, p$$

All customers from group p and higher who are present in the queue upon our tagged customer's arrival must certainly get served before he does; from Little's result we know that on the average there will be $\lambda_i W_i$ customers from the ith group present in the queue when our tagged customer arrives and therefore,

$$\bar{N}_{ip} = \lambda_i W_i \qquad i = p, p+1, \ldots, P \tag{3.26}$$

Similarly all customers from groups $p+1, p+2, \ldots, P$ who join the system while our tagged customer is in the queue will also be served before he is; since he spends on the average W_p sec in the queue and since each group's arrival process is independent of queue size we know that there will on the average be $\lambda_i W_p$ customer arrivals from the ith group while our tagged customer waits on queue. Therefore,

$$\bar{M}_{ip} = \lambda_i W_p \qquad i = p+1, p+2, \ldots, P \tag{3.27}$$

Thus for the nonpreemptive HOL system Eq. (3.11) becomes

$$W_p = W_0 + \sum_{i=p}^{P} \bar{x}_i \lambda_i W_i + \sum_{i=p+1}^{P} \bar{x}_i \lambda_i W_p \qquad p = 1, 2, \ldots, P \tag{3.28}$$

By straightforward arguments we have arrived at the set of defining equations for our unknowns W_p. Solving Eq. (3.28) for W_p we have

$$W_p = \frac{W_0 + \sum_{i=p+1}^{P} \rho_i W_i}{1 - \sum_{i=p}^{P} \rho_i} \qquad p = 1, 2, \ldots, P \tag{3.29}$$

This set of equations may be solved recursively with no difficulty since we have a triangular set; that is, we first find W_P and from this find W_{P-1}, and so on. We find it convenient to define

$$\sigma_p = \sum_{i=p}^{P} \rho_i \tag{3.30}$$

Solving recursively, we obtain the solution

$$W_p = \frac{W_0}{(1-\sigma_p)(1-\sigma_{p+1})} \qquad p = 1, 2, \ldots, P \tag{3.31}$$

This last equation was one of Cobham's principal contributions to this problem, and its form is rather suggestive. In particular we see the effect of those customers of equal or higher priority *present* in the queue when our customer arrives as given by the denominator term $1-\sigma_p$ and also the effect of customers of higher priority *arriving* during our customer's queueing time as given by the denominator term $1-\sigma_{p+1}$. Furthermore we notice that W_p does not depend on customers from lower priority groups (that is, for $i = 1, 2, \ldots, p-1$) except for their contribution to the numerator W_0. Thus the solution given in Eq. (3.31) demonstrates that some W_p may be finite (for $p \geq$ some critical value) while other lower priority groups may be experiencing unstable (unbounded) queueing times.† Figure 3.3(a) demonstrates this stable–unstable behavior for a system with $P = 5$ groups. In this figure we have plotted the normalized waiting time $W_p/\bar{x}$ since this is a useful dimensionless form. Also in this figure we gain our first experience with the conservation law [Eq. (3.16)] by observing the dashed line that represents the average waiting time for the FCFS system; this dashed curve is a plot of Eq. (3.18). In Figure 3.3(b) the same curves are shown on an expanded scale; there one may observe the way in which the conservation law is functioning. In particular if one measures these curves, it will be seen that $\sum_p (\rho_p/\rho) W_p = W_0/(1-\rho)$; for $\rho > 1$ it is clear that this average blows up to infinity.

From the method described in Section 3.3 it is possible to find the distribution of waiting time for each priority group. Let us denote the Laplace transform for the pth group's waiting time in queue by $W_p^*(s)$. The solution is [CONW 67, KEST 57]:

$$W_p^*(s) = \frac{(1-\rho)[s + \lambda_H - \lambda_H G_H^*(s)] + \lambda_1 [1 - B_1^*(s + \lambda_H - \lambda_H G_H^*(s))]}{s - \lambda_p + \lambda_p B_p^*(s + \lambda_H - \lambda_H G_H^*(s))} \tag{3.32}$$

† We note that the pth group experiences finite waiting time so long as $\rho < 1 + \rho_1 + \rho_2 + \cdots + \rho_{p-1}$.

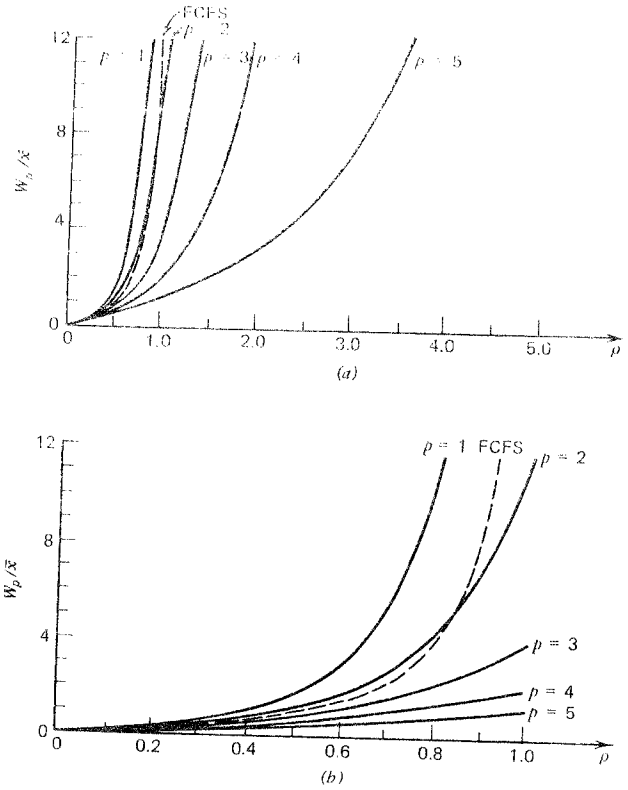

Figure 3.3 HOL with no preemption: $P = 5$, $\lambda_p = \lambda/5$, $\bar{x}_p = \bar{x}$.

where

$$\lambda_H = \sum_{i=p+1}^{P} \lambda_i \tag{3.33}$$

$$\lambda_L = \sum_{i=1}^{p-1} \lambda_i \tag{3.34}$$

$$B_H^*(s) = \sum_{i=p+1}^{P} \frac{\lambda_i}{\lambda_H} B_i^*(s) \tag{3.35}$$

$$B_L^*(s) = \sum_{i=1}^{p-1} \frac{\lambda_i}{\lambda_L} B_i^*(s) \tag{3.36}$$

$$G_H^*(s) = B_H^*(s + \lambda_H - \lambda_H G_H^*(s)) \tag{3.37}$$

and of course by our usual notation $B_i^*(s)$ corresponds to the Laplace transform of the pdf for the ith group's service time. In these definitions we have used the subscript H to denote the set of priority classes higher than our tagged unit and the subscript L to denote those lower. We observe that the definition in Eq. (3.37) is the same as the functional equation given in (3.14); also the form for $W_p^*(s)$ is surprisingly similar to that for the Pollaczek–Khinchin transform equation for waiting time in the FCFS system. Specifically we note that in the case $P = 1$ then $\lambda_H = \lambda_L = B_H^*(s) = B_L^*(s) = G_H^*(s) = 0$ and $\lambda_p = \lambda$, and so $W_1^*(s)$ reduces to the P-K transform equation for the FCFS system, as of course it must. From Eq. (3.32) one may calculate the mean waiting time given in Eq. (3.31) by differentiation.

Since the various priority groups receive different grades of service, and since they may each have different distributions of service time, then in a real sense we are discriminating on the basis of (distribution of) service time; therefore, whereas the conservation law certainly holds, the distribution for number in system will differ from that of the FCFS system. Let us in fact attempt to discriminate completely on the basis of *exactly known* service times. We may accomplish this as follows with the HOL model defined above. In particular let us create a priority queueing discipline in which highest priority is given to that job with the shortest service time (that is, an SJF system). This was considered in [PHIP 56] in which a continuum of priority classes was defined such that the class index p was defined to be any strictly decreasing function of the service time $\tilde{x}$. Thus we have a model in which an entering customer whose service time is known to be exactly x sec joins the queue behind all other customers with service times less than (or equal to) x and in front of all customers in the queue with service times greater than x (note that in the case where the pdf of overall service time has impulses, then any ties are broken by the FCFS rule). This is an M/G/1 queueing system in which we assume that customer service times are chosen from $B(x)$ prior to their arrival and therefore they may be ordered in the queue immediately upon entry as described above. Customers whose service times fall in the interval $x < \tilde{x} \le x + dx$ are all grouped into the same priority group, and the service time density associated with this group is merely a unit impulse at x sec of service; of course the fraction of customers who fall in this group is merely $b(x)\,dx$, and this will be an infinitesimal quantity unless $B(x)$ has a discontinuity at x. Such is the nature of our continuum of priority groups. Let us now calculate the average waiting time $W(x)$ for a customer whose service time lies in the interval $(x, x + dx)$. Recall that the priority index p is a strictly decreasing function of service time $\tilde{x}$.

Therefore in the limit Eq. (3.30) becomes

$$\sum_{i=p}^{P} \rho_i \to \int_{x=0}^{x^+} \rho(y)\, dy$$

where $\rho(x) = \lambda(x)x$ and $\lambda(x) = \lambda b(x)$; this is the correct expression for ρ since the average service time for such customers is exactly x sec and the average arrival rate of such customers is $\lambda\, dB(x)/dx$. Therefore Eq. (3.31) takes on the following limiting value for $W(x)$:

$$W(x) = \frac{W_0}{\left[1-\lambda\int_0^x yb(y)\,dy\right]\left[1-\lambda\int_0^{x^+} yb(y)\,dy\right]} \qquad (3.38)$$

and we note that the denominator reduces to $[1-\lambda\int_0^x yb(y)\,dy]^2$ when $B(x)$ is continuous at x. Here, as in the discrete case, our solution applies only for those "priority" groups that enjoy finite average waiting times. Thus Eq. (3.38) gives the average wait in an SJF queueing discipline for a customer whose service time is x. Note for a customer with a very long service time that $\lim_{x\to\infty} W(x) = W_0/(1-\rho)^2$, whereas for extremely short customers, we have $\lim_{x\to 0} W(x) = W_0$.

Let us now consider an HOL system with a *preemptive* queueing discipline. For this case we assume that the preemption is of the preemptive resume type. The approach here is much like that described in Section 3.2 and proceeds as follows. Recalling that T_p is the average of the total time spent in system by our tagged customer from group p, we recognize that his average delay consists of three components. The first is his average service time $\bar{x}_p$. Second, there is the delay due to the service (work) required by those customers of equal or higher priority whom he finds in the system; by our conservation results, we see that our tagged customer finds an average amount of work in the system equal to $(\sum_{i=p}^{P} \lambda_i \overline{x_i^2}/2)/(1-\sigma_p)$, which must be done before he gets served [the mean work backlog is equal to the mean wait in M/G/1 and so we recognize this term as equal to the expression $W_0/(1-\rho)$ for a system fed only by groups $p, p+1, \ldots, P$; the other groups are completely "invisible" to our tagged customer!]. Third, he will be delayed by any customers who enter the system before he leaves and who are members of strictly higher priority groups; the average number of such arrivals from group i must be $\lambda_i T_p$, each of which delays our tagged customer by an average of $\bar{x}_i$ sec. Thus we may say

$$T_p = \bar{x}_p + \frac{\sum_{i=p}^{P} \lambda_i \overline{x_i^2}/2}{1-\sigma_p} + \sum_{i=p+1}^{P} \rho_i T_p$$

The solution for T_p is therefore

$$T_p = \frac{\bar{x}_p(1-\sigma_p) + \sum_{i=p}^{P} \lambda_i \overline{x_i^2}/2}{(1-\sigma_p)(1-\sigma_{p+1})} \qquad (3.39)$$

Let us now pose an interesting *optimization* problem whose solution is within our grasp. It seems natural for us to ask for some guidance in assigning external priorities to customers. We consider the nonpreemptive case below. Let us assume that there is a system cost (rate) of C_p dollars for each second of delay suffered by each customer from priority group p; it is then clear that the average cost per second to the system, which we denote by C, must be

$$C = \sum_{p=1}^{P} C_p \bar{N}_p$$

where $\bar{N}_p$ is merely the average number of type p customers in the system. Of course from Little's result we know that regardless of the queueing discipline it must be that $\bar{N}_p = \lambda_p T_p = \lambda_p[W_p + \bar{x}_p]$ and so we have

$$C = \sum_{p=1}^{P} \rho_p C_p + \sum_{p=1}^{P} C_p \lambda_p W_p \qquad (3.40)$$

We desire to find that nonpreemptive (work-conserving) queueing discipline which minimizes C. We will solve this problem for a given M/G/1 system with P priority groups, an average arrival rate of λ_p type-p customers per second, and a service time distribution for type-p customers given by $B_p(x)$. Let us rewrite Eq. (3.40) to bring the *constant* sum to the left-hand side as follows:

$$C - \sum_{p=1}^{P} \rho_p C_p = \sum_{p=1}^{P} (C_p/\bar{x}_p)(\rho_p W_p)$$

and it is the sum on the right-hand side that we must minimize by an appropriate choice of queueing discipline. Let $f_p = C_p/\bar{x}_p$ (a given quantity) and $g_p = \rho_p W_p$ (a design variable through W_p), and so we are asking to minimize the following sum of products $\sum_{p=1}^{P} f_p g_p$. However, from the conservation law given in Eq. (3.16) we know that

$$\sum_{p=1}^{P} \rho_p W_p = \text{constant with respect to queue discipline} \qquad (3.41)$$

That is, we wish to minimize the "area" under the product of two functions, one of which itself has a constant area. Now if we reorder the subscripts so that

$$f_1 \le f_2 \le \cdots \le f_P \qquad (3.42)$$

we see that the optimum way in which we can match the terms g_p against

the terms f_p [under the constraint in Eq. (3.41)] is to assign as little "mass" as possible in g_P to match the largest term f_P. Having done this we will then assign as little of the remaining mass in g_{P-1} against the next largest term f_{P-1}, and so on. Now from its definition ($g_P = \rho_P W_P$) we see that since ρ_P is a given constant then we will minimize g_P by minimizing W_P. The nonpreemptive work-conserving queueing discipline that minimizes W_P is clearly HOL with the highest priority group corresponding to P (as usual). Having accomplished this, W_{P-1} may next be minimized by making this the second highest priority group in an HOL system, and so on. Thus we see that *the solution to our optimization problem is that, of all the possible nonpreemptive work-conserving queueing disciplines, the HOL discipline with the ordering given in Eq. (3.42)* is that which minimizes the average cost given in Eq. (3.40)!* This proof depended upon the conservation law but also could have been established by the more usual interchange argument [COX 61] (in which only stationary disciplines are considered—a sufficient class as recently shown [LIPP 75]). It is truly amazing that such a result is obtainable and so simply.

3.7. TIME-DEPENDENT PRIORITIES

The reason for imposing a priority structure on the customer arrivals is to provide preferential treatment to the "higher priority" groups at the expense of the "lower priority" groups. We have shown above in the case of linear cost rates that the HOL priority system is optimum in that it minimizes the average (linear) cost. However, this linear cost function is not always suitable and in fact the appropriate cost function is often unknown. In spite of this, the practical world is abundant with examples of priority queueing systems for which decisions have been made regarding the relative desired performance among classes (and which therefore imply some form of cost function, perhaps unknown to the users or the system). For example, most military systems use an HOL discipline with preemption permissible by the highest priority groups and usually with four or five groups in all. Another example is in the servicing of automobiles at the repair station in which the mechanic selects from among those automobiles waiting for service that one with perhaps the shortest (or perhaps the most expensive) service requirement. Often, therefore, rather than specifying the cost function, one is willing to specify the relative waiting times among the various priority groups. For example, a system designer may be required to synthesize a priority queueing

* Taking $\bar{x}_p = 1/\mu_p$, we have $f_p = \mu_p C_p$, and so this optimum ordering is referred to as "the μC rule."

discipline in which the desired performance is given as the ratios W_{p-1}/W_p ($p = 1, 2, \cdots, P-1$). The designer is then faced with achieving these ratios for a given specification of the customer behavior; that is, we assume the arrivals are of the M/G/1 type with given quantities λ_p, $\bar{x}_p$, and $B_p(x)$. Therefore from their definitions the quantities ρ_p, σ_p, and W_0 are also specified [in fact, $B_p(x)$ need not be specified but only $\bar{x}_p$ and $\overline{x_p^2}$ are necessary to determine these three system parameters]. From Eq. (3.31), therefore, the behavior of W_p is completely determined for HOL and as a consequence so is the behavior of the ratios W_{p+1}/W_p. Unfortunately this freezes the design and so the independently specified ratios cannot, in general, be achieved for the HOL system!

Therefore we must introduce some additional degrees of freedom into our priority queueing discipline if we are to meet the required specifications. The time-dependent priority system described below provides a set of variable parameters b_p where $0 \le b_1 \le b_2 \le \cdots \le b_P$; these parameters are at the disposal of the designer in adjusting these relative waiting times [KLEI 64a, b].

Let us assume that some tagged customer arrives at time τ and is assigned at time t a priority $q_p(t)$ calculated from

$$q_p(t) = (t - \tau) b_p$$

where t ranges from τ until the time at which this customer's service is completed. We consider the following nonpreemptive system. Whenever the service facility is ready for a new customer, that customer with the highest instantaneous priority $q_p(t)$ is then taken into service [that is, at time t, a customer with priority $q'(t)$ is given preferential treatment over a customer with priority $q(t)$ where $q'(t) > q(t)$]. Whenever a tie for the highest priority occurs, the tie is broken by an FCFS rule. We note that higher priority customers gain priority at a faster rate (b_p) than lower priority customers.

Figure 3.4 shows an example of the manner in which this priority

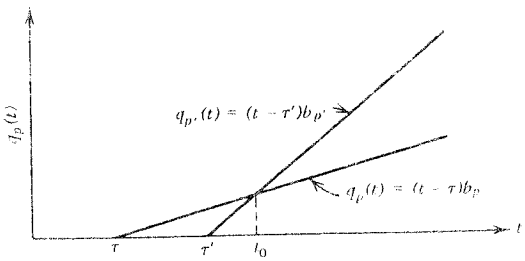

Figure 3.4 Interaction between priority functions for the delay-dependent priority system.

structure allows interaction between the priority functions for two customers. Specifically, at time τ, a customer from priority group p arrives, and attains priority at a rate equal to b_p. At time τ' a different customer enters from a higher priority group p'; that is, $p' > p$. When the service facility becomes free, it next chooses that customer in the queue with the highest instantaneous priority. Thus, in our example, the first customer will be chosen in preference to the second customer if the service facility becomes free at any time between τ and t_0 (in spite of the fact that the first customer is from a "lower" priority class); but, for any time after t_0, the second customer will be chosen in preference to the first.

We study this system for the case of exponential service times. We use the method in Section 3.2 and are faced immediately with calculating the quantities $\bar{N}_{ip}$ and $\bar{M}_{ip}$. We begin with the calculation of $\bar{M}_{ip}$ and refer the reader to Figure 3.5. Clearly, $\bar{M}_{ip} = 0$ for $i \le p$ since no later arrivals with smaller (or equal) slope can ever "catch up" to the tagged customer. Now consider the arrival of a p-type customer, the tagged customer, 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 3.5. In looking for $\bar{M}_{ip}$, we must calculate how many i-type customers (for $i > p$) 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 customers that arrive in the time interval $(0, V_i)$ will satisfy these conditions. Let us calculate the value of V_i. Clearly,

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

and so

$$V_i = W_p \left[1 - \frac{b_p}{b_i} \right]$$

Therefore, with an input rate λ_i for the type-i customers, we find that

$$\bar{M}_{ip} = \lambda_i V_i$$

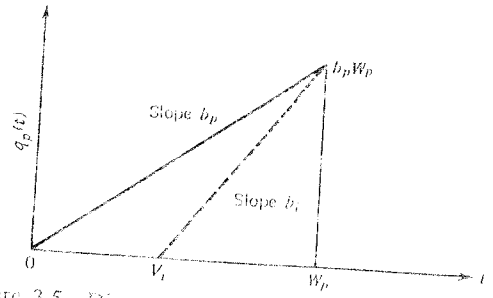

Figure 3.5 Diagram of priority, $q_p(t)$, for obtaining $\bar{M}_{ip}$.

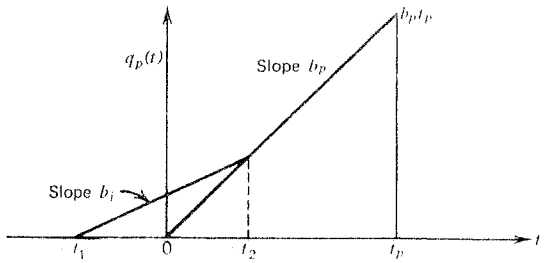

Figure 3.6 Diagram of priority, $q_p(t)$, for obtaining $\bar{N}_{ip}$.

and so

$$\bar{M}_{ip} = \lambda_i W_p \left[1 - \frac{b_p}{b_i} \right] \qquad \text{for all } i > p \tag{3.43}$$

We now prove that $\bar{N}_{ip} = \lambda_i W_p (b_i / b_p)$ for $i \le p$. Consider again that a type-p customer, the tagged customer, arrives at time $\tau = 0$ and spends a total time t_p in the queue. His attained priority at the time of his acceptance into the service facility will be $b_p t_p$, as shown in Figure 3.6. Upon his arrival, the tagged customer finds n_i type-i customers already in the queue. Let us consider one such type-i customer, as shown in the figure, which arrived at $t = -t_1$. In looking for $\bar{N}_{ip}$, we must calculate how many type-i customers arrive before $t = 0$, are still in the queue at $t = 0$, and obtain service before the tagged customer does. It is obvious from the figure that a type-i customer that arrives at time $-t_1$ ($t_1 > 0$) and that waits in the queue a time $w_i = w_i(t_1)$ such that $t_1 < w_i(t_1) \le t_1 + t_2$ will satisfy these conditions. Obviously, $w_i(t_1)$ must not exceed $t_1 + t_2$ since otherwise the i-type customer will be of lower priority than the tagged customer, and will therefore fail to meet the conditions stipulated above. Let us solve for t_2. Clearly,

$$b_p t_2 = b_i (t_1 + t_2)$$

and so

$$t_2 = \left[\frac{b_i}{b_p - b_i} \right] t_1$$

or

$$t_1 + t_2 = \left[\frac{b_p}{b_p - b_i} \right] t_1 \tag{3.44}$$

It is therefore clear that the expected number, $\bar{N}_{ip}$, of i-type customers that are in the queue at $t = 0$ and that also obtain service before the tagged customer does, can be expressed as

$$\bar{N}_{ip} = \int_0^\infty \lambda_i P\left\{ t < w_i(t) \le \left[\frac{b_p}{b_p - b_i} \right] t \right\} dt \tag{3.45}$$

where $\lambda_i\,dt$ is the expected number of i-type customers that arrived during the time interval $(-t-dt,-t)$ and where $P\{t-w_i(t)=[b_i/(b_p-b_i)]t\}$ is the probability that a customer who arrived in that interval spends at least t and at most $[b_i/(b_p-b_i)]t$ see in the queue. Equation (3.45) can be written as

$$\bar{N}_{ip}=\lambda_i\int_0^\infty[1-P(w_i\le t)]\,dt-\lambda_i\int_0^\infty\left[1-P\left\{w_i\le\left[\frac{b_i}{b_p-b_i}\right]t\right\}\right]dt$$

$$=\lambda_i\int_0^\infty[1-P(w_i\le t)]\,dt-\lambda_i\left[1-\left(\frac{b_i}{b_p}\right)\right]\int_0^\infty[1-P(w_i\le y)]\,dy$$

where we have made the change of variable $y=[b_i/(b_p-b_i)]t$. Now, since

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

(for w_i a non-negative random variable), and since in our notation $W_i=E[w_i]$, we obtain

$$\bar{N}_{ip}=\lambda_i W_i-\lambda_i\left[1-\frac{b_i}{b_p}\right]W_i$$

and therefore

$$\bar{N}_{ip}=\lambda_i W_i\frac{b_i}{b_p}\qquad\text{for all } i\le p \tag{3.46}$$

Furthermore, it is clear from Eq. (3.26) that $\bar{N}_{ip}=\lambda_i W_i$ for $i\ge p$ since our tagged customer can never catch up with these higher priority customers (all of whom are present upon his arrival).

Having derived expressions for $\bar{N}_{ip}$ and $\bar{M}_{ip}$ we may now substitute for these quantities into Eq. (3.11) and obtain

$$W_p=\frac{W_0+\sum_{i=p}^{P}\rho_i W_i+\sum_{i=1}^{p-1}\rho_i W_i(b_i/b_p)}{1-\sum_{i=p+1}^{P}\rho_i[1-(b_p/b_i)]}\qquad p=1,2,\ldots,P \tag{3.47}$$

This set of P linear equations in the W_p's is sufficient to solve our problem. However, it is possible to create a much simpler triangular set of equations from these by making use of the conservation law given in Eq. (3.16). Specifically we may rewrite the first sum in the numerator from Eq. (3.47) as follows:

$$\sum_{i=p}^{P}\rho_i W_i=\frac{\rho W_0}{1-\rho}-\sum_{i=1}^{p-1}\rho_i W_i$$

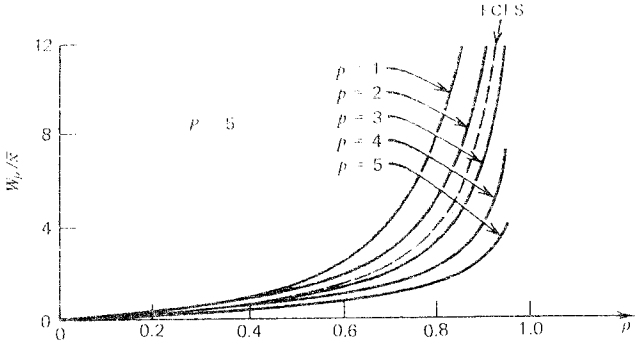

Figure 3.7 $W_p/\bar{x}$ for the time-dependent priority system with no preemption. $P=5$, $\lambda_p=\lambda/5$, $\bar{x}_p=\bar{x}$.

and when this substitution is made we have

$$W_p=\frac{[W_0/(1-\rho)]-\sum_{i=1}^{p-1}\rho_i W_i[1-(b_i/b_p)]}{1-\sum_{i=p+1}^{P}\rho_i[1-(b_p/b_i)]}\qquad p=1,2,\ldots,P \tag{3.48}$$

which is the main result for this nonpreemptive time-dependent discipline. It is interesting to note the extremely simple dependence that W_p has on the parameters b_i, namely these parameters only appear as ratios [KLEI 64a, b].

The typical behavior for this time-dependent queueing discipline is shown in Figure 3.7. The dashed curve shown is that for the FCFS system and once again shows the effect of the conservation law on priority queueing disciplines.

Thus we have analyzed a queueing discipline that provides a free set of parameters b_p/b_{p+1} that may be used to meet the specified system performance requirements given as W_p/W_{p+1} for $p=1,2,\ldots,P-1$. We see that only $P-1$ performance ratios may be specified and that we have exactly this many degrees of freedom for meeting that specification; the Pth condition is forced upon us as a scaling factor (that is, the conservation law) for all the waiting times and of course is a function of the utilization of the system (it is also clear that W_P for this class of systems can never lie below the corresponding curve for the HOL system, because of the ultimate preference given to the highest priority group in HOL).

A natural extension to this discipline is one in which a customer's priority increases in proportion to some *arbitrary power* of his elapsed

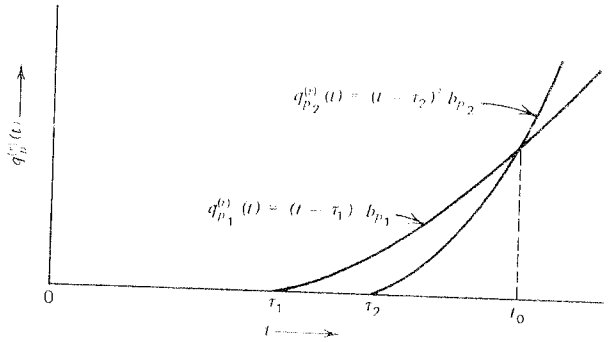

Figure 3.8 Coupling of different priority units.

time, rather than the first power as above [KLEI 67a]. We now address ourselves to this "generalization." Thus we define, for any non-negative number r, an rth-order time-dependent priority discipline as one that calculates the priority $q_p^{(r)}(t)$ at time t associated with a customer arriving at time τ as follows:

$$q_p^{(r)}(t) = (t-\tau)^r b_p$$

Further define $W_p^{(r)}$ as the expected value of the time spent in the queue of an rth-order system for a unit from group p.

The coupling between customers of different priority classes is illustrated in Figure 3.8. As for $r=1$, we see that it is possible for customers to change their relative positions in the queue. It should be noted that there can be at most only *one* interchange between any two customers and it is this property which simplifies the analysis.

Consider two generalized time-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 Exercise 3.14, we prove that if we choose

$$\left(\frac{b_p}{b_{p+1}}\right)^{1/r} = \left(\frac{b'_p}{b'_{p+1}}\right)^{1/r'} \qquad p = 1, 2, \ldots, P-1 \tag{3.49}$$

then

$$W_p^{(r)} = W_p^{(r')} \tag{3.50}$$

Thus 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 Eq. (3.49). The case for $r_0 = 1$ has already been treated above and so, in order to obtain $W_p^{(r)}$, we appeal to these results and obtain (see Exercise 3.14) the main result: for an

rth-order delay-dependent priority system without preemption, and $0 \le \rho < 1$,

$$W_p^{(r)} = \frac{[W_0/(1-\rho)] - \sum_{i=1}^{p-1} \rho_i W_i [1-(b_i/b_p)^{1/r}]}{1 - \sum_{i=p+1}^{P} \rho_i [1-(b_p/b_i)^{1/r}]} \qquad p = 1, 2, \ldots, P \tag{3.51}$$

This result we see is in basically the same form as $r=1$. Furthermore from the result in Eq. (3.50) we see that no greater generality for W_p is afforded with arbitrary r since they are all equivalent to our earlier result for $r=1$.

However, there is insight to be gained from this generalization. In Figure 3.9 we show $q_p^{(r)}(t)$ for a customer arriving at time τ, with r as a parameter. We see that

$$\lim_{r\to 0} q_p^{(r)}(t) = b_p u_{-1}(t-\tau) \tag{3.52}$$

where $u_{-1}(t-\tau)$ is the unit step function occurring at time τ. Thus for $r=0$, an entering customer from group p is assigned a fixed value of priority equal to b_p. This is the HOL system studied above. Moreover, as $r \to \infty$, $q_p^{(r)}(t)$ becomes a step function of infinite height at time $\tau+1$. Thus, customers that have been in the system for more than 1 sec have infinite priority and those that have been in the system for less than 1 sec have zero priority. Remembering that an FCFS criterion is used to break a tie, the limit as r approaches infinity is seen to be a strict FCFS system. These

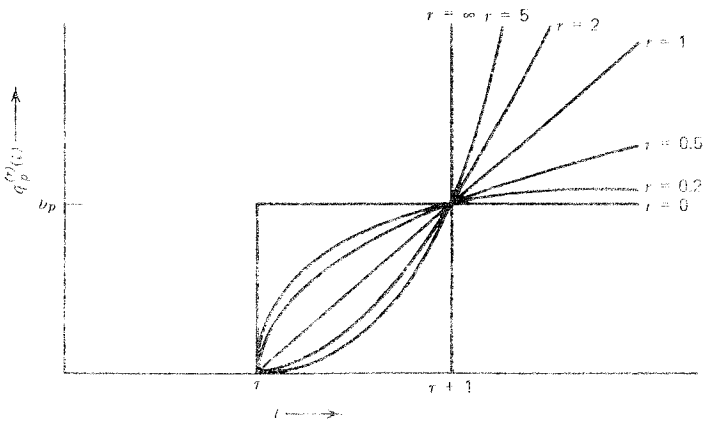

Figure 3.9 $q_p^{(r)}(t)$ for several r

two limiting cases can also be obtained by taking the limit of $W_p^{(r)}$ as follows. From Eq. (3.51), for $b_p < b_{p+1}$ $(p = 1, 2, \ldots, P-1)$,

$$\lim_{r \to 0} W_p^{(r)} = \lim_{(b_p/b_{p+1})^{1/r} \to 0} W_p^{(r)} = \frac{W_0/(1-\rho) - \sum_{i=1}^{p-1} \rho_i W_i}{1 - \sum_{i=p+1}^{P} \rho_i}$$

Solving this last set of recursive equations yields

$$\lim_{r \to 0} W_p^{(r)} = \frac{W_0}{\left(1 - \sum_{i=p}^{P} \rho_i\right)\left(1 - \sum_{i=p+1}^{P} \rho_i\right)} \tag{3.53}$$

which is the same result obtained for HOL. We also note that $\lim_{r \to \infty} (b_p/b_{p+1})^{1/r} = 1$ and so

$$\lim_{r \to \infty} W_p^{(r)} = \frac{W_0}{1-\rho} \tag{3.54}$$

which is the result for FCFS.

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 the HOL system and as $r \to \infty$ we obtain the FCFS system. For $r = 1$ we have the first-order time-dependent system. In Figure 3.10 we illustrate the general behavior of the expected wait on queue as we vary our priority discipline over the class of rth-order systems for $0 \le r$. We show the case with $P = 5$, $b_p/b_{p+1} = \frac{1}{2}$, $\rho_p = \rho/5$, $\bar{x}_p = \bar{x}$, for $p = 1, 2, \ldots, 5$, $\rho = 0.95$, and $W_0 = 1$. The dashed line in this figure demonstrates the conservation law for this particular case. The wide dispersion of $W_p^{(r)}$ among the priority groups shown in Figure 3.10 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 we have shown, 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.

It is interesting to note that the class of rth-order delay-dependent priority systems covers the spectrum from that queueing discipline which separates priority groups in the greatest possible extent (that is, HOL) to the discipline that does not separate them at all (that is, FCFS).

Comparing the performance of the HOL system in Figure 3.3 with the performance of the time-dependent system in Figure 3.7 we observe that a certain generality seems to be lacking in the latter curves; namely all priority groups appear to saturate at the same point $\rho = 1$. This lack of generality is only an illusion. Indeed stable performance for higher

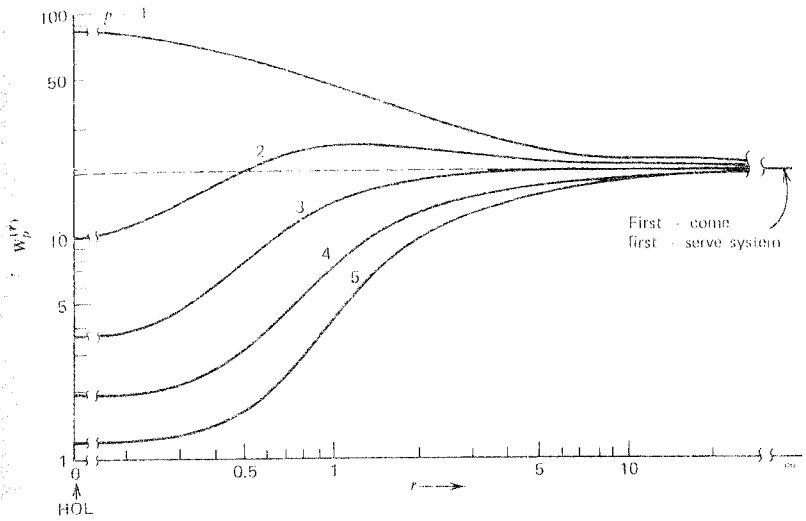

Figure 3.10 $W_p^{(r)}$ versus $r(b_p/b_{p+1} = 0.5,\ \rho = 0.95)$.

priority groups while lower groups are experiencing infinite average waits may be realized with the time-dependent discipline by permitting certain of the ratios b_p/b_{p+1} to approach zero; this effectively separates the $(p+1)$st group (and all higher) from the pth group (and all lower) in an HOL fashion [KLEI 66]. So for example in Figure 3.11 we show the performance of this more generalized time-dependent priority queueing discipline for which we have chosen the parameters $P = 25$, $\lambda_p = \lambda/25$, $\bar{x}_p = \bar{x}$ and have forced the creation of five HOL groups; within each HOL group are five priority groups that interact in a fashion similar to that shown in Figure 3.7.

Other "dynamic" priority queueing disciplines have been considered and the reader is referred to references [JACK 60, 61, 62].

3.8. OPTIMUM BRIBING FOR QUEUE POSITION

For those queueing disciplines so far studied (and for most of those described in the literature) the relative priority given to any customer is completely out of his individual control. The customer in effect has no choice as to which priority group he must join.

In this section we shift the emphasis somewhat, and allow each entering customer to "buy" his relative priority by means of a bribe [KLEI 67b]. The size of the bribe will be determined, in general, from certain economic factors inherent in the population of customers; in particular, the greater

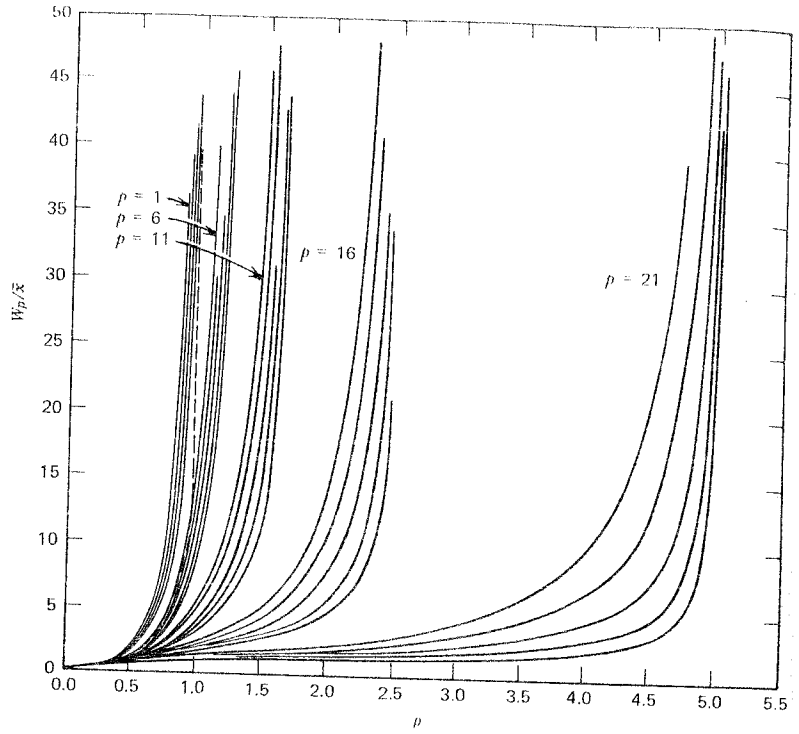

Figure 3.11 $W_p/\bar{x}$ for the mixed HOL and time-dependent system with no preemption.

the wealth of a customer and the greater his dislike of waiting on queue, the greater will be his bribe.

We consider an M/G/1 system with Poisson arrivals at a mean rate of λ customers per second, and an arbitrary PDF for service time $B(x)$ with a mean service time of $\bar{x}$ sec. Let a customer's bribe, given by Y, be a random variable with an arbitrary distribution function $\beta(y) = P[Y \leq y]$. We assume that the arrival time, the service time, and the bribe are all independent random variables for each customer and are independent of the values chosen for all other customers.

The system operates as follows: A new arrival to the system offers* a non-negative bribe Y to the "queue organizer." This customer is then

*This bribe may be thought of as given before the customer sees the length of the queue [in which case the distribution $\beta(x)$ reflects his measure of wealth and impatience] and is therefore independent of queue length.

placed in position on the queue so that all those customers whose bribes $Y' \geq Y$ are in front of him and all these with bribes $Y'' < Y$ are behind him. Newly entering customers may therefore be placed in front of, or behind this customer, depending upon their bribe. Each time the service facility completes work on some customer (who then leaves the system), it then accepts into service the customer at the front of the queue. Once in service, a customer cannot be ejected until he is completely serviced (that is, nonpreemptive*). Whenever customers give identical bribes, they are serviced in a first-come–first-serve order.

We define, for $\varepsilon > 0$, the left and right limits of $\beta(y)$ as

$$\beta(y^-) = \lim_{\varepsilon \to 0} \beta(y - \varepsilon)$$

$$\beta(y^+) = \lim_{\varepsilon \to 0} \beta(y + \varepsilon)$$

Let $W(y) \triangleq$ average waiting time (in queue) for a customer whose bribe $Y = y$. For such a customer, (say, the tagged customer), we now calculate $W(y)$ using the method as described in Section 3.2. The tagged customer must, on the average, wait a time W_0 before the customer who is in service upon his arrival is finished. In addition, he must wait until service is given to all those customers still in queue who arrived before he did and whose bribes equaled or exceeded his. The expected number of such customers whose bribes lie in the region $(u, u + du)$ is, by Little's result,

$$\lambda(u) W(u)\, du$$

where

$$\lambda(u) = \frac{\lambda\, d\beta(u)}{du}$$

Each such customer causes the tagged customer to wait an average of $\bar{x}$ sec. Furthermore, the tagged customer must wait until service is given to all those who enter the system while he is on the queue and whose bribes exceed his. The expected number whose bribes lie in the interval $(u, u + du)$ and who arrive during his average wait $W(y)$ is

$$\lambda(u) W(y)\, du$$

Each such customer also adds $\bar{x}$ sec to the tagged customer's average wait. Combining these three contributions to the tagged customer's

*A preemptive case is studied in [KLEI 67b].

average wait, we get*

$$W(y) = W_0 + \int_{y^-}^{\infty} \bar{x}\lambda(u)W(u)\,du + \int_{y^-}^{y} \bar{x}\lambda(u)W(y)\,du$$

or

$$W(y) = \frac{W_0 + \int_{y^-}^{\infty} \rho W(u)\,d\beta(u)}{1 - \int_{y^-}^{\infty} \rho\,d\beta(u)}$$

where $\rho = \lambda\bar{x}$. Since $\beta(\infty) = 1$, we have

$$W(y) = \frac{W_0 + \rho\int_{y^-}^{\infty} W(u)\,d\beta(u)}{1 - \rho + \rho\beta(y^-)} \tag{3.55}$$

In Exercise 3.15 we ask the reader to show that the solution to this integral is simply

$$W(y) = \frac{W_0}{[1 - \rho + \rho\beta(y^-)][1 - \rho + \rho\beta(y^+)]} \tag{3.56}$$

We also note that at those bribes of value y where $\beta(y)$ is continuous the solution becomes

$$W(y) = \frac{W_0}{[1 - \rho + \rho\beta(y)]^2} \tag{3.57}$$

It is interesting to note once again that the only way in which the service time distribution $B(x)$ enters the solution is through its first and second moments; this is no surprise to us for M/G/1 systems.

Note, in general, that we obtain finite average waits for all those customers offering bribes greater than y_{crit} where

$$y_{crit} = \begin{cases} 0^- & \rho < 1 \\ \beta^{-1}\left(\dfrac{\rho - 1}{\rho}\right) & \rho \geq 1 \end{cases} \tag{3.58}$$

and where $\beta^{-1}(u)$ is that value of y for which $\beta(y) = u$. This behavior is not unlike that of the HOL system for $\rho \geq 1$. We note in the limit as the bribe approaches infinity that the average waiting time approaches W_0.

Let us consider some special cases now. First in the case of a constant bribe y_0 (the same for all customers) as given by

$$\beta(y) = \begin{cases} 0 & y < y_0 \\ 1 & y \geq y_0 \end{cases}$$

* The lower limits of y^- and y^+ come about since all ties are broken on a first-come–first-serve basis. (Also, W_0 is merely $\lambda\overline{x^2}/2$ as usual.)

we have

$$W(y_0) = \frac{W_0}{1 - \rho} \tag{3.59}$$

Since all bribes are the same (resulting in no effective bribe at all), Eq. (3.59) should correspond to the well-known result for an FCFS M/G/1 queue, which it does. Second, for the case where $\beta(y)$ is continuous at the origin, giving $\beta(0) = 0$, we see that

$$W(0) = \frac{W_0}{(1 - \rho)^2} \tag{3.60}$$

This behavior at zero bribe should describe the waiting time for the lower priority group of a $P = 2$ HOL system where the arrival rate of this lower priority group is negligible compared to the total arrival rate. Indeed, as can be seen from Eq. (3.31) with $P = 2$ and $\rho_1 \ll \rho_2$, the equations above are consistent. We also observed this behavior following Eq. (3.38) for $\lim_{x\to\infty} W(x)$. Last, when only a finite (countable) set of bribes are allowed (say at the values y_p), then we have a discrete distribution which yields

$$W(y_p) = \frac{W_0}{\left[1 - \rho\sum_{i=p+1}^{P} \Delta\beta(y_i)\right]\left[1 - \rho\sum_{i=p}^{P} \Delta\beta(y_i)\right]} \tag{3.61}$$

where $\Delta\beta(y_i) \triangleq \beta(y_i^+) - \beta(y_i^-)$. This equation corresponds exactly to the result for HOL.

As soon as we introduce the notion of a bribe, we must then consider other cost factors as well. In particular, let us define a random impatience factor $\tilde{\alpha}$ (≥ 0) that measures how many dollars* it costs a customer for each second that he spends in the queue. We now introduce a cost function $C(\alpha)$ defined as

$$C(\alpha) = y_\alpha + \alpha W(y_\alpha) \tag{3.62}$$

where, again,

α = value taken on by a customer's impatience factor $\tilde{\alpha}$ (dollar/sec),
y_α = bribe offered by a customer whose impatience factor $\tilde{\alpha} = \alpha$,
$W(y_\alpha)$ = average waiting time (in queue) for a customer whose bribe is y_α.

Thus $C(\alpha)$ is the sum of the customer's bribe (in dollars) and his cost of waiting (in dollars). We assume that customers have (self-) assigned values of $\tilde{\alpha}$ before they enter the system, and that the population of customers, as a whole, produces a probability distribution $P(\alpha)$ on the

* This cost may be measured in terms of customer inconvenience or impatience, if you will.

random variable $\tilde{\alpha}$, that is, $P(\alpha)=P[\tilde{\alpha}\le\alpha]$. The queueing models here are the same as those considered earlier where now the bribe is some (deterministic) function of the random variable $\tilde{\alpha}$. We have thus shifted emphasis from the situation in which a customer offers a random bribe to a situation where the customer's bribe is functionally related to his (random) impatience factor $\tilde{\alpha}$.

We pose the following optimization problem: Find that function y_α which minimizes the expected cost C, that is,

$$\underset{y_\alpha}{\text{minimize}}\left[C\triangleq\int_0^\infty C(\alpha)\,dP(\alpha)\right] \tag{3.63}$$

subject to an average bribe constraint equal to B, that is,

$$B=\int_0^\infty y_\alpha\,dP(\alpha) \tag{3.64}$$

We must therefore choose y_α to minimize

$$C=\int_0^\infty [y_\alpha+\alpha W(y_\alpha)]\,dP(\alpha) \tag{3.65}$$

Because of the average bribe constraint, this is equivalent to minimizing

$$C-B=\int_0^\infty \alpha W(y_\alpha)\,dP(\alpha)$$

Define

$$\rho(\alpha)=\rho\frac{dP(\alpha)}{d\alpha} \tag{3.66}$$

We may interpret the quantity $\rho(\alpha)\,d\alpha$ as the fraction of time that the server is busy serving customers whose impatience factor lies in the interval $(\alpha,\alpha+d\alpha)$. Using Eq. (3.66) we then find that for $0<\rho<1$,

$$C-B=\frac{1}{\rho}\int_0^\infty \alpha\rho(\alpha)W(y_\alpha)\,d\alpha \tag{3.67}$$

Now, the continuous form of the conservation law may be written as

$$\int_0^\infty \rho(\alpha)W(y_\alpha)\,d\alpha=\frac{\rho}{1-\rho}W_0 \tag{3.68}$$

We note that minimizing Eq. (3.67) involves finding that function y_α such that the product of $\rho(\alpha)W(y_\alpha)$ and α has minimum area. However, Eq. (3.68) states that the first of these functions must itself have a *constant* area. Since $\rho(\alpha)$ is independent of y_α, we must look for conditions on $W(y_\alpha)$. Now, using the same argument as that for proving HOL optimum for linear costs (in which we matched an increasing sequence against a

decreasing sequence) we see that a necessary and achievable (see below) condition on $W(y_\alpha)$ is that it decreases with α, that is,

$$\frac{dW(y_\alpha)}{d\alpha}<0 \tag{3.69}$$

for all $\alpha\notin S$ [where the set S has the property $\int_S dP(\alpha)=0$]. Here the increasing function is α itself. Condition (3.69) may be rewritten as

$$\frac{dW(y_\alpha)}{dy_\alpha}\frac{dy_\alpha}{d\alpha}<0 \tag{3.70}$$

From Eq. (3.56) we have (letting $y_\alpha=y$)

$$\frac{dW(y)}{dy}=-\rho W_0\frac{A(y^+)[d\beta(y^-)/dy]+A(y^-)[d\beta(y^+)/dy]}{[A(y^+)A(y^-)]^2}$$

where $A(u)\triangleq 1-\rho+\rho\beta(u)$. Now since $\beta(u)\le 1$ and $\rho<1$ then $A(y^+)A(y^-)>0$. Also since $\beta(u)$ is a distribution function then $d\beta(u)/du\ge 0$. This implies that for all values of its argument $W(y_\alpha)$ has a nonpositive derivative, that is

$$\frac{dW(y_\alpha)}{dy_\alpha}\begin{cases}<0 \text{ at those } y \text{ for which } \dfrac{d\beta(y)}{dy}>0\\ =0 \text{ at those } y \text{ for which } \dfrac{d\beta(y)}{dy}=0\end{cases} \tag{3.71}$$

From Eqs. (3.70) and (3.71), then, we have that our necessary condition on y_α becomes

$$\frac{dy_\alpha}{d\alpha}>0 \qquad \text{for } \alpha\notin S \tag{3.72}$$

That this last is achievable is obvious for a large family of functions (for example, $y_\alpha=\alpha$). From this family, however, we may use only those functions satisfying Eq. (3.64); such functions clearly exist [for example, see Eq. (3.73) below].

Consider an interval $\alpha_1<\alpha<\alpha_2$ in which $P(\alpha)$ is constant. Clearly y_α can be arbitrary in any such interval without affecting C; the same is true at any *point* α for which $P(\alpha)$ is continuous. But such regions are in the set S. However, for the sets* S_1 (defined by $\alpha_1-\varepsilon\le\alpha\le\alpha_1$) and S_2 (defined by $\alpha_2\le\alpha\le\alpha_2+\varepsilon$), in which $P(\alpha)$ is assumed to be increasing, we require that Eq. (3.69) holds and also that

$$W(y_\alpha)>W(y_\alpha)$$

* Here $\varepsilon>0$.

where $a \in S_1$ and $b \in S_2$. This last is true for the same reasons leading up to Eq. (3.69), namely, that in order to minimize $C - B$, we must reduce $W(y_\alpha)$ as α increases.

To demonstrate that Eq. (3.72) is also sufficient we consider Eq. (3.65). The first term merely gives B, which is independent of y_α, and the second term depends only upon the *relative* size of the bribes and not upon the absolute bribe itself. However, Eq. (3.72) gives a complete description of the rank ordering of the bribes. Consequently the necessary and sufficient condition for y_α to be an optimum bribing function is merely that it satisfy Eqs. (3.64) and (3.72).

Thus the solution to the minimization problem set forth in Eqs. (3.63) and (3.64) restricts y_α to be a strictly increasing function of α for $\alpha \notin S$. Having constrained only the mean bribe, we get only a *condition* on y_α rather than an explicit functional form; indeed, the solution is independent of the exact form of y_α, as long as it is strictly increasing with α. Thus for the purposes of calculation and example we may choose some (simple) relation, such as the following linear one:

$$y_\alpha = K\alpha$$

Applying the mean bribe constraint, we get

$$B = K\int_0^\infty \alpha \, dP(\alpha)$$

Letting A be the average impatience factor, we get from the last two equations

$$y_\alpha = \frac{B}{A}\alpha \tag{3.73}$$

This then is an optimal bribing function.

In order to obtain some insight into the behavior of the optimum bribing procedure and the cost function, we offer the following example. Consider a system with an exponentially distributed bribe, namely,

$$B(y) = 1 - e^{-\sigma y} \qquad \sigma \geq 0,\ y \geq 0 \tag{3.74}$$

We may immediately calculate the waiting time $W(y)$ from Eq. (3.57) as

$$W(y) = \frac{W_0}{(1-\rho e^{-\sigma y})^2} \tag{3.75}$$

Using our optimum bribing rule given in Eq. (3.73) we find that the distribution of impatience factor $P(\alpha)$ that gives rise to the bribing distribution in Eq. (3.74) must be

$$P(\alpha) = 1 - e^{-(B/A)\sigma\alpha}$$

and then the average cost gives

$$C = \frac{1}{\sigma} + \frac{AW_0}{\rho}\log_e \frac{1}{1-\rho} \tag{3.76}$$

where, of course, A is the average impatience factor and $B = 1/\sigma$ is the average bribe.

The optimization described above is a global optimization and seeks bribing functions that minimize the total average cost. Recently, in [BALA 72], a similar (nonpreemptive) bribing system was studied in which conditions were found for which a customer would offer a bribe so as to minimize his own expected cost (disregarding the global minimum). Most of the considerations in [BALA 72] center around the discussion of *stable* bribing policies; roughly speaking, a bribing policy is stable if when all customers follow this policy then it does not pay for any individual customer to deviate from it. Upon their arrival and prior to making their bribe, customers are informed of the bribes given by all other customers in the system (and therefore, the queue size is also given). The first result obtained is that in the system G/G/1 for $\rho < 1$, with finite second moment of service time and with $C(\alpha)$ as given in Eq. (3.62) then a global optimal policy is one in which all customers should give zero bribe; however, this policy is clearly unstable since an infinitesimal bribe puts a customer at the head of the queue. For the M/M/1 queue the bribing policy b_k (bribe size when the queue size is k) is stable if and only if

$$(1-\rho)\max_k \Delta b_k \leq \alpha\bar{x} \leq \frac{(1+\rho)(1-\rho)}{\rho}\min_k \Delta b_k \tag{3.77}$$

where $\Delta b_k = b_k - b_{k-1}$. We are here implying that each customer has the same impatience factor $\tilde{\alpha} = \alpha$. Furthermore, it can be shown for the system G/M/1 that the bribing policy in which $b_{2k} = b_{2k+1}$ but which is strictly increasing on the even integers $(2k)$ is stable if and only if

$$\frac{1}{2}\max_k \Delta b_k \leq \frac{\alpha\bar{x}}{1-\rho} \leq \frac{1}{1+p_1+p_2}\min_k \Delta b_k \tag{3.78}$$

and p_i is the probability that at least i new customers arrive between an arrival instant and the completion of the service in progress. Considerations for the M/G/1 system are considerably more complex. In [BALA 72] the preemptive resume case is also considered. See also [BALA 73]. In [ADIR 72] locally optimal bribing policies are described for M/M/1 (both with and without preemption) as well as optimal pricing policies for the server.

3.9. SERVICE-TIME-DEPENDENT DISCIPLINES

We have seen earlier in Section 3.4 that queueing disciplines that do not discriminate in any way on the basis of service time must all have the same average waiting time. Beyond that we have seen some examples (e.g., HOL) in which priority depends in some incidental way on service time. In this section we mention some results in which more explicit use is made of a customer's service time. This discussion is rather abbreviated in this chapter but forms a point of departure for Chapter 4 when we consider models for computer time-sharing in which great effort is devoted to creating strong discrimination on the basis of required service.

One feels intuitively that giving preferential treatment to shorter jobs tends to reduce the overall average waiting time as well as the average number in a priority queueing system. In fact we have seen one example of this result in Section 3.6 where Eq. (3.42) determined the correct ordering of priorities in the optimum HOL system under linear costs; we note that if all costs are identical (that is, $C_p = C_0$ for $p = 1, 2, \ldots, P$) then this ordering is strictly on the basis of shortest average job. In fact in the continuous case [see Eq. (3.38)] we found that the SJF discipline gave the smallest possible average waiting time for nonpreemptive disciplines (i.e., apply the μC rule to this case).

There are a number of interesting disciplines based on customer service time and we list some of these below along with those we have studied (we use notation common in the applications and in the scheduling theory literature [CONW 67]):

1. FCFS: first-come-first-serve
2. LCFS: last-come-first-serve
3. SPT: shortest-processing-time-first (same as SJF)
4. SRPT: shortest-remaining-processing-time-first
5. SEPT: shortest-expected-processing-time-first
6. SERPT: shortest-expected-remaining-processing-time-first

We now state, without proof, results among these various disciplines. Since these studies arise not only from the study of queueing systems but also from the study of scheduling systems, we also consider the case where no arrivals are permitted to enter (the case for scheduling) but rather all jobs that require processing are available at the start of the "busy period." (We choose to refer to service now as "processing time" as is common in the literature.) For the case of no arrivals, there is an additional discipline that we must consider which has recently been introduced by Sevcik [SEVC 74], namely,

7. SIPT: shortest-imminent-processing-time-first

The SIPT discipline operates as follows: if the ith customer has a distribution of processing time $B_i(x)$ and has already received x sec of service, and if the permitted points in time when this ith customer may be preempted are $t_{i1}, t_{i2}, \ldots$, then the priority $q_i(x)$ of this job is calculated under this discipline as follows:

$$\frac{1}{q_i(x)} = \min_{(t_{ij} > x)} \left[\frac{\int_x^{t_{ij}} [1 - B_i(y)]\, dy}{B_i(t_{ij}) - B_i(x)} \right]$$

The customer whose function $q_i(x)$ is maximized is defined to have the largest priority; the above ratio is the expected time spent on customer i if he is allowed at most an amount of service equal to $t_{ij} - x$.

For some of these seven disciplines, we are interested in defining related disciplines in which *cost* enters the picture. Often, we are concerned with linear cost functions (that is, costs which are linear with average waiting time). The new class of disciplines we wish to introduce involves forming one of the measures mentioned above (as for example, SEPT) and dividing this measure for each job by the cost rate associated with that job (thus, for example, forming the new discipline SEPT/C). We will need this additional definition for the case of SPT/C, SEPT/C (note that this is the μC rule), SRPT/C, and SIPT/C.

Let us now comment on some of the known results. In all cases, we assume a conservative system (no creation or destruction of work—specifically, no cost for preemption and no idle server when jobs are present).

First, we consider the case of no arrivals and costs that are linear with average waiting time. In the case of exactly known service times, it is known (see, for example, p. 26ff [CONW 67] for a proof and discussion of this result) that SPT/C is optimum (i.e., it minimizes the average cost). If only the distribution of service time for each job is known, then, for the nonpreemptive case, SEPT/C minimizes the average cost [SMIT 56]; in the preemptive case (where the set of permissible preemption points may be specified), then SIPT/C scheduling is optimum [SEVC 74].

Let us now consider the more interesting case of arrivals (i.e., queueing systems). We seek the optimum scheduling rule (but restrict ourselves to the case of priority disciplines only—that is, rules that evaluate a job's priority based only on that job's parameters) for given types of cost functions [linear, convex, concave, minimum variance of time in system (σ_s^2), maximum σ_s^2]. In a recent paper by Schrage [SCHR 74], a very nice summary of several of these optimal scheduling disciplines is given. He considers three possible information states, namely, (1) exact service time information given; (2) distribution of service time only given, or (3) no

Table 3.1.
Optimal Service-Time-Dependent Scheduling Algorithms

	(1) PROCESSING TIMES KNOWN	(2) ONLY DISTRIBUTION OF PROCESSING TIMES KNOWN	(3) NO INFORMATION REGARDING PROCESSING TIMES
Nonpreemptive	SPT/C for M/G/1 linear costs [PHIP 56, ACZE 60, FIFE 65, HARR 72][a]	SEPT/C for M/G/1 linear costs [PHIP 56, ACZE 60, FIFE 65, HARR 72][a]	FCFS for G/G/1 convex costs [HAJI 71]
	SPT/C for G/G/1 linear costs two classes only [WOLF 70, SCHR 74]	SEPT/C for G/G/1 linear costs two classes only [WOLF 70, SCHR 74]	FCFS for G/G/1 minimum σ_s^2 [KING 62]
			LCFS for G/G/1 maximum σ_s^2 [TAMB 68]
Preemptive	SRPT for G/G/1 identical linear costs ($C_p = C_0$) [SCHR 68]	SEPT/C for G/M/1 linear costs [SCHR 74]	FCFS for G/IFR/1 convex costs [JACK 61, SCHR 74][b]
	SRPT/C for M/G/1 linear costs ([AVI 64, ETSC 66, JAIS 68] for two classes) [SCHR 74][b]	SIPT/C for M/G/1 linear costs [SEVC 74][b]	LCFS for G/M/1 concave costs [SCHR 74][b]
			FB for G/DFR/1[c] linear costs [KALR 71]—for M/DFR/1 [SCHR 74][b]

[a] Also see heuristic proof that the μC rule is optimal in Section 3.6.
[b] Conjecture only (with heuristic proof in some cases).
[c] See Chapter 4 below for a definition of FB.

information regarding servivice times given. Observe that case (2) is the most general; case (1) is clearly a special (degenerate distribution) example of (2).* Also, case (3) is the situation with only one class of customer. The results and appropriate references to these results are given in the table on p. 146; the reader is urged to see these references for further details and restrictions on these optimality results.

Another comparison among some of these queueing disciplines is given in [SUZU 70]. Let $D_1 \rightarrow D_2$ denote the fact that the average wait using discipline D_1 is greater than or equal to that for D_2. Then, the following relationships hold:

where LPT(LRPT) means longest-(remaining)-processing-time-first, and RS means random order of service.

There are cases in which exact processing times are unavailable but more than just the distribution of service times is known. For example it may be possible to separate customers' required processing times into "large" and "small." In particular, many examples indicate [CONW 67] that this separation into two groups provides a considerable reduction in mean waiting times as opposed to the FCFS system (see Exercises 3.8 and 3.9).

As mentioned earlier, other processing-time-dependent queueing disciplines will be considered next in Chapter 4.

REFERENCES

ACZE 60 Aczel, M. A., "The Effect of Introducing Priorities," *Operations Research,* **8,** 730–733 (1960).

ADIR 72 Adiri, I., and V. Yechiali, "Optimal Pricing and Priority Purchasing Policies," IBM Research Report RC-3581, September 2, 1972.

AVI 63 Avi-Itzhak, B., and P. Naor, "Some Queueing Problems with the Service Station Subject to Breakdown," *Operations Research,* **11,** 303–320 (1963).

AVI 64 Avi-Itzhak, B., I. Brosh, and P. Naor, "On Discretionary Priority Queueing," *Zeitschrift für angewandte Mathematik und Mechanik,* **6,** 235–242 (1964).

BALA 72 Balachandran, K. R., "Purchasing Priorities in Queues," *Management Science,* **18,** No. 5, 319–326 (1972).

* Note that SEPT/C becomes SPT/C and SERPT/C becomes SRPT/C for known service times.

BALA 73 Balachandran, K. R., and J. C. Lukens, "Stable Pricing Policies in Service Systems," Report MS 1, College of Industrial Management, Georgia Institute of Technology, October 1973.

BRUM 69 Brumelle, S. L., "Some Inequalities for Multi-Server Queues," ORC 69-17, Operations Research Center, University of California, Berkeley, 1969.

COBH 54 Cobham, A., "Priority Assignment in Waiting Line Problems," *Operations Research*, **2,** 70–76 (1954).

COLE 71 Cole, G. C., *Computer Network Measurements: Techniques and Experiments*, School of Engineering and Applied Science, University of California, Los Angeles, UCLA-ENG-7165, 1971.

CONW 67 Conway, R. W., W. L. Maxwell, and L. W. Miller, *Theory of Scheduling*, Addison-Wesley (Reading, Mass.), 1967.

COX 61 Cox, D. R., and W. L. Smith, *Queues*, Methuen (London) and Wiley (New York), 1961.

CRAB 73 Crabill, T. B., D. Gross, and M. J. Magazine, "A Survey of Research on Optimal Design and Control of Queues," Serial T-280, School of Engineering and Applied Science, The George Washington University, June 1, 1973.

ETSC 66 Etschmaier, "Discretionary Priority Processes," M.S. Thesis, Case Institute of Technology, Cleveland, Ohio, 1966.

FIFE 65 Fife, D. W., "Scheduling With Random Arrivals and Linear Loss Functions," *Management Science*, **11,** No. 3, 429–437 (1965).

GAVE 62 Gaver, D. P., Jr., "A Waiting Line with Interrupted Service, Including Priorities," *Journal of the Royal Statistical Society, Series B*, **24,** 73–90 (1962).

HAJI 71 Haji, R., and G. F. Newell, "Optimal Strategies for Priority Queues with Nonlinear Costs of Delay," *SIAM Journal of Applied Mathematics*, **20,** 224–240 (1971).

HARR 72 Harrison, J. M., "Dynamic Scheduling of a Multi-Class Queue, I: Problem Formulation and Descriptive Results," Technical Report #36, and "Dynamic Scheduling of a Multi-Class Queue, II: Discount Optimal Dynamic Policies," Technical Report #37, Department of Operations Research, Stanford University, Stanford, California, June 1972.

JACK 60 Jackson, J. R., "Some Problems in Queueing with Dynamic Priorities," *Naval Research Logistics Quarterly*, **7,** 235–249 (1960).

JACK 61 Jackson, J. R., "Queues with Dynamic Priority Discipline," *Management Science*, **8,** No. 1, 18–34 (1961).

JACK 62 Jackson, J. R., "Waiting-Time Distributions for Queues with Dynamic Priorities," *Naval Research Logistics Quarterly*, **9,** 31–36 (1962).

JAIS 68 Jaiswal, N. K., *Priority Queues*, Academic Press (New York), 1968.

KALR 71 Kalro, A. L., "Optimal Processor Scheduling in a Computer Time Shared System," ORC 71-25, University of California, Berkeley, California, September 1971.

KEIL 62 Keilson, J., "Queues Subject to Service Interruption," *Annals of Mathematical Statistics*, **33,** 1314–1322 (1962).

KEST 57 Kesten, H. and J. Th. Runnenberg, "Priority in Waiting-Line Problems I and II," *Nederlandse Akademie van Wetenschappen, Amsterdam, Proceedings, Series A*, **60,** 312–324, 325–336, (1957).

KING 62 Kingman, J. F. C., "The Effect of the Queue Discipline on Waiting Time Variance," *Proceedings of the Cambridge Philosophical Society*, **58,** 163–164 (1962).

KLEI 64a Kleinrock, L., *Communication Nets: Stochastic Message Flow and Delay*, McGraw-Hill (New York), 1964, reprinted by Dover Publications, Inc., (New York), 1972.

KLEI 64b Kleinrock, L., "A Delay Dependent Queue Discipline," *Naval Research Logistics Quarterly*, **11,** 329–341 (1964).

KLEI 65 Kleinrock, L., "A Conservation Law for a Wide Class of Queueing Disciplines," *Naval Research Logistics Quarterly*, **12,** 181–192 (1965).

KLEI 66 Kleinrock, L., "Queueing with Strict and Lag Priority Mixtures," *Proceedings of the 4th International Conference on Operational Research*, Boston, Mass. K-I-46 to K-I-67, 1966.

KLEI 67a Kleinrock, L., and R. P. Finkelstein, "Time Dependent Priority Queues," *Operations Research*, **15,** 104–116 (1967).

KLEI 67b Kleinrock, L., "Optimum Bribing for Queue Position," *Operations Research*, **15,** 304–318 (1967).

KLEI 75 Kleinrock, L., *Queueing Systems, Vol. I: Theory*, Wiley Interscience, (New York), 1975.

LIPP 75 Lippman, S. A., "On Dynamic Programming with Unbounded Rewards." *Management Science*, 21, No. 11, 1225–1233 (1975).

PHIP 56 Phipps, T. E., Jr., "Machine Repair as a Priority Waiting-Line Problem," *Operations Research*, **4,** 76–85 (1956).

PRAB 73 Prabhu, N. U., and S. Stidham, Jr., "Optimal Control of Queueing Systems," Technical Report No. 186. Dept. of Operations Research, Cornell University, June 1973.

REED 74 Reed, F. C., "Difference Equations and the Optimal Control of Single Server Queueing Systems," Technical Report No. 23, Stanford University, Dept. of Operations Research, March 22, 1974.

SCHR 68 Schrage, L., "A Proof of the Optimality of the Shortest Remaining Processing Time Discipline," *Operations Research*, **16,** 687–690 (1968).

SCHR 70 Schrage, L., "An Alternative Proof of a Conservation Law for the Queue G/G/1," *Operations Research*, **18,** 185–187 (1970).

SCHR 74 Schrage, L., "Optimal Scheduling Disciplines for a Single Machine Under Various Degrees of Information," Working Paper, Graduate School of Business, University of Chicago, 1974.

SEVC 74 Sevcik, K., "A Proof of the Optimality of 'Smallest Rank' Scheduling," *Journal of the Association for Computing Machinery*, **21,** 66–75 (1974).

SMIT 56 Smith, W. E., "Various Optimizers for Single-Stage Production," *Naval Research Logistics Quarterly*, **3**, 59–66 (1956).

SUZU 70 Suzuki, T., and K. Hayashi, "On Queue Disciplines," *Journal of the Operations Research Society of Japan*, **13**, 43–58 (1970).

TAMB 68 Tambouratzis, D. G., "On the Property of the Variance of the Waiting Time of a Queue," *Journal of Applied Probability*, **5**, 702–703 (1968).

TORB 73 Torbett, E. A., "Models for the Optimal Control of Markovian Closed Queueing Systems with Adjustable Service Rates," Technical Report No. 20, Dept. of Operations Research, Stanford University, January 15, 1973.

WOLF 70 Wolf, R. W., "Work Conserving Priorities," *Journal of Applied Probability*, **7**, 327–337 (1970).

EXERCISES

3.1. Using the notion of residual life, show that $W_0 = \lambda \overline{x^2}/2$.

3.2. Consider an M/G/1 system with 2 priority groups and some unspecified queueing discipline which is work-conserving. We are given that

$$W_2 = \frac{W_0}{1 - \alpha\rho_1 - \beta\rho_2}$$

where $\rho_p = \lambda_p \bar{x}_p$ is the utilization factor for the pth group ($p = 1, 2$) and $o < \alpha < 1$, $0 < \beta < 1$. Find W_1 in terms of ρ_1, ρ_2, α, β, and W_0.

3.3. Find the mean and variance for Y_b and Y_c in Figure 3.1.

3.4. For M/G/1 we wish to compare FCFS with LCFS.

(a) Show that $W_{\text{FCFS}} = W_{\text{LCFS}}$ $(= W)$ by using the moment-generating properties of $W^*(s)$.

(b) Similarly show that $\sigma^2_{\text{FCFS}} = (1-\rho)\sigma^2_{\text{LCFS}} - \rho W^2$ ■

3.5. Consider an M/M/1 nonpreemptive HOL system. Let j be the smallest integer such that $\sum_{i=j}^{P} \rho_i < 1$. Solve for W_p $(p = j, j+1, \cdots, P)$. Note that $W_p = \infty$ for $p < j$.

3.6. For the system described in the previous problem, establish a conservation for the sum

$$\sum_{p=j}^{P} \rho_p W_p$$

3.7. Calculate W_p for the nonpreemptive HOL system from Eq. (3.32).

3.8. Consider a nonpreemptive HOL system with $P = 2$ constructed as follows. We assume that service times for all customers are drawn from $B(x)$, *but are known when a customer arrives*. Let x_0 be a number defining the boundary between the two groups, that is, if $x < x_0$, then a job falls in group $p = 2$ and if $x \geq x_0$ it is placed in group $p = 1$.

(a) Show that

$$W = \sum_{p=1}^{2} \frac{\lambda_p}{\lambda} W_p = \frac{W_0}{1-\rho}\left[\frac{1-\rho B(x_0)}{1-\rho_2}\right]$$

(b) Prove that this simple discrimination is an improvement for W over FCFS.

3.9. Consider an M/G/1 system with a $P = 2$ nonpreemptive HOL priority discipline. Let

$$W = \frac{\lambda_1}{\lambda} W_1 + \frac{\lambda_2}{\lambda} W_2$$

Assume $\rho < 1$ [COLE 71].

(a) Prove that

$$W = W_{\text{FCFS}} \frac{1 - \lambda_2 \bar{x}}{1 - \lambda_2 \bar{x}_2}$$

where

$$\bar{x} = \frac{\lambda_1}{\lambda} \bar{x}_1 + \frac{\lambda_2}{\lambda} \bar{x}_2$$

and W_{FCFS} is the mean wait in an FCFS system.

(b) Suppose now that group 2 consists of all jobs with service time $\leq \tau$, and group 1 has $\bar{x} > \tau$, where $\bar{x}$ has the general distribution $B(x)$. Let τ_0 be the optimum value of this threshold such that W is minimized, and let $W_{\min}$ be this minimum average wait.

(i) Show that

$$\frac{W_{\min}}{W_{\text{FCFS}}} = \frac{\bar{x}}{\tau_0}$$

(ii) Show that τ_0 is defined through

$$\frac{1}{\rho} = 1 + \frac{\frac{\lambda_1}{\lambda}(\bar{x}_1 - \tau_0)}{\tau_0 - \bar{x}}$$

3.10. Consider a two priority M/G/1 system for which $W_0 = 2$, $\rho_1 = \rho_2 = \frac{1}{4}$.

(a) Suppose $W_1 = 5$. Find W_2.

(b) If the system is HOL (nonpreemptive), find W_1 and W_2.

(c) If the system is FCFS, find W_1 and W_2.

(d) If the system is LCFS, find W_1 and W_2.

3.11. Consider a $P = 2$ nonpreemptive priority queueing system with $\lambda_1 = \lambda_2 = 1$ and $\bar{x}_1 = \frac{1}{2}$ and $\bar{x}_2 = \frac{1}{4}$.

(a) Design a system which achieves a performance ratio $W_2/W_1 = \alpha < 1$.

(b) Suppose a customer enters at some random time and must wait for service until the system empties. Give an expression for the ratio of this customer's average waiting time to his average wait in an FCFS system with the same input.

3.12. Consider a delay-dependent discipline for which $0 \geq b_1 \geq b_2 \geq \cdots \geq b_P$. Find the set of simultaneous equations similar to Eq. (3.48) that define W_p $(p = 1, 2, \ldots, P)$.

3.13. Consider a $P = 2$, rth-order time-dependent priority discipline with the following cost function (for some constant, m):

$$C = \sum_{p=1}^{2} C_p \lambda_p [W_p^{(o)}]^{m+1}$$

(a) For a given pair (b_1, b_2), express $W_p^{(o)}$ in terms of W_0, ρ_p, b_p, and r $(p = 1, 2)$.

(b) For $0 \leq (b_1/b_2)^{1/r} < 1$, find the optimum value for b_1/b_2 so as to minimize C.

3.14. Consider the rth-order time-dependent priority discipline. We wish to prove that Eq. (3.50) follows from Eq. (3.49).

(a) Let $r' = 1$ with no loss of generality. Consider any three customers whose priority functions (may) intersect each other. Show that the intersection times for the general r case will be exactly the same for the $r' = 1$ case if Eq. (3.49) holds.

(b) Use induction to establish the equivalence of all intersection times for the general case of M customers to establish Eq. (3.50) (i.e., equivalence of intersection times implies equivalence of order of service).

(c) Now, in addition prove Eq. (3.51).

3.15. **(a)** Show that Eq. (3.50) is indeed the solution (for the average wait conditioned on a bribe of value x) to Eq. (3.55).

(b) Show that the average cost C given in Eq. (3.76) is correct for M/M/1.

(c) For bribes uniformly distributed over the interval $[0, M]$, find the average cost C.

3.16. Consider a nonpreemptive M/M/m system in which the average service time for the ith server is $1/\mu_i$ where $1/\mu_1 < 1/\mu_2 < \cdots < 1/\mu_m$. That is, the ith server is faster than the $(i+1)$st on the average. Customers join the queue in order of arrival. When the ith server becomes free, he offers his services to the first queued customer; if this customer refuses service from him (i.e., the customer is holding out for a faster server), he then offers his services to the second queued customer, and so on. If no queued customer accepts, he remains idle until possibly some newly arriving customer accepts him. Each customer uses a (local) strategy that minimizes his average time in system. No service times are known to any customers (not even their own service time). Let k_i be that position in the queue at which a customer should first accept the ith server [KLEI 64a].

Show that the critical positions k_i must satisfy

$$k_i < \frac{S_{i-1}}{\mu_i} \leq k_i + 1$$

where $S_i = \mu_1 + \mu_2 + \cdots + \mu_i$ and $k_1 = 1$.

3.17. Consider an M/G/1 system with N queues that are labeled $Q1$, $Q2, \ldots, QN$. Arriving customers join the tail of $Q1$ and after receiving service [from distribution $B(x)$], they join the tail of $Q2$, and so on, finally they join QN, receive service, and then depart at last. Each customer therefore receives N independent services from $B(x)$. We also assume a priority ordering among the queues, which we denote by $P = \{q_1, q_2, \cdots, q_N\}$, which implies that Qi has priority over Qj if $q_i < q_j$; for example, $P = \{1, 3, 2\}$ implies that $Q1$ has highest priority, $Q3$ is next, and $Q2$ is last. This is the order in which a queue is selected for service whenever a service completion occurs. Within a given queue, service is FCFS.

Let $p_k = P[k$ "services" in system at departure instants], where each newly arriving customer counts for N "services." Let $Q(z) = \sum_{k=0}^{\infty} p_k z^k$. Let $\rho = \lambda \bar{x} N$

(a) Find $Q(z)$.

Now let $N = 2$ and $P = \{2, 1\}$. Let

$$R(z_1, z_2) = \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} p(k_1, k_2) z_1^{k_1} z_2^{k_2}$$

where $p(k_1, k_2) = P[k_1$ customers in $Q1$ and k_2 customers in $Q2]$ (this too is calculated at departure instants).

(b) For the given priority ordering, what possible values can k_2 take?

(c) Show that the answer to (b) allows us to write

$$p(k_1, k_2) = p_{2k_1+k_2}$$

(d) Show that

$$R(z_1, z_2) = \frac{(1-\rho)B^*(\lambda - \lambda z_1)[z_2 - z_1 + (1-z_2)B^*(\lambda - \lambda z_1)]}{[B^*(\lambda - \lambda z_1)]^2 - z_1}$$

[*Hint:* Consider the series obtained for the sum $Q(z) + Q(-z)$ and for the difference $Q(z) - Q(-z)$].

(e) From (d), find $E[k_2]$.

Let us now consider $P = \{1, 2\}$.

(f) Find T, the average time a customer spends in the system.

3.18. As a variation on the cost function given in Eq. (3.65), consider an FCFS M/M/1 system for which $C = \mu C_s + \lambda T \alpha$ where C_s is the cost per unit of service rate for a server. C is thus the average cost of running the facility. Find the optimum value of μ that minimizes C.

3.19. Show that W_{SPT}, the mean wait for a shortest-processing-time-first (same as SJF) discipline is related to T_{SRPT}, the mean time in system for a shortest-remaining-processing-time-first discipline for M/M/1 as follows:

$$W_{SPT} = \rho T_{SRPT}$$

For an M/M/1 system with $\lambda = 0.9$ and $\mu = 1.0$, calculate and compare W_{SPT}, W_{SRPT}, and W_{FCFS}.

3.20. Consider an M/G/1 system operating under the shortest-remaining-processing-time-first (SRPT) discipline in which all customer service times are known ahead of time. A new customer will preempt a customer in service only if his service time is less than that which remains for the customer in service. At a service completion, the shortest job is served next.

Show that the mean time in system T_{SRPT} is

$$T_{SRPT} = \int_{y=0}^{\infty} \int_{x=0}^{y} \frac{dx}{1 - \lambda f_1(x)} dB(y) + \frac{\lambda}{2} \int_{y=0}^{\infty} \frac{f_2(y) + y^2[1 - B(y)]}{[1 - \lambda f_1(y^-)][1 - \lambda f_1(y^+)]} dB(y)$$

where

$$f_i(y) = \int_0^y u^i dB(u)$$

3.21. Consider a sequence of arrival times {1, 3, 4, 5, 7, 8, 13, 15, 17, 23, 27} and a corresponding sequence of service times {5, 6, 2, 3, 2, 1, 2, 2, 2, 3, 1}. Assume the system is empty prior to the arrival at time 1.

(a) Calculate the average wait for these eleven customers when the discipline is FCFS.

(b) Repeat for LCFS.

(c) Repeat for SPT (SJF).

(d) Repeat for LPT.

(e) Repeat for SRPT.

(f) Repeat for LRPT.

(g) Confirm the $D_1 \rightarrow D_2$ relationship given in the text for these six cases.