Proof: Let the queueing system be empty at $t = 0$ and let

$$P_n(t \mid k) = P\{\text{at time } t \text{ there are exactly } n \text{ users being served} \mid \text{exactly } k \text{ arrivals in } [0, t]\}$$

Then, since arrivals occur according to a Poisson process,

$$P_n(t) = \sum_{k=0}^{\infty} P_n(t \mid k) \frac{(\lambda t)^k e^{-\lambda t}}{k!} \tag{4.89}$$

But, as we have seen in Chapter 2, if there are exactly k arrivals from a Poisson process in $[0, t]$, the unordered arrival times are uniformly, independently distributed over $[0, t]$. So, we now consider a random user arriving in the interval $[0, t]$ (see Figure 4.13). If the random user arrives with x more time left in $[0, t]$, the probability that he is still being serviced at t is $[1 - F_S(x)]$. But the unconditional probability he arrives in $[t - x, t - x + dx]$ is dx/t. Thus, the probability that a random user who arrives any time in $[0, t]$ is still being serviced at time t is

$$\alpha = \int_0^t \frac{1 - F_S(x)}{t}\, dx \tag{4.90}$$

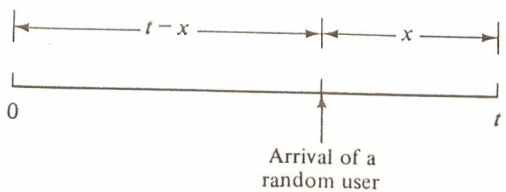


FIGURE 4.13 User arrival in $[0, t]$.

We now use the fact that the unordered arrival times are independent. Given k arrivals in $[0, t]$, the probability that there are n customers being serviced at t $(n \leq k)$ is

$$P_n(t \mid k) = \binom{k}{n} \alpha^n (1 - \alpha)^{k-n} \tag{4.91}$$

Substituting (4.90) and (4.91) into (4.89), we obtain (4.88)—after some algebraic manipulation. The steady-state result (4.87) is derived just by letting $t \longrightarrow \infty$ in (4.88).

As we have already noted earlier, in order to apply a $M/G/\infty$ (or $M/M/\infty$) model, it is not really necessary to have an infinite number of servers. In practice, all that is needed is that the actual number of servers be sufficiently large or that workload be sufficiently small so that queues almost never form. Fire department operations, for example, possess this characteristic. It is highly unlikely that in a major city there will ever be insufficient fire engines or fire ladders to dispatch to a major fire; for if the fire stations in a given area run out of fire companies as a result of one or more multiple alarm fires, then other fire companies are dispatched to the scene, first from neighboring stations and eventually, if necessary, from nearby cities or communities. Assuming then that fire alarms in a region are generated according to a Poisson process, service times for different fire alarms are statistically independent and identically distributed, fire alarm rates and service rates are constant in time, and an infinite number of fire-fighting units are available, one can use the $M/G/\infty$ model to estimate the distribution of the number of fire-fighting units which are busy at any one time in an area. Although the validity of some of the assumptions above may be questioned, the results provided by such a model[9] [CHAI 71] have been shown to provide excellent approximations to the true distribution [IGNA 78].

4.8.3 G/G/1 System

We consider next a system consisting of a single server with independent and identically distributed user interarrival times, independent and identically distributed service times and unlimited queueing capacity. This is the $G/G/1$ system.

Let us now call X the random variable that represents interarrival time and use $f_X(x)$, $1/\lambda$, and σ_X^2 for the pdf, the expected value and the variance of X, respectively. Similarly, as we have already done, we shall use S, $f_S(s)$, $1/\mu$, and σ_S^2 as the corresponding symbols for the service times.

Practically no easily usable exact results exist for this model because of the analytical difficulties that we outlined earlier. However, some useful bounds have been developed in recent years for the quantities $\bar{L}$, $\bar{L}_q$, $\bar{W}$, and $\bar{W}_q$.

For general $G/G/1$ systems (no restrictions on the interarrival or on the service time pdf's) a most useful upper bound [MARS 68] and a (much less useful) lower bound [MARC 78] have been obtained. These bounds state that for the average steady-state waiting time in queue, $\bar{W}_q$, we have

$$\frac{\rho^2(1 + C_S^2) - 2\rho}{2\lambda(1 - \rho)} \leq \bar{W}_q \leq \frac{\lambda(\sigma_X^2 + \sigma_S^2)}{2(1 - \rho)} \tag{4.92}$$

where C_s $(= \sigma_s \mu)$ is the coefficient of variation for the service times [as used in (4.79a)] and, as usual, $\rho = \lambda/\mu$. The condition for the existence of steady state is $\rho < 1$.

[9] The referenced model, due to J. Chaiken, is a modified $M/G/\infty$ model which accounts for the fact that two or more fire units are often simultaneously dispatched to a fire alarm.

That the lower bound in (4.92) is not particularly tight becomes obvious from the fact that, even at very high utilization rates ($\rho \doteq 1$), this bound is trivial (i.e., takes negative values) unless $C_s > 1$. But for C_s to be greater than 1, it must be that the service time pdf must be "more random" than the negative exponential pdf (which has $C_s = 1$). This would be rather unusual in practical urban service system problems.

Fortunately, a simple and quite tight lower bound has been obtained [MARS 68] for a particular subclass of $G/G/1$ queueing systems. This subclass, happily, encompasses most of the cases that one is likely to encounter in practice. The subclass of $G/G/1$ queueing systems that we refer to consists of those systems for which the interarrival time pdf $f_X(x)$ has the following property: For *all* values of a constant t_0 (≥ 0), it is true that

$$E[X - t_0 \,|\, X > t_0] \leq \frac{1}{\lambda} \tag{4.93}$$

Although (4.93) may appear complicated at first sight, it really imposes a very simple condition: if it is known that any given interarrival gap lasted more than a time t_0, then (4.93) requires that the expected length of the *remaining* time, $X - t_0$, in that gap be less than the unconditional expected length of the gap,[10] $E[X]$ ($= 1/\lambda$). One of the basic properties of the negative exponential random variable, as we know, is that (4.93) is an equality for this random variable.

When condition (4.93) is satisfied, the following is true:

$$B - \frac{1+\rho}{2\lambda} \leq \bar{W}_q \leq B \tag{4.94}$$

where B is the upper bound listed in (4.92), that is,

$$B = \frac{\lambda(\sigma_X^2 + \sigma_S^2)}{2(1-\rho)}.$$

As usual, the relationships $\bar{L}_q = \lambda \bar{W}_q$, $\bar{W} = \bar{W}_q + 1/\mu$, and $\bar{L} = \lambda \bar{W}$ hold, and thus upper and lower bounds can be obtained for the quantities $\bar{L}$, $\bar{W}$, and $\bar{L}_q$ as well from the bounds above. To appreciate how good the bounds in (4.94) are, we can look at the form that (4.94) takes for the average queue length, $\bar{L}_q$. We have

$$\lambda B - \frac{1+\rho}{2} \leq \bar{L}_q \leq \lambda B$$

which means that the difference between the upper and the lower bounds is $(1 + \rho)/2$, and since $0 < \rho < 1$, this difference is always between $\frac{1}{2}$ and 1.

[10] Probability distributions that have this property are sometimes referred to as "decreasing mean residual life" (DMRL) distributions, for obvious reasons.

Thus, we are able to determine the average queue length to within an accuracy of between 0.5 and 1 user (depending on the value of ρ), which is truly excellent for any practical application. In fact, the two bounds, remarkably, get closer to each other, on a *percentage* basis, as $\rho \to 1$, because $\bar{L}_q$ then becomes large, and a difference of 1 between the two bounds is thus insignificant.

It has already been stated that the case which satisfies condition (4.93) and, consequently, for which tight upper and lower bounds can be used is the most common in practice. The reason is that most "well-behaved" arrival-time distributions satisfy (4.93). Such, for instance, would be the case for uniform or triangular or beta-type pdf's, which often are reasonably good approximations of many general interarrival time pdf's. Only a few common continuous random variables, notably those in the hyperexponential family, which are "more random"—informally speaking—than the negative exponential random variable (see Problem 4.16), do not, in fact, satisfy (4.93). In practice, it is usually simple to recognize those distributions which do not satisfy (4.93). This is especially true of random variables taking on discrete values only.

Consider, for instance, the random variable Y with the probability mass function shown in Figure 4.14. For the value $t_0 = 1$, we have

$$E[Y - t_0 \,|\, Y > t_0] = E[Y - 1 \,|\, Y > 1] = P\{Y = 10\} \cdot (10 - 1) = 1 \cdot 9 = 9$$

On the other hand, $E[Y] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 10 = 5.5$. Therefore, (4.93) is *not* satisfied by random variable Y.

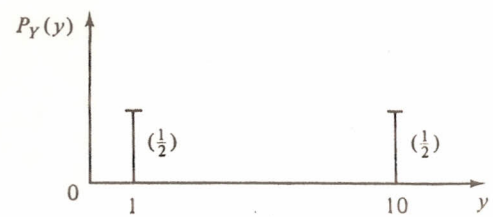


FIGURE 4.14 Random variable that does not satisfy condition (4.93).

Heavy-traffic approximation. Another important practical result which is available for the $G/G/1$ system is known as the *heavy-traffic approximation* [KING 62]. As the name suggests, this is a result that applies for values of ρ which are close to 1 and, consequently, it provides estimates for waiting times when it is known that waiting times are large.

The heavy-traffic approximation states that when $\rho \to 1$, the distribution of steady-state waiting time in queue in a $G/G/1$ system is approximately

negative exponential with mean value

$$\bar{W}_q = \frac{\lambda(\sigma_X^2 + \sigma_S^2)}{2(1-\rho)} \tag{4.95}$$

where λ, ρ, σ_X^2, and σ_S^2 are as defined above.

This is a remarkable result, since it provides not just an estimate of a moment of waiting times but of the actual *distribution* itself under very general conditions.

Note that the above expression for $\bar{W}_q$ is identical to the expression for the upper bound on $\bar{W}_q$ for $G/G/1$ systems in (4.92). Thus, the upper bound of (4.92) improves as an approximate estimate of $\bar{W}_q$ as $\rho \longrightarrow 1$. From (4.95) and (4.94) we can also reach the following important practical conclusion:

> The average waiting time for $G/G/1$ queueing systems becomes dominated by a $(1-\rho)^{-1}$ term under steady-state conditions, as the utilization ratio approaches 1. Thus, the type of behavior first indicated by the simple $M/M/1$ queueing system is also present for entirely general arrival- and service-time distributions.

4.8.4 $G/G/m$ Queueing Systems

About the only general results that have been obtained to date for the $G/G/m$ case are in the form of quite loose upper and lower bounds on average steady-state queueing characteristics [BRUM 71]. These bounds are often computed by, first, comparing a $G/G/m$ system with a $G/G/1$ system that has the same "service potential" as the $G/G/m$ system (i.e., the single server in $G/G/1$ works m times as fast as each of the servers in $G/G/m$) and, then, by using the earlier results on $G/G/1$.

The best generally applicable bounds on the average waiting time in queue which have been published to date for $G/G/m$ systems give the inequalities

$$\bar{W}_q^1 - \frac{(m-1)\mu E[S^2]}{2m} \leq \bar{W}_q \leq \frac{[\sigma_X^2 + (1/m)\sigma_S^2 + ((m-1)/m^2)(1/\mu^2)]\lambda}{2(1-\lambda/m\mu)} \tag{4.96}$$

where μ, σ_S^2, and $E[S^2]$ are the service rate, variance of service time, and second moment of service time, respectively, for *each* of the m servers. $\bar{W}_q^1$ is the average waiting time for a $G/G/1$ system with a service time described by a random variable $S^* = S/m$ (i.e., with service m times faster than that

of each of the m servers in the $G/G/m$ system) and with an arrival process identical to that for the $G/G/m$ system.

For $\bar{W}_q^1$ one should obviously use either an exact expression, if one is available, or, as is more likely, a lower bound on W_q^1 by using (4.92) or, if applicable, (4.94). For example, for the $M/G/m$ queueing system, one should use the exact expression (4.81) for W_q^1 with $1/m\mu$ and σ_S^2/m^2, for the expected value and variance of the service times, respectively.

Finally, a result analogous to the heavy-traffic approximation for $G/G/1$ systems has also been derived recently [KOLL 74] for $G/G/m$ systems. This result states:

> For $\lambda/m\mu \longrightarrow 1$ in a $G/G/m$ system, the waiting time in queue under steady-state conditions assumes a distribution that is approximately negative exponential with mean value
>
> $$W_q \simeq \frac{[\sigma_X^2 + (\sigma_S^2/m)]\lambda}{2(1-\lambda/m\mu)} \tag{4.97}$$

Note once more that expected waiting time is dominated by a $(1-\rho)^{-1}$ term, as ρ approaches 1 ($\rho = \lambda/m\mu$ for multiserver systems).

4.9 QUEUEING SYSTEMS WITH PRIORITIES

In all the cases that we have discussed so far, the order in which the users of a service facility get access to that service facility has been determined by queue disciplines such as FCFS, LCFS, and SIRO (see also Section 4.2). A common feature of these disciplines is that users are characterized by the order of their arrival at the queueing system, but no distinction is made among different *classes* of users with reference to the type or the length of service they request. It is natural to ask what effect these disciplines have on the expected total system time, $\bar{W}$ (or on $\bar{W}_q$ or $\bar{L}$ or $\bar{L}_q$), experienced by users of a queueing system.

Suppose, for example, that in the case of the single operator with an infinite number of lines at the center for emergency calls which we discussed earlier (Section 4.6), one wanted to choose among the FCFS, SIRO, and LCFS disciplines. To maintain a FCFS (or LCFS) sequencing of calls, it is necessary to purchase an electronic call-sequencing device. Does such a device contribute anything to improving the quality of service as perceived by callers to our emergency center?

To answer this question, one can reexamine the analysis that led to (4.38), the expression for $\bar{W}$ for that $M/M/1$ system with infinite capacity. Note that the state-transition diagram of Figure 4.6 is unaffected by the queue discipline

in use—all that matters is the total number of calls on hand. Thus, the analysis leading to (4.38) is identical for FCFS, SIRO, and LCFS, and consequently, at least for this example,

$$\bar{W}_{\text{FCFS}} = \bar{W}_{\text{LCFS}} = \bar{W}_{\text{SIRO}} \tag{4.98}$$

Moreover, since Little's formula (4.10) is valid for any queue discipline (see Section 4.4), the same is true for $\bar{L}$, $\bar{L}_q$, and $\bar{W}_q$.

Does this mean that callers will be indifferent to the queue discipline? Probably not, for the *distributions* of the total system times and of the waiting times (but *not* of the number of calls in the system or in queue) will be different in the three cases. To understand why, consider a random caller who at the instant when she is connected with the emergency center is told that k other callers are also waiting for service, while the operator is currently busy with another call. In a FCFS queue this information is valuable to the new caller in estimating her expected waiting time, in a SIRO queue less so, and in a LCFS queue it is practically useless. We infer from the above that, for this example,

$$\text{Var}\,(W_{\text{FCFS}}) < \text{Var}\,(W_{\text{SIRO}}) < \text{Var}\,(W_{\text{LCFS}}) \tag{4.99}$$

where Var (W) denotes the variance of the total system occupancy time.

One can extend the same reasoning that led to (4.98) to any queueing system, and indeed the following statement can be made [KLEI 76]:

> As long as the queueing discipline sequences the users of a queueing system in a way that is independent of their service time (or any measure of their service time), the distribution of the number in the system—and the expected total system time and waiting time—will be invariant to the queue discipline.

Note that the foregoing statement applies to other conceivable queue disciplines in addition to FCFS, SIRO, and LIFO.

The validity of (4.99) for any queueing system can also be argued intuitively (the inequalities become equalities for systems with an infinite number of servers). However, in all but a few cases—such as the $M/G/1$ system—it is very difficult to obtain Var (W_{SIRO}) or Var (W_{LCFS}), even with advanced techniques.

In practice, though, numerous queueing systems sequence the access of users to service facilities by using criteria related to the length of service or to the type of service requested. The latter criterion (type of service requested) is particularly often used in the case of urban services and, especially, of emergency urban services. For instance, police car dispatchers almost always

classify reported incidents into one of several categories and accord higher priority to some categories over others (e.g., calls about "crimes in progress" receive more immediate attention than reports of "missing, probably stolen items"). Similarly, in emergency medical rooms of large hospitals, certain types of patients are given higher priority than other types. We shall examine next the queueing characteristics of systems of this type.

4.9.1 Preemptive and Nonpreemptive Priorities

Priorities that differentiate among classes of users are generally classified as *preemptive* or *nonpreemptive*. In either case, as soon as service to any particular user has been completed, the system chooses as the next user a representative of the highest-priority class present. (Priorities *within* classes are determined according to FCFS, LCFS, SIRO, or some other discipline.) The difference, however, lies in the fact that in systems using preemptive priorities, high-priority users are never kept waiting in favor of lower-priority ones. That is, the system stands ready to interrupt service to any present occupant of the service facility, immediately upon arrival of a user belonging to a class with higher priority than that of the present facility occupant. By contrast, nonpreemptive systems never interrupt a service to a user once that service has begun—even if a higher-priority user arrives at the system while this service is going on.

For preemptive systems, what happens to users who get "ejected" from the service facility is in itself a matter to be specified. In some systems, service to the ejected user—once that user eventually regains hold of the service facility—continues right from the point where it was interrupted. In other systems, service may have to be restarted from scratch. (Note that for users whose service times are negative exponential, these two cases are indistinguishable.) It is also conceivable that an ejected user may be assigned to a priority class higher than his former one, in compensation for being ejected.

It should also be noted that many queueing systems—especially in the urban environment—operate with different priority rules for different user classes. For instance, it is possible that users in a class with very high-priority status may obtain preemptive service, whereas users of medium importance may enjoy only nonpreemptive priority over users in the low-priority classes. Such is the case in police dispatching where an incident reporting "a police officer in danger" is almost always accorded preemptive priority, while other types of high-priority incidents may be nonpreemptive. Finally, it is also possible that different queue disciplines (e.g., FCFS, SIRO, etc.) may be used *within* different classes of users at the same queueing system. Thus, users who belong to, say, class A may queue up according to a FCFS order, whereas users in another class, B, may be served in random order.

If nothing else, it should be clear from the above that there exist a bewil-

dering number of variations of queueing systems with class priorities. Of those variations, queueing theorists have studied with some success a few and have derived an interesting (but hardly exhaustive) set of results. We shall review now some of the most important and useful of those results, always with reference to the type of queueing model shown in Figure 4.15. Facility users in this model are separated into a number, say r, of distinct user classes. Each class is assigned a priority number k, $k = 1, 2, \ldots, r$, for use of the service facility. By convention, the *smaller* the priority number, the *higher* the priority of the class.

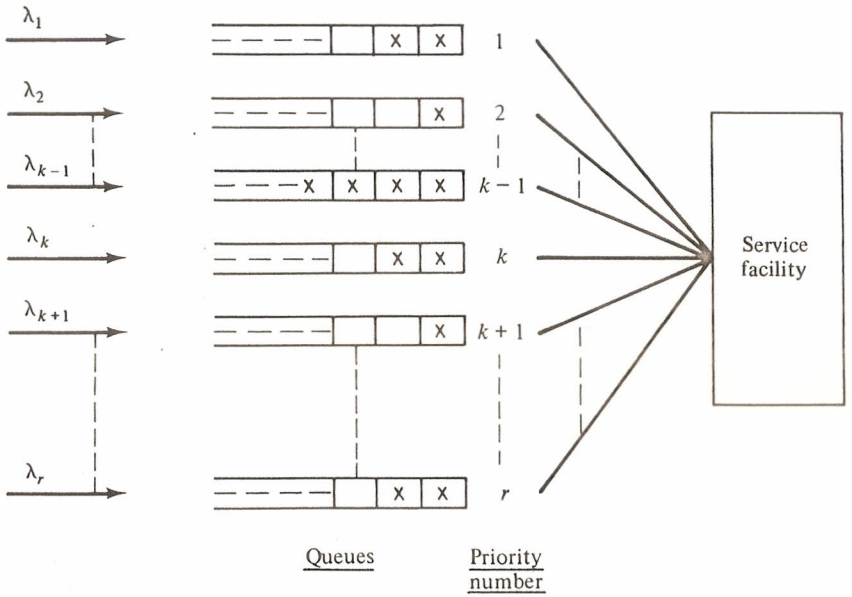


FIGURE 4.15 Schematic representation of a queueing system with r priority classes.

Each of the r queues will be assumed to run on a FCFS basis, but any given priority class cannot obtain access to the service facility unless no other user belonging to a higher-priority (lower-k) class is present in the queues. However, whether user service is ever interrupted or not will depend on whether the queueing system uses preemptive or nonpreemptive priorities.

Finally, we note at the outset that in all cases that will be examined in this section, it is assumed that *arrivals* for each priority class are *Poisson* with arrival rate λ_k for priority class k.

Nonpreemptive priorities in a M/G/1 system. We shall consider first the case in which the queueing model of Figure 4.15 contains a single server that operates under a nonpreemptive priority regime. Moreover, we shall assume

that the random variable S_k, which describes the service time for users in priority class k, has a general probability density function, expected value $1/\mu_k$, and second moment $E[S_k^2]$. We shall also define the quantity $\rho_k \triangleq \lambda_k/\mu_k$. Thus, the queueing system of Figure 4.15 is a $M/G/1$ system with utilization ratio given by

$$\rho = \rho_1 + \rho_2 + \cdots + \rho_r \tag{4.100}$$

The system queueing capacity is assumed to be infinite.

Under these assumptions, we shall now proceed to derive an expression for the average waiting time in queue for a user in priority class k, $\bar{W}_{qk}$. To do this, let us consider the arrival of a random user from class k at the queueing system. We can immediately write an expression for $\bar{W}_{qk}$ (in steady state) as follows:

$$\bar{W}_{qk} = \bar{W}_0 + \sum_{i=1}^{k} \frac{1}{\mu_i} \bar{L}_{qi} + \sum_{i=1}^{k-1} \frac{1}{\mu_i} \bar{M}_i \tag{4.101}$$

where $\bar{W}_0$ = expected remaining time in service for the user who occupies the server at the time when the new user (from class k) arrives at the queueing system

$\bar{L}_{qi}$ = expected number of users in priority class i who are already waiting in queue at the instant when the new user (from class k) arrives

$\bar{M}_i$ = expected number of users in priority class i who will arrive while the newly arrived user (from class k) waits in the queue

To comprehend the meaning of (4.101) it is important to notice that the first summation on the right includes user classes 1 through k (*including* k) since users from these classes who are already waiting when our new user arrives at the system will be served before the new user. By contrast the new user takes precedence over all users of classes $k + 1$, $k + 2, \ldots, r$ who are already waiting. Therefore, these classes do not affect the expected waiting time of the new user and thus do not appear in (4.101). Note also that the second summation is for classes 1 through $k - 1$ (does *not* include k) since users from these classes who arrive while our new user from class k is waiting will take precedence over the user from class k.

We now set out to evaluate the quantities $\bar{W}_0$, $\bar{L}_{qi}$, and $\bar{M}_i$ in (4.101). Beginning with $\bar{W}_0$, note that if the server at the instant of the new user's arrival is occupied by a user from priority class i, we have [see our discussion of random incidence in Chapter 2 and, in particular, (2.66)]

$$(\bar{W}_0 \mid i) = \frac{E[S_i^2]}{2E[S_i]} = \frac{\mu_i E[S_i^2]}{2}$$

But since we have Poisson arrivals for users of class k, the probability that a user of class i will be occupying the server at the instant of our random arrival is simply equal to the fraction of time, ρ_i, that users of type i occupy the server. Thus,

$$\bar{W}_0 = \sum_{i=1}^{r} \rho_i(\bar{W}_0 \mid i) = \sum_{i=1}^{r} \frac{\rho_i \mu_i E[S_i^2]}{2} = \sum_{i=1}^{r} \frac{\lambda_i E[S_i^2]}{2} \tag{4.102}$$

For the value of $\bar{L}_{qi}$, we can use our fundamental relation (4.11) and write

$$\bar{L}_{qi} = \lambda_i \bar{W}_{qi} \tag{4.103}$$

(Note that the $\bar{W}_{qi}$ are still unknown.) Finally, for the expected number of users of class i, $\bar{M}_i$, who arrive during the time when the new user from class k is waiting in line, it is clear that since the arrival process for type i users is Poisson with rate λ_i,

$$\bar{M}_i = \lambda_i \bar{W}_{qk} \tag{4.104}$$

Substituting (4.103) and (4.104) in (4.101),

$$\bar{W}_{qk} = \bar{W}_0 + \sum_{i=1}^{k} \rho_i \bar{W}_{qi} + \bar{W}_{qk} \sum_{i=1}^{k-1} \rho_i \tag{4.105}$$

where $\bar{W}_0$ is given by (4.102). Solving now (4.105) for $\bar{W}_{qk}$, we obtain

$$\bar{W}_{qk} = \frac{\bar{W}_0 + \sum_{i=1}^{k} \rho_i \bar{W}_{qi}}{1 - \sum_{i=1}^{k-1} \rho_i} \qquad \text{for } k = 1, 2, \ldots, r \tag{4.106}$$

We can now solve (4.106) recursively, beginning with $\bar{W}_{q1}$, then with $\bar{W}_{q2}$, and so on. After a moderate amount of algebra (try deriving $\bar{W}_{q1}$ and $\bar{W}_{q2}$), it becomes clear that

$$\bar{W}_{qk} = \frac{\bar{W}_0}{(1 - a_{k-1})(1 - a_k)} \qquad \text{for } k = 1, 2, 3, \ldots \tag{4.107}$$

where, for convenience, we have used the notation[11]

$$a_k \triangleq \sum_{i=1}^{k} \rho_i \tag{4.108}$$

and $\bar{W}_0$ is as given by (4.102).

Expression (4.107) is probably the best-known result in the literature on queueing systems with priorities. From it and through use of (4.11), (4.10), and (4.13), one can obtain expressions for $\bar{L}_{qk}$, $\bar{W}_k$, and $\bar{L}_k$ for all classes of

[11] In the expression for $\bar{W}_{q1}$, we must set $a_0 = 0$.

users k. It should also be emphasized that (4.107) demonstrates clearly the fact that in a nonpreemptive priority system, steady state may be reached for some priority classes and not for others. From (4.107), priority class k will reach steady state as long as $a_k < 1$, since for that condition $\bar{W}_{qk}$ is positive and finite. That is, if there is an integer p ($1 \leq p \leq r$) such that $\rho_1 + \rho_2 + \cdots + \rho_p < 1$ while $\rho_1 + \rho_2 + \cdots + \rho_p + \rho_{p+1} \geq 1$, then the p highest-priority classes reach steady-state delays while users in classes $p + 1$ through r experience unbounded waiting times. In this case, (4.107) must be modified slightly to account for the fact that, *in steady state*, priority classes 1 through p occupy the server for a fraction of time equal to ρ_k each ($k = 1, 2, \ldots, p$); class $p + 1$ occupies the server for a fraction of time equal to $1 - a_p$; and, finally, classes $p + 2$ through r do not ever obtain access to the server. We then have

$$\bar{W}_{qk} = \begin{cases} \dfrac{\sum_{i=1}^{p} \dfrac{\rho_i E[S_i^2]}{2E[S_i]} + \dfrac{(1 - a_p)E[S_{p+1}^2]}{2E[S_{p+1}]}}{(1 - a_{k-1})(1 - a_k)} & \text{for } k \leq p \\ \infty & \text{for } k > p \end{cases} \tag{4.107a}$$

Note that (4.107a) reduces to (4.107) for $p = r$. Note also that the W_{qk} *do not* depend on users in priority classes lower than k, except for the contribution of these users to the numerator of (4.107) or of (4.107a).

4.9.2 Important Optimization Result

A question that naturally arises in the design of queueing systems with priorities is how these priorities should be assigned to accomplish some desirable objective with regard to the performance of the queueing system. Obviously, the answer to this question will depend on the characteristics of the queueing system at hand and on the nature of the desirable objective. In general, questions of this type are among the most difficult to deal with in queueing theory.

A most useful result in this area has been obtained for the nonpreemptive $M/G/1$ priority system which we have just analyzed. We consider again r distinct classes of users with Poisson arrivals (at a rate λ_k for class k) and general service times (with mean $1/\mu_k$ for class k) and with the priority arrangement described through Figure 4.15. Let us assume that for each class of users there is a cost c_k ($k = 1, 2, \ldots, r$) for each unit of time that a user from this class spends in the system (in queue or in service).

Suppose, then, that the objective is to minimize the average cost to all users per unit of time (for steady-state conditions); that is, we wish to minimize

$$C = \sum_{i=1}^{r} c_i \bar{L}_i = \sum_{i=1}^{r} c_i \rho_i + \sum_{i=1}^{r} c_i \lambda_i \bar{W}_{qi} \tag{4.109}$$

Then the following holds true:

Theorem: To minimize (4.109) compute for each class k, the ratio $f_k = c_k/(1/\mu_k)$. Then assign priorities according to the relative magnitudes of the r ratios: specifically, the higher the value of the ratio, the higher the priority of the class.

In other words, we *reorder* the subscripts of the ratios f_k so that

$$f_1 \geq f_2 \geq \cdots \geq f_r \tag{4.110}$$

Then the class of users that corresponds to the ratio f_1 (i.e., that incurs the highest cost per unit of service time) is assigned top priority for access to the server, the class corresponding to the ratio f_2 second highest priority, and so on. Note that this result holds only for the case when the costs of system time (or of waiting time) increase linearly with system (or waiting) time for *all* classes of users.

No formal proof of the theorem will be presented here (the reader may wish to consult [KLEI 76], for one). It is simple, however, to argue intuitively for the theorem's validity by recognizing that minimizing C in (4.109) is the same as minimizing the *expected* area between the "cost inflow" and "cost outflow" curves in Figure 4.16. This figure is a modified version of Figure 4.3 in which the vertical axis represents "cost" instead of "number of users." We have no control over the "cost inflow" curve whose expected rate of growth per unit of time is constant and equal to $\sum_{i=1}^{r} c_i\lambda_i$. However, we can maximize the expected rate of growth of the "cost outflow" curve by always

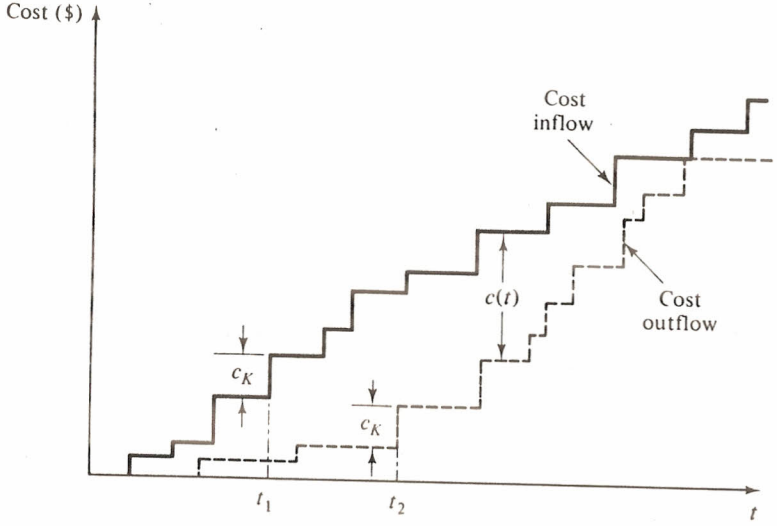


FIGURE 4.16 Cost inflow and outflow at a queueing system. At t_1 a user with cost c_k per minute arrives. The same user leaves the system at t_2. At time t, cost is "accumulated" at the rate $c(t)$ per unit of time.

selecting among the available users the one that "expels" expected cost from the system at the maximum rate per unit of time. This is equivalent to choosing the user with the maximum $c_k \cdot \mu_k$ (= rate of cost "expulsion" per unit time resulting from serving a user of class k) which confirms the theorem. It should also be noted that the first of the two terms on the right-hand side of (4.109) is a constant that is independent of the priority assignments. Therefore, to minimize (4.109) one must minimize the second sum on the right.

The following corollary follows from the theorem.

Corollary: To minimize the *average waiting time* for all users in the system, assign priorities according to the expected service times for each user class: the shorter the expected service time, the higher the priority of the class.

To prove the corollary, all we have to do is set $c_k = 1$ for all k in the theorem (this is equivalent to assigning equal weight to a unit of time's waiting for all classes of users). Then $f_k = \mu_k$ and the result of the corollary follows.

The corollary we have just discussed is sometimes referred to as the "shortest-expected-processing-time-first" (SEPT) rule. Note that, contrary to FCFS, SIRO, and LCFS, SEPT affects the distribution of the number of users present in the queueing system and the expected waiting and total system times by using a characteristic of the length of service time as the criterion for sequencing.

4.9.3 Nonpreemptive Priorities in a $M/M/m$ System

In attempting to extend expressions such as (4.107) to the case of queueing systems with many servers, one encounters the same problems that arise in trying to extend the analysis of $M/G/1$ queueing systems to $M/G/m$ queueing systems. Thus, exact results for multiserver queueing systems with nonpreemptive priorities are available *only* for the case of *negative exponential* service times ($M/M/m$ systems). Moreover, the mathematical analysis becomes intractable unless all r priority classes of users have the same mean service time, μ^{-1}.

Under these assumptions—and using the same notation as for the single-server system—it is not difficult to show (see Problem 4.8) that, as before,

$$\bar{W}_{qk} = \frac{\bar{W}_0}{(1 - a_{k-1})(1 - a_k)} \qquad \text{for } k = 1, 2, \ldots, r \tag{4.111}$$

where $\bar{W}_0$ is now defined as the expected value of the remaining time until any one of the m servers becomes free following the arrival of a new user (of class k) at the queueing system.[12] (If one or more of the servers are free at

[12] As usual for multiserver systems, we define $\rho_i = \lambda_i/m\mu$

the instant of the new arrival, that remaining time is equal to zero.) From the theory of $M/M/m$ queueing systems and, in particular, from expression (4.44), we then have (assuming $\lambda_1 + \lambda_2 + \cdots + \lambda_r < m\mu$)

$$\bar{W}_0 = P\{\text{all servers are busy}\} \cdot E[\text{remaining time} \mid \text{all servers are busy}]$$
$$= \left[\sum_{n=m}^{\infty} P_n\right] \cdot \frac{1}{m\mu} = \frac{P_0}{m\mu} \sum_{n=m}^{\infty} \frac{(\lambda/\mu)^n}{m^{n-m}m!} = \frac{P_0(\lambda/\mu)^m}{m!(1 - \lambda/m\mu)m\mu} \tag{4.112}$$

where P_0 is given by (4.46) and $\lambda = \sum_{k=1}^{r} \lambda_k$.

This queueing model—and expression (4.111)—is among the most applicable in urban operations research. Many urban services can be viewed as multiserver systems with Poisson demands of several types and with (nonpreemptive) priority rules that determine the order in which different types of demands will be serviced. In fact, the model places no restrictions on how the types of demand will be defined. Thus, one can, for instance, classify demands according to the region of a city from which they originate or according to the type of service requested or both (e.g., λ_{ij} might be the rate at which j-alarm fires are generated from region i in a city).

The most restrictive assumption in the model is that service times to all types of demands are assumed to be identical, negative exponential random variables, but even in cases when this is not quite true, expression (4.111) can still be used to obtain some idea of the level of delay experienced by different types of demands for various priority rankings and for different values of m (see also Chapter 5 in [LARS 72b]).

4.9.4 Preemptive Priorities

Some results also exist for the case of systems with preemptive priorities, both for the *preemptive-resume* and the *preemptive-repeat* types. The former refers to the situation in which service to an ejected user, once that user regains access to a server, continues from the point where it was interrupted earlier. By contrast, in the latter type of preemptive system, all service received already is "lost" once a user is ejected from service prior to service completion.

The one result that will be presented here for the preemptive case once again fits the model of Figure 4.15 with a single server. However, it is now assumed that the service time distribution is negative exponential and, in addition, that all classes of users have the *same* expected service rate μ. As we have noted already, the preemptive resume and the preemptive repeat cases are now identical because of the lack of memory of the server.

For this model it can be shown (see, e.g., [KLEI 76]) that

$$\bar{W}_k = \frac{1/\mu}{(1 - a_{k-1})(1 - a_k)} \qquad \text{for } k = 1, 2, \ldots, r \tag{4.113}$$

where $\bar{W}_k$, as usual, signifies the *total* expected system time[13] for a user in class k and, as in the previous section, $a_k = \rho_1 + \rho_2 + \cdots + \rho_k$. More results for preemptive priority systems are derived in Problem 4.9.

Example 2: Repair Work with Priorities

Consider a repair crew charged with performing work for vehicles of the local urban transit authority. Vehicles are separated into two types and breakdowns occur in a Poisson manner at rates of λ_1 and λ_2 for the two types of vehicles, respectively. Repair times are negative exponential and the average service time is the same, $1/\mu$, for both types. Assuming that $\lambda_1 + \lambda_2 < \mu$, we wish to compare the expected system occupancy time (time spent waiting for repair plus time under repair) for each type of vehicle for:

1. The case where type 1 vehicles enjoy preemptive priority over type 2 vehicles.
2. The case where type 1 vehicles enjoy nonpreemptive priority over type 2 vehicles.
3. The case where no priorities exist and breakdowns are repaired in a FCFS fashion, irrespective of the type of vehicle.

Solution

Below we use $\lambda = \lambda_1 + \lambda_2$, $\bar{W}_1$ and $\bar{W}_2$ for the expected system times for case 1, $\bar{W}_1^*$ and $\bar{W}_2^*$ for case 2, and $\bar{W}$ for case 3. Then we have, from (4.107), (4.113), and (4.38):

$$\bar{W}_1 = \frac{1/\mu}{1 - \rho_1} = \frac{1}{\mu - \lambda_1}$$

$$\bar{W}_2 = \frac{1/\mu}{(1 - \rho_1)(1 - \rho_1 - \rho_2)} = \frac{\mu}{(\mu - \lambda_1)(\mu - \lambda)}$$

$$\bar{W}_1^* = \frac{(\rho_1/\mu) + (\rho_2/\mu)}{1 - \rho_1} + \frac{1}{\mu} = \frac{1}{\mu}\frac{\mu + \lambda_2}{\mu - \lambda_1}$$

$$\bar{W}_2^* = \frac{(\rho_1/\mu) + (\rho_2/\mu)}{(1 - \rho_1)(1 - \rho_2)} + \frac{1}{\mu} = \frac{1}{\mu}\left[\frac{\mu^2 - \lambda_1(\mu - \lambda)}{(\mu - \lambda_1)(\mu - \lambda)}\right]$$

and

$$\bar{W} = \frac{1}{\mu - \lambda}$$

It then follows that

$$\bar{W}_1 < \bar{W}_1^* < \bar{W} < \bar{W}_2^* < \bar{W}_2 \tag{4.114}$$

as we would expect.

[13]Note that (4.107) and (4.111) refer to expected *waiting* time.

Furthermore,

$$\bar{W} = \frac{\lambda_1}{\lambda}\bar{W}_1 + \frac{\lambda_2}{\lambda}\bar{W}_2 = \frac{\lambda_1}{\lambda}\bar{W}_1^* + \frac{\lambda_2}{\lambda}\bar{W}_2^* = \frac{1}{\mu - \lambda} \qquad (4.115)$$

That is, the average system occupancy time for *all* breakdowns is identical for the three cases. The relationship shown in (4.115) is a simple example of what are often referred to in queueing theory as "conservation relations." A more general result along these lines and some interesting related questions are developed in [KLEI 76].

4.10 QUEUEING NETWORKS

We have now completed our review of queueing systems (which we defined in Section 4.2 as consisting of one service facility that contains a number of identical servers). In this section we turn our attention to queueing networks (i.e., to sets of interconnected queueing systems). Interconnected in this case implies any combination of *in series* and *in parallel* arrangements.

It has already been indicated or implied several times in this chapter that many urban service systems can be viewed as queueing networks. So far, we have seen many results which are useful in analyzing individual components of these networks. The art of queueing network analysis consists of combining the results (and the analytical techniques) that apply to individual components and drawing conclusions that describe the properties of the complete urban service system under consideration.

The word "art" has been used intentionally above. For one should keep in mind that queueing theory offers very few general results that apply expressly to queueing networks. Therefore, in solving problems that involve such networks, much depends on the ingenuity of the analyst in choosing the "right" simplifying assumptions that preserve the essence of the problem while making the calculation of an approximate solution possible.

In the next two sections, we shall first present what is perhaps the single most useful general result in the analysis of queueing networks. We shall then illustrate by example a widely applicable approach to the analysis of an extensive family of queueing network models.

4.10.1 Important Property of *M/M/m* Queueing Systems

For $M/M/m$ queueing systems with infinite queue capacity, we now state a property that often plays an important simplifying role in the analysis of queueing networks. It is sometimes referred to as the *equivalence property* for $M/M/m$ systems.

Let the arrival process at a $M/M/m$ queueing system with infinite queue capacity have parameter λ. Then, under steady-state conditions (i.e., for $\lambda < m\mu$), the departure process from the queueing system is also Poisson with parameter λ.

The proof of this property is quite involved. However, it is relatively straightforward for the special case $M/M/1$ (see Problem 4.10).

The implications of the property for the analysis of queueing networks are quite obvious. If some "component" (facility) of a queueing network can be modeled as a $M/M/m$ system with infinite capacity, the "output" of this "component" is also Poisson with parameter λ. That is, users will leave this specific facility according to a probability distribution *identical* to the probability distribution for the arrival of users at the facility.

Thus, if the served users of the $M/M/m$ facility are subsequently routed to another facility, the arrival process to this other facility is a Poisson process. In fact, it is clear that if

1. the queueing network consists of, say, K facilities in series (Figure 4.17), each of which contains $m_1, m_2, \ldots, m_K$ $(m_i = 1, 2, 3, \ldots)$ identical servers with negative exponential service and rates μ_i $(i = 1, 2, \ldots, K)$;
2. there is infinite queueing space between successive facilities; and
3. the arrival process for the first facility (facility 1 in Figure 4.17) is Poisson with rate λ;

then, under steady-state conditions, the queueing network can be analyzed as K independent $M/M/m$ queues and the results of Section 4.6 are directly

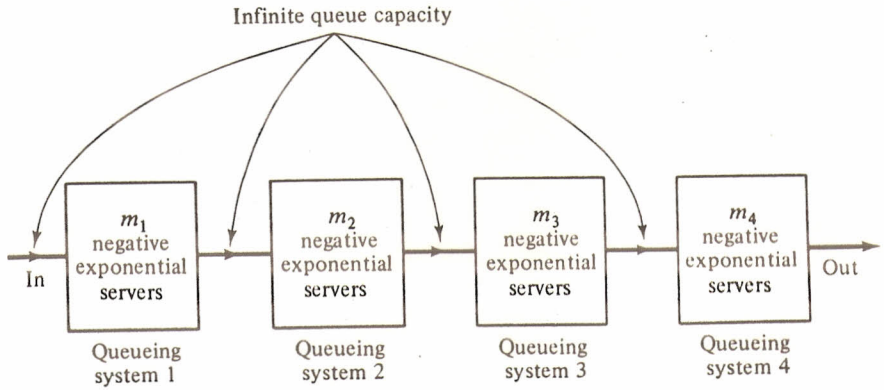


FIGURE 4.17 Poisson input at rate λ at queueing system 1 will result in a Poisson input at rate λ for all four queueing systems.

applicable. The condition for steady state in this case is that $\lambda < m_i\mu_i$ for all i ($i = 1, 2, \ldots, K$).

Three additional notes:

1. It is not necessary that *all* users proceed in series from one facility to the next. For instance, in a "feedforward" network (i.e. a network without any feedback loops) of the type shown in Figure 4.18, departing users from a facility may be divided among N subsequent facilities according to the probabilities p_j ($j = 1, 2, \ldots, N$; $\sum_{j=1}^{N} p_j = 1$). Here p_j represents the probability that a user departing the facility will go to facility j, with the routing of each user determined independently. Since the streams of events that result from the subdivision (in the way described above) or from the mixing together of Poisson streams are also Poisson, the $M/M/m$ analysis can also be used in such, more complicated cases.[14] Thus, as an example for the network shown in Figure 4.18, all flows between facilities are Poisson as long as the arrival processes represented by λ_1, λ_2, and λ_3 are Poisson and as long as the assumptions in the equivalence property hold for each queueing system separately.
2. The equivalence property does not hold for the case of $M/M/m$ queueing systems with *finite* queue *capacity*. It is simple using a state-transition diagram to argue intuitively why this is so. Queueing networks composed of $M/M/m$ queues of this type are often analyzed using the state-transition diagram approach that we shall describe below.
3. Unfortunately, the "negative" statement of the equivalence property is also true: it can be shown [GROS 74] that the *only* type of service distribution which results in a Poisson output, if the input to the queueing system is Poisson, is the negative exponential distribution. Thus, a necessary condition for the output (the departure process) of a $M/G/m$ queueing system to be Poisson is that "$G = M$" (i.e., that the service time distribution be negative exponential). The only exception to this rule is the $M/G/\infty$ queueing system. As a result of this negative side of the equivalence property, the presence in a queueing system of even a single facility with service times that are not negative exponential often creates serious analytical problems.[15]

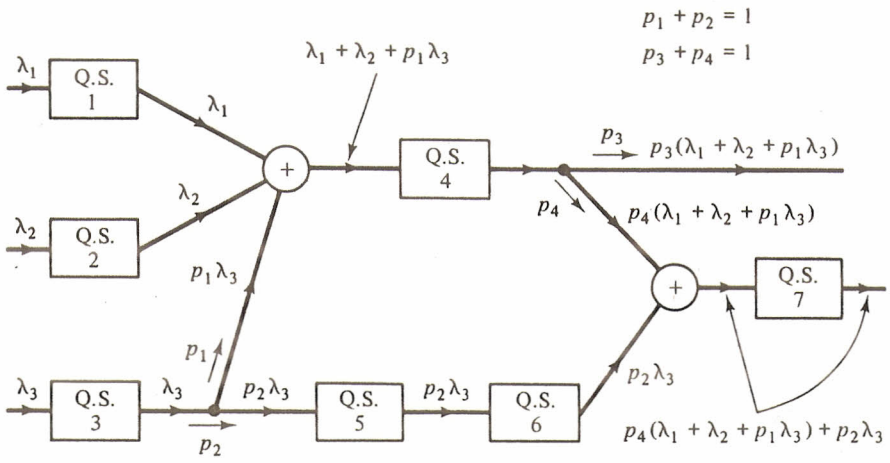


FIGURE 4.18 If the conditions of the equivalence property are satisfied for queueing systems 1 through 7, then all the inputs and outputs in the queueing network above are Poisson with the indicated rates.

[14] The reader should be aware that many "subtleties" arise in the analysis not only of queueing networks that permit some types of feedback but also of some types of feedforward networks (see [KLEI 75] [KLEI 76] and, especially, [DISN 79]). For instance, in networks with feedback, the flow of units within the network may *not* be Poisson.

4.10.2 State-Transition-Diagram Approach to Networks with Blocking Effects

Whenever the arrival of users at a queueing network constitutes a Poisson process and each queueing system in the network has negative exponential service times, it is possible, *at least in principle*, to obtain the properties of the queueing network under steady-state conditions by using the following two-step approach:

STEP 1: Prepare a state-transition diagram that shows all the possible states that the queueing *network* (i.e., the *collection* of queueing systems) can be in and the transitions between states in steady state.

STEP 2: Write and solve the balance equations for the steady-state probabilities of the queueing *network*.

This approach will be one of the fundamental ideas that Chapter 5 will develop with respect to a specific but very rich family of queueing systems. In this section we shall only illustrate the approach with reference to series queueing networks with *blocking* effects. One of the main points that the

[15] For practical purposes, however, the output of a $M/G/m$ queueing system can often be assumed approximately Poisson for large m. This is true due to a limit theorem stating that, when a large number of renewal processes are pooled together, the resulting process approaches Poisson, irrespective of the type of the individual renewal processes.

reader should take note of is how this approach can deal with interactions between the component queueing systems of the network.

Example 3: In-Series Servers with No Waiting Space

Consider the queueing network shown in Figure 4.19. Arrivals at the first facility are Poisson with rate λ. The two facilities contain single identical servers with negative exponential service times and mean service time $1/\mu$. No queues are allowed in front of facility 1 or between the two facilities. Thus, facility 1 is "blocked" whenever it has completed service to a user while facility 2 is occupied with another user. As a result, prospective users of the queueing network are turned away not only when, on arrival, they find facility 1 busy, but also when they find it blocked.

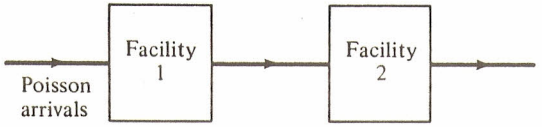


FIGURE 4.19 Two service facilities in series.

Solution

To describe the state of the network we now need two *index numbers:* one to indicate the state of facility 1 and the second to indicate the state of facility 2. Facility 2 can be either empty (0 users in the facility) or contain 1 user; facility 1 can be empty (0 users) or full *and* busy (1 user) or idle *but* full due to blocking from another user in facility 2. We indicate this last condition by the letter b, for "blocked." Thus, the possible network states are 00, 01, 10, 11, and $b1$, with the first index number, by convention, denoting the state of facility 1. Note that state $b0$ cannot exist.

A state-transition diagram for the network of Figure 4.19 is now shown in Figure 4.20. Using the diagram and recalling the "rate in" = "rate out" approach that we have taken in order to write steady-state equations of balance, we have

$$\mu P_{01} - \lambda P_{00} = 0 \tag{4.116a}$$

$$\mu P_{10} + \mu P_{b1} - (\mu + \lambda)P_{01} = 0 \tag{4.116b}$$

$$\lambda P_{00} + \mu P_{11} - \mu P_{10} = 0 \tag{4.116c}$$

$$\lambda P_{01} - 2\mu P_{11} = 0 \tag{4.116d}$$

$$\mu P_{11} - \mu P_{b1} = 0 \tag{4.116e}$$

$$P_{00} + P_{01} + P_{10} + P_{11} + P_{b1} = 1 \tag{4.117}$$

where the meaning of each of the steady-state probabilities P_{ij} is obvious.

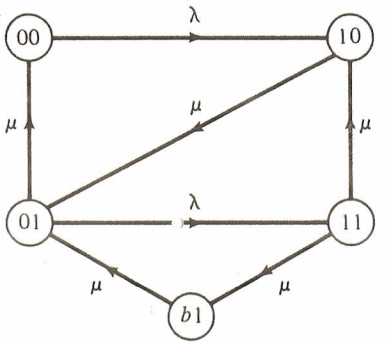


FIGURE 4.20 State-transition diagram for the example for Figure 4.22.

Using (4.117) and any four of the five equations in (4.116) and solving for the five steady-state probabilities, we obtain

$$P_{00} = \frac{2}{F}$$

$$P_{01} = \frac{2\rho}{F}$$

$$P_{10} = \frac{\rho^2 + 2\rho}{F}$$

$$P_{11} = P_{b1} = \frac{\rho^2}{F}$$

where $F = 3\rho^2 + 4\rho + 2$ and, as usual, $\rho = \lambda/\mu$.

Quantities of interest may now be computed using the steady-state probabilities. For instance, for the mean number of users in the system, $\bar{L}$, we have

$$\bar{L} = P_{01} + P_{10} + 2(P_{11} + P_{b1}) = \frac{5\rho^2 + 4\rho}{F}$$

Of more interest, however, are the effects of blocking (and of the zero queueing capacity) on the performance of the queueing network. For example, the mean number of busy servers in the network is

$$\overline{BS} = P_{01} + P_{10} + P_{b1} + 2P_{11} = \frac{4\rho^2 + 4\rho}{F}$$

$$\doteq \begin{cases} 2\rho - 2\rho^2 & \text{for } \rho \ll 1 \\ \frac{8}{9} & \text{for } \rho = 1 \\ \frac{4}{3} - \frac{4}{9\rho} & \text{for } \rho \gg 1 \end{cases} \tag{4.118}$$

while the fraction of potential users lost is given by

$$f = P_{10} + P_{b1} + P_{11} = \frac{3\rho^2 + 2\rho}{F} \doteq \begin{cases} \rho - \frac{1}{2}\rho^2 & \text{for } \rho \ll 1 \\ \frac{5}{9} & \text{for } \rho = 1 \\ 1 - \frac{2}{3\rho} & \text{for } \rho \gg 1 \end{cases} \quad (4.119)$$

It is interesting to observe how $\overline{BS}$ and f change as additional queueing slots are provided in front of one or both service facilities.

Theoretically, this type of approach can, as noted above, be applied to any queueing network—no matter how complex—as long as the queue capacities for every element of the network are *finite*.[16] When all queue capacities are finite, there is a finite number of possible network states and, consequently, a finite number of balance equations which determine the steady-state probabilities (see Problem 4.13).

In practice, however, the number of states increases rapidly with the number of network elements, while the writing and solving of steady-state balance equations turns progressively more difficult as the complexity of network topologies increases. To appreciate this, imagine the simple network of Figure 4.19 but extended by two more facilities in series: thus, we have an in-series network consisting of four single-server facilities with no queueing slots available in front of the first service facility or in the space between facilities. It then turns out that a full 34 different network states exist, with each state requiring four index numbers (one for each of the four servers).

The approach can also be applied to networks where some or all queue capacities are infinite. However, the number of balance equations is also infinite in this case, since we must write one balance equation for each state. Under these conditions it is possible to obtain closed-form solutions for the steady-state probabilities only if some type of "structure" is detected in the infinite equations. [We did observe such structures in our discussion of birth-and-death queueing systems (Sections 4.5 and 4.6) and were thus able to solve a system with an infinite number of balance equations.] The hypercube queueing model, which is presented in Chapter 5, is an excellent example of a queueing system in which this structure-detection approach applies.

[16] Remember that this discussion assumes Poisson arrivals at the network and negative exponential service times for all servers.

4.11 TIME-DEPENDENT ANALYSIS OF THE *M/M/m* QUEUEING SYSTEM

The applications of queueing theory that we have discussed so far in this chapter have almost invariably used the assumption of steady state and thus have ignored the time dimension. There are many situations, however, where it is precisely the fluctuation of congestion effects with time that we are most interested in.

Consider, for example, the problem of determining personnel requirements for the operation of toll booths at a bridge, tunnel, or highway leading into a city. It would clearly be foolish to maintain the same number of open booths from, say, 2 to 4 A.M. as during the peak traffic hours of 7 to 9 A.M. This situation calls for an analysis that recognizes explicitly the fact that demand (i.e., the rate at which cars arrive at the toll-collection area) varies considerably over time. Similarly, in deciding on the number of police patrol units to be deployed over a 24-hour period, a police administrator would do well to take into account the fact that the rate of calls to police dispatchers is at a peak between 5 P.M. and midnight in most cities and then falls rapidly to very low levels during the early morning hours.

One way to deal with such problems is to subdivide the time period of interest into shorter time periods, during which the demand rates and the service rates can be considered approximately constant and perform a separate queueing analysis for each of these shorter time periods, assuming that the steady-state results are valid within each of them. There are numerous examples of this type of analysis, including a classical one [EDIE 54] that was applied to the toll-booth problem that we have just described.

This approach, however, may sometimes be inadequate: the demand and/or service rates may be changing too rapidly over time for the steady-state expressions (which assume the existence of "long-term equilibrium conditions") to be valid. The approach also fails if it happens that for one or more of the short periods of time *average* demand exceeds the service rate—a case for which no steady-state expressions exist.

In some cases of this type, numerical solution techniques can often be of some help. We shall describe below one such technique as it applies to $M/M/m$ queueing systems.

The analysis below will be performed with reference to the center-for-emergency-calls example (see Section 4.6), for which it will be assumed that:

1. The arrival of calls at the $M/M/m$ system is a Poisson process with a known time-dependent rate $\lambda(t)$ for $0 \leq t \leq T$.
2. The m operators/servers are independent and identical, and service

times are independent and negative exponential with mean μ^{-1} (μ is *not* a function of time).

3. The queueing system has a capacity of K ($K \geq m$) and calls that find the system full are lost.
4. The state of the system at time $t = 0$ is known probabilistically (see below for further explanation).

Let $P_n(t)$, $i = 0, 1, 2, \ldots, K$, denote the probability that at time t there are n calls present—being answered or waiting for an operator. Using the state-transition diagram 4.9, it is then simple to write a system of first-order differential equations that describe this queueing system [cf. (4.20) and (4.21)]:

$$\frac{dP_0(t)}{dt} = -\lambda(t)P_0(t) + \mu P_1(t) \tag{4.120a}$$

$$\frac{dP_n(t)}{dt} = \lambda(t)P_{n-1}(t) - [\lambda(t) + n\mu]P_n(t) + (n+1)\mu P_{n+1}(t)$$

$$\text{for } 1 \leq n \leq m-1 \tag{4.120b}$$

$$\frac{dP_n(t)}{dt} = \lambda(t)P_{n-1}(t) - [\lambda(t) + m\mu]P_n(t) + m\mu P_{n+1}(t)$$

$$\text{for } m \leq n \leq K-1 \tag{4.120c}$$

$$\frac{dP_K(t)}{dt} = \lambda(t)P_{K-1}(t) - m\mu P_K(t) \tag{4.120d}$$

Given a set of probability values $P(0)$ that describe the state of the queueing system at $t = 0$ [i.e., $P(0) = \{P_0(0), P_1(0), \ldots, P_K(0)\}$ such that $\sum_{n=0}^{K} P_n(0) = 1$], the $K + 1$ equations (4.120) can be solved iteratively on a computer. To do this, one must choose a time interval Δt which is sufficiently small to be consistent with the Poisson assumptions for the arrival and service processes (i.e., the probability of two or more user arrivals or service completions during Δt must be very small) and then use the approximation

$$P_n(t + \Delta t) = P_n(t) + \frac{dP_n(t)}{dt} \cdot \Delta t \qquad n = 0, 1, 2, \ldots, K \tag{4.121}$$

In this way, beginning with $P(0)$, one first solves for $P(\Delta t) = \{P_0(\Delta t), P_1(\Delta t), P_2(\Delta t), \ldots, P_K(\Delta t)\}$, then for $P(2\Delta t)$, $P(3\Delta t)$, and so on, for the whole period of interest T. This type of numerical solution can be accomplished with the aid of standardized computer programs (such as the IBM Continuous System Modeling Program—CSMP) which are designed specifically for accurate iterative solution of systems of differential equations such as (4.120).

Once the probabilities $P_n(t)$ have been computed, other quantities of interest can be derived as well. For instance, for the expected number of calls *in queue* at time t, we have

$$L_q(t) = \sum_{n=m+1}^{K} (n - m)P_n(t) \tag{4.122}$$

and for the expected waiting time *in the queue* for an accepted call that arrives at time t,

$$W_q(t) = \frac{1}{m\mu} \sum_{n=m}^{K-1} (n - m + 1)P_n(t) \tag{4.123}$$

Exercise 4.7 Argue the validity of (4.123). Why not use the expression $W_q(t) = L_q(t)/\lambda(t)$ in this case?

In urban problems, the foregoing time-dependent queueing model arises most often in situations where the demand rate, $\lambda(t)$, is periodic with period $\tau = 24$ hours. This, for example, would be the case for the toll-booth and police-allocation problems described at the beginning of this section. In that case it is reasonable to expect and it has been shown [KOOP 72] that the numerical solution of (4.120) will also become periodic eventually, independent of the starting conditions $P(0)$, as long as[17]

$$\bar{\lambda} = \frac{1}{\tau} \int_0^{\tau} \lambda(t)\, dt < m\mu \tag{4.124}$$

In that case, we shall have $P_n(t + \tau) = P_n(t)$ for all t greater than some threshold value. For relatively low utilization systems, the probabilities $P_n(t)$ will converge to the periodic solution very quickly (e.g., after a few hours of the first day when $\tau = 24$ hours.) Convergence, however, cannot be verified until (4.120) have been solved for a second (or third, or more, if necessary) day.

Although, in order to obtain numerical solutions of (4.120) one must naturally have a finite number ($= K + 1$) of equations, one can still use a trial-and-error approach to obtain solutions for cases where no calls/users can be turned away [provided that (4.124) holds]. All one has to do is specify K to be sufficiently large so that at no time does $P_K(t)$, the probability of a full system, exceed some acceptable value [e.g., $P_K(t) < 10^{-4}$ for all t]. Problems with K as high as 600 have been solved at reasonable computer cost [HENG 75]. An application of the $M/M/m$ time-dependent model to the deployment of police patrol cars through the day in New York City is described in [KOLE 75].

[17] If (4.124) does not hold, the queueing system will eventually become saturated.

References

[BARN 78] BARNETT, A. I., "Control Strategies for Transport Systems with Nonlinear Waiting Costs," *Transportation Science, 12* (1), 119–136 (February 1978).

[BRUM 71] BRUMELLE, S. L., "Some Inequalities for Parallel Server Queues," *Operations Research, 19*, 402–413 (1971).

[CHAI 71] CHAIKEN, J. M., *The Number of Emergency Units Busy at Alarms Which Require Multiple Servers*, Report R-531, Rand Corporation, Santa Monica, Calif., 1971.

[DISN 79] DISNEY, R. L., *Queueing Networks*, Department of Industrial Engineering and Operations Research, Virginia Polytechnic Institute and State University, Blacksburg, Va., 1979.

[EDIE 54] EDIE, L. C., "Traffic Delays at Toll Booths," *Operations Research, 2*, 107–138 (1954).

[GROS 74] GROSS, D., AND C. M. HARRIS, *Fundamentals of Queueing Theory*, Wiley, New York, 1974.

[HENG 75] HENGSBACH, G., AND A. R. ODONI, *Time-Dependent Estimates of Delays and Delay Costs at Major Airports*, Report R75-4, MIT Flight Transportation Laboratory, Cambridge, Mass., 1975.

[IGNA 78] IGNALL, E. J., P. KOLESAR, AND W. E. WALKER, "Using Simulation to Develop and Validate Analytic Models: Some Case Studies," *Operations Research, 26*, 237–253 (1978).

[KING 62] KINGMAN, J. F. C., "On Queues in Heavy Traffic," *Journal of the Royal Statistical Society, Series B, 24*, 383–392 (1962).

[KLEI 75] KLEINROCK, L., *Queueing Systems, Volume I: Theory*, Wiley, New York, 1975.

[KLEI 76] KLEINROCK, L., *Queueing Systems, Volume II: Computer Applications*, Wiley, New York, 1976.

[KOLE 75] KOLESAR, P. J., K. L. RIDER, T. B, CRABILL, AND W. E. WALKER, "A Queueing-Linear Programming Approach to Scheduling Police Patrol Cars," *Operations Research, 23*, 1045–1062 (1975).

[KOLL 74] KÖLLERSTRÖM, J., "Heavy Traffic Theory for Queues with Several Servers: I," *Journal of Applied Probability, 11*, 544–552 (1974).

[KOOP 72] KOOPMAN, B. O., "Air Terminal Queues under Time-Dependent Conditions," *Operations Research, 20*, 1089–1114 (1972).

[LARS 72a] LARSON, R. C., "Improving the Effectiveness of New York City's 911," in *Analysis of Public Systems*, A. W. DRAKE, R. L. KEENEY, AND P. M. MORSE, eds., MIT Press, Cambridge, Mass., 1972.

[LARS 72b] LARSON, R. C., *Urban Police Patrol Analysis*, MIT Press, Cambridge, Mass., 1972.

[LITT 61] LITTLE, J. D. C., "A Proof of the Queueing Formula $L = \lambda W$," *Operations Research, 9*, 383–387 (1961).

[MARC 78] MARCHAL, W. G., "Some Simpler Bounds on the Mean Queueing Time," *Operations Research, 26*, 1083–1088 (1978).

[MARS 68] MARSHALL, K. T., "Some Inequalities in Queueing," *Operations Research, 16*, 651–665 (1968).

[OSUN 72] OSUNA, E. E., AND G. F. NEWELL, "Control Strategies for an Idealized Public Transportation System," *Transportation Science, 6* (1), 52–72 (February 1972).

[STID 74] STIDHAM, S., JR., "A Last Word on $L = \lambda W$," *Operations Research, 22*, 417–421 (1974).

Problems

4.1 *Model of a taxi station* Consider a taxi station where taxis looking for passengers and passengers looking for taxis arrive according to Poisson processes, with mean rates per minute of 1 and 1.25. A taxi will wait no matter how many other taxis are in line, but an arriving passenger waits only if the number of passengers already waiting for taxis is two or less. Assuming steady-state conditions, find:

a. The mean number of taxis waiting for passengers.

b. The mean number of passengers waiting for taxis.

c. The mean number of passengers who in the course of an hour do not join the waiting line because at least three passengers were already waiting.

4.2 *Queueing system with balking* We have already noted (cf. Section 4.6) that it is usually difficult to obtain closed-form solutions for queueing systems in which there is user balking or where the rate of service depends on the number of users present. There are exceptions to this.

a. Consider a single-server system with infinite system capacity, Poisson arrivals, and negative exponential service times for which the rates of user

arrivals and of service [cf. (4.58)] are

$$\lambda_n = \frac{\lambda}{n+1} \qquad n = 0, 1, 2, \ldots$$
$$\mu_n = \mu \qquad n = 1, 2, 3, \ldots$$

(In the above, λ is the arrival rate when the system is empty.) Show that

$$P_n = \frac{(\lambda/\mu)^n}{n!} e^{-\lambda/\mu} \qquad n = 0, 1, 2, \ldots$$

What is the condition for steady state in this case? Find an expression for ρ.

b. Suppose we now assume that there is no balking ($\lambda_n = \lambda$) but that the service rate depends on the number of users present ($\mu_n = c_n \cdot \mu$ for $n = 1, 2, \ldots$). Find a form for c_n such that the expression for P_n becomes identical, as in part (a). To what classical queueing system is this new system equivalent, as far as the service rate is concerned?

c. Find $\bar{L}$ for case (a). Show that, for this case,

$$\bar{W} = \frac{\lambda}{\mu^2(1 - e^{-\lambda/\mu})}$$

4.3 *Repairs of MTA buses* The metropolitan transit authority of a region wishes to establish a crew of auto mechanics that will be responsible for repairing the authority's buses. The crew is stationed at a single location.

Bus breakdowns occur randomly (Poisson process) at a mean rate of one per hour. The time required to fix a bus has a negative exponential distribution (regardless of crew size). The expected repair time required by a one-worker crew would be 2 hours.

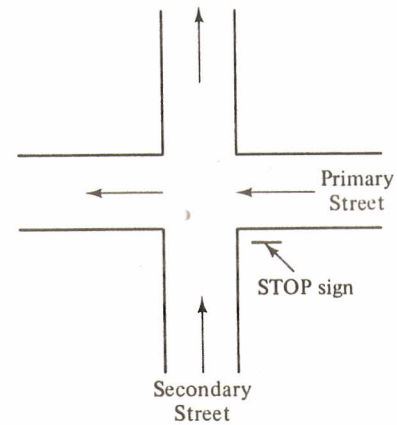
The cost per hour for each member of a repair crew is \$10.00. The cost that is attributable to not having a bus in use (i.e., a bus standing at the bus repair shop) is estimated to be \$40.00 per hour. (Both men and buses are on 8-hour days.)

Assume that the mean service rate of the repair crew is proportional to its size. What should the crew size be in order to minimize the expected total cost of this operation per hour? Repeat this question but with the mean service rate proportional to the square root of the crew size.

4.4 *Waiting at a street intersection* Consider the intersection of two city streets shown in Figure P4.4(1). Both streets are one-way. One of them is designated as the "primary" street and vehicles on it have priority at the intersection. On the "secondary" street there is a stop sign at the intersection.

Consider a car on the secondary street that arrives at the intersection at a random time while no other car on the secondary street is waiting to cross the intersection. Assume that:

1. The car on the secondary street arrives at the intersection at a random time.



2. The vehicles on the primary street do not slow down or yield to vehicles on the secondary street at the intersection.
3. The headways, H, between vehicles on the primary street are independent and identically distributed random variables with pdf $f_H(t)$, expected value $E[H]$, variance σ_H^2, and so on.
4. The car on the secondary street will cross the intersection as soon as a time gap greater than t_0 (a constant) is perceived before arrival of the next vehicle on the primary street (assume that drivers perceive such things correctly). Note that a car on the secondary street may cross the intersection immediately upon arrival there, if the remaining time until the arrival of the next primary-street car at the intersection is greater than t_0.

Figure P4.4(2) illustrates this whole situation. Let

Y = time between the instant when the car on the secondary street first arrives at the intersection and the instant when it begins crossing

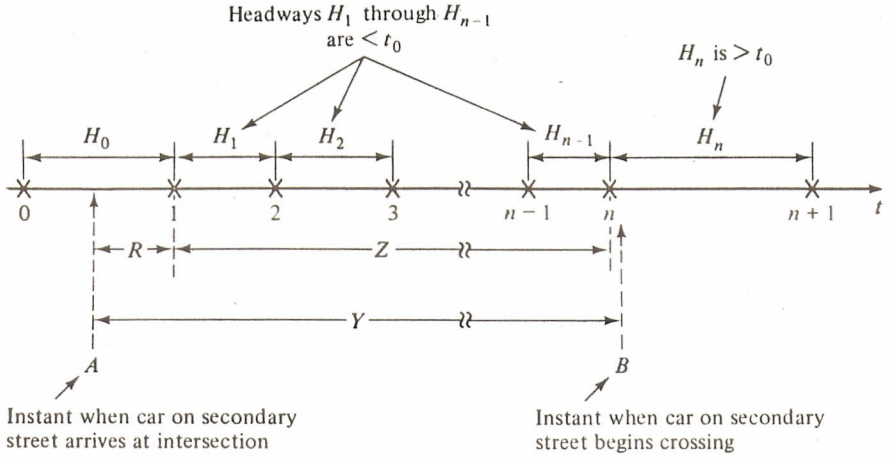
a. Derive an expression for $E[Y]$ in terms of $f_H(t)$ (and its moments) and t_0.

b. Derive an expression for σ_Y^2.

c. Apply your results of parts (a) and (b) to the case where the headways are negative exponentially distributed [i.e., $f_H(t) = \lambda e^{-\lambda t}$ for $t \geq 0$]. Show that

$$E[Y] = \frac{1}{\lambda}[e^{\lambda t_0} - (1 + \lambda t_0)]$$
$$\sigma_Y^2 = \frac{1}{\lambda^2}(e^{2\lambda t_0} - 2\lambda t_0 e^{\lambda t_0} - 1)$$

d. Apply your results of parts (a) and (b) to the case where $t_0 = 4$ seconds and H is uniformly distributed between 2 and 10 seconds.



X = instants when cars on primary street reach the intersection

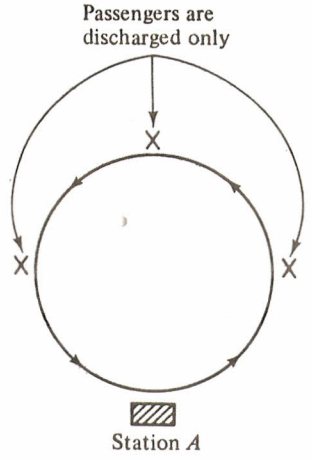
R = remaining time until first primary-street car's arrival

$Y = R + Z$

e. Assuming that all drivers on the secondary street use the same t_0, does your result of part (a) apply to all drivers on the secondary street *once they become first in line*, or only to those drivers who find no one else waiting at the stop sign at the time of their arrival at the intersection? Please explain in a couple of sentences (no mathematical analysis). What about the case of pedestrians crossing a street with no traffic lights (and drivers who do not slow down)?

4.5 *Single-bus transportation system* This problem illustrates, once more, one of the main themes of this book—that the "obvious" thing to do does not always produce the best results. For several types of urban transportation systems (e.g., buses, elevator banks, subways, etc.) it is sometimes better to delay some vehicles than to let them proceed, as soon as possible, with a "trip." This will result in more regular headways (between the passage of vehicles from stops), which, in turn, improves overall system performance. Consider as an example the following simple situation. A bus "system" consists of a single bus that operates on a route (Figure P4.5) with a single stop (station A) for picking up passengers, who are then delivered to other stops along the route. (This could be a primitive model for a local bus system in a suburban community during the evening rush hour. The single pickup stop would be at the train station where commuters from a central business district return from work.) Define H to be the headway between successive departures of the bus from station A. Assume that:

1. The time interval, X, between the instant when the bus leaves station A and the instant when it returns there for the first time is a discrete random



variable with

$$P\{X = x\} = \begin{cases} 0.5 & \text{for } x = 1 \\ 0.5 & \text{for } x = 9 \end{cases}$$

2. Passengers board the bus instantaneously once it returns to station A and the capacity of the bus is sufficiently large so that no one is ever left out.
3. Passengers arrive at station A to catch the bus randomly, according to some probabilistic process which is independent of the location of the bus at any given time.

a. Find the expected time, $E[W]$, that a random passenger will spend at station A if the bus always leaves station A immediately after it arrives there and the passengers board it (instantaneously).

b. Repeat part (a) for the case in which the bus is *held* at the station for three extra time units whenever it returns to station A only one time unit after leaving:

$$H = \begin{cases} 4 & \text{if } x = 1 \\ 9 & \text{if } x = 9 \end{cases}$$

Note that $E[W]$ has decreased now despite the fact that the frequency of bus service has decreased as well.

c. Assume that it has been decided to use the dispatching strategy

$$H = \begin{cases} a_0 & \text{if } x = 1 \\ 9 & \text{if } x = 9 \end{cases}$$

at station A, where a_0 is an unknown constant. Determine the value of a_0 that minimizes $E[W]$.

In general, it has been shown that, for any pdf $f_X(x)$ for X, the *optimal* headway strategy for this problem is to set $H = \text{Max}\ (a_0^*, X)$, where a_0^* is the optimal value of the constant a_0 (note that X is allowed to be any random variable in this case but that the result above is limited to single-bus systems).

d. In part (c) you found that $a_0^* = E^*[W]$, where $E^*[W]$ is the minimum value of $E[W]$. This is not an accident but a general property of optimal headway strategies for this problem [OSUN 72]. Does this suggest a good iterative procedure for solving part (c)?

e. Repeat parts (a) and (c) for the case $f_X(x) = e^{-x}$ for $x \geq 0$.

A good review and more general results for problems of this type has been published by Barnett [BARN 78].

4.6 *Spatially distributed queue with random server location* In this problem we examine a spatially distributed queueing system similar to that of Example 1, Section 4.7. In the process we shall examine how the "advantage of central location" (i.e., the benefits of positioning a facility/server at the geographical center of an urban area) behaves as a function of system utilization. In this problem the server will be randomly located in the district, in contrast to Example 1, in which it was centrally located.

We shall examine here the case in which the rectangular district of Figure 4.10 is patrolled by a police car which is dispatched to incidents within the district. Incidents are served in a FCFS manner, with the patrol car traveling from incident to incident whenever there is a backlog of calls for police assistance. At times when no pending calls exist and the patrol car is free, it remains *stationary at the location of the last incident* that it served, waiting to be dispatched to the next call for assistance. All other assumptions in the problem (e.g., right-angle travel, uniformly distributed incidents in the district, Poisson demands, service time on the scene, travel speeds, etc.) are identical to those of Example 1. It will be assumed, however, that the dimensions of the district are now $2X_0 \times 2Y_0$ miles. This is done to make the results of this problem comparable to those of Example 1, where the ambulance must make *round trips* between the hospital and the incidents that it serves.

a. Obtain expressions for ρ, $\bar{L}$, $\bar{L}_q$, $\bar{W}$, and $\bar{W}_q$ in this case. To avoid any confusion, we specify that the patrol car "begins serving" a particular incident, as soon as the patrol car begins its trip toward that incident. You should also assume that successive service times (= travel time + time on the scene) are statistically independent. [In truth, there is a slight correlation between successive service times (why?).]

b. Assuming exactly the same numerical values as in Example 1 (numerical example), prepare a table for ρ, $\bar{L}$, $\bar{L}_q$, $\bar{W}$, and $\bar{W}_q$, as the rate of calls per hour, λ, increases. Compare these values to those for the ambulance service of Example 1. How do these values differ as λ increases ($\rho \longrightarrow 1$)?

4.7 *Planning for a large parking garage* A parking garage is planned for a large airport which is wholly owned by a city. This garage will serve the needs of "park-

and-fly" passengers at one major section of the airport. It has been decided that the garage's capacity will be geared to the peak season of the airport, which consists of about five consecutive months. The planners believe that the demand (i.e., the number of cars seeking a parking spot at that section of the airport) can be reasonably modeled as a Poisson process with a mean, λ, of about 2,500 cars per day. The time during which any given parking spot is occupied by a car is assumed to be a random variable with some arbitrary pdf, a mean $(1/\mu) = 30$ hours, and finite variance. It has been decided, after much debate, to provide sufficient capacity at the garage so that "a motorist will be able to find a free space there 98 percent of the time"—with a new system planned to direct motorists to areas with free parking spaces. Assuming that drivers who are informed, at the time when they seek entry, that the garage is full will become discouraged and go somewhere else:

a. Estimate approximately how many parking spaces should be provided at the garage. Justify your reasoning, possibly with reference to some queueing theory model.

b. Based on your analysis for part (a)—and provided that all other assumptions in the model are reasonable—discuss which is more important in planning for the size of the garage: the accuracy of the estimated values of λ and $1/\mu$, or the probability of parking availability sought (e.g., 98 percent)?

4.8 *Derivation of expected waiting times* The method used in Section 4.9 for deriving the expected waiting time $\bar{W}_{qk}$ for class k users of a $M/G/1$ system with nonpreemptive priorities is very helpful for many problems involving various queue disciplines or priorities.

a. Use this method to derive an expression for $\bar{W}_q$, the expected waiting time at a $M/G/1$ system with a last-come, first-served (LCFS) queue discipline. Assume infinite queue capacity. Your result should of course be the same as (4.81), the expression for $\bar{W}_q$ for a $M/G/1$ system with FCFS queue discipline.

b. Use this method to derive expressions (4.111) and (4.112) for $\bar{W}_{qk}$ in the case of a $M/M/m$ queueing system with nonpreemptive priorities.

c. Use the same method for deriving expression (4.113) for $\bar{W}_{qk}$ in a $M/M/1$ system with preemptive priorities.

4.9 *Hospital emergency ward with preemptive priorities* The lone doctor in the Puddleduck City Hospital emergency ward encounters two types of patients: emergency and nonemergency patients, who arrive at independent Poisson rates λ_1 and λ_2, respectively. Their treatment times are independent and exponentially distributed with parameters μ_1 and μ_2, respectively.

If an emergency patient arrives during the treatment of a nonemergency patient, the latter's treatment is immediately stopped in favor of the emergency patient. The interrupted treatment is resumed (i.e., from the point at which it was interrupted) when there are no emergency patients present. The treatment of an emergency

patient is *never* interrupted, and within the two groups there is a first-come, first-served discipline.

a. Find the mean number of emergency patients in the system ($\bar{L}_1$).

b. Show that the mean number of nonemergency patients in the system is

$$\bar{L}_2 = \frac{\rho_2}{1-\rho_1-\rho_2}\left[1+\frac{\mu_2}{\mu_1}\left(\frac{\rho_1}{1-\rho_1}\right)\right]$$

where

$$\rho_1+\rho_2<1 \qquad \rho_1=\frac{\lambda_1}{\mu_1} \qquad \rho_2=\frac{\lambda_2}{\mu_2}$$

c. Assume that $\lambda_1/\mu_1=\lambda_2/\mu_2<\frac{1}{2}$. Compare the mean waiting times of both types of patient with and without the preemptive priority system.

Hints: The time a nonemergency patient spends in the emergency ward consists of three components:

1. The time between her arrival and the next time that a nonemergency patient receives attention (T_1).
2. The time to complete treatment on all the nonemergency patients in front of her.
3. The time from her first receiving attention until the time that the treatment is completed (completion time, T).

Show that

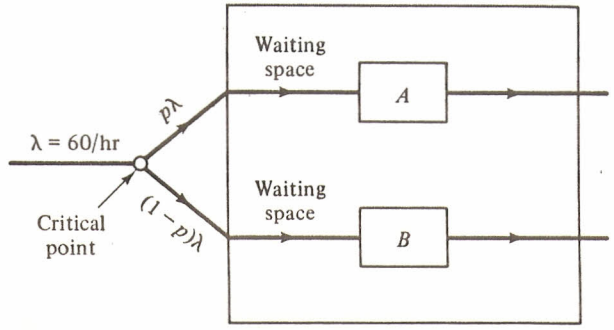
$$E[T_1]=\frac{\lambda_1}{\mu_1^2(1-\rho_1)^2} \qquad E[T]=\frac{1}{\mu_2(1-\rho_1)}$$

4.10 *"Output" of a M/M/1 system* In this problem you are asked to prove the important theorem of Section 4.10 for the case of a $M/M/1$ queueing system. Let a queueing system with a single server have Poisson arrivals at a rate λ, infinite queue capacity, and negative exponential service time with mean $1/\mu$ ($\mu>\lambda$). Show that in the steady state the "output" stream leaving this queueing system is also Poisson at a rate λ.

Hint: What is the pdf for the time between service completions when the server is continually busy?

4.11 *"User-optimal" versus "system-optimal" equilibrium* Two different facilities, A and B, provide the same type of service. Each facility contains a single server with service times distributed as negative exponential random variables. The mean service times are $1/\mu_1=1$ minute and $1/\mu_2=4$ minutes at facilities A and B, respectively. Other than the different expected length of service times, the service that A and B provide is identical (in terms of quality, cost to the user, etc.). (Think, for instance, of two truck-weighing and inspection facilities that differ only by the rate at which they inspect trucks.)

A combined total of $\lambda=60$ customers per hour wish to avail themselves of the service provided by facilities A and B. The situation is pictured in Figure P4.11. Arrivals of customers at the *critical point* are Poisson. At that point each customer makes a choice, independently of all others, on which facility he or she is going to use. This choice is made *without any knowledge* of the status of the queue in front of either facility. Let p denote the probability that a random customer chooses facility A (and $1-p$ that he or she chooses B).



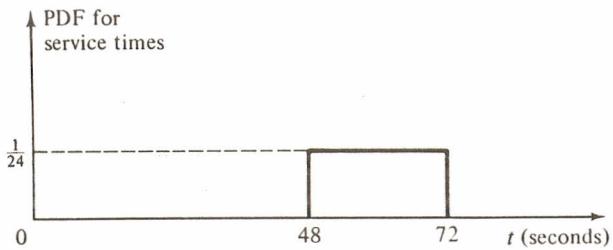
a. Consider the case when the same pool of customers use facilities A and B on a repeated basis (say, once every day). (Think, for instance, of suburban commuters in a town with two major access roads to the central business district. Each of these commuters decides every day, independently of all others, which of the two roads he or she is going to use that day, without knowledge of traffic conditions on either road.) Let us assume that each user is only concerned with minimizing his or her expected time spent in a service facility, $\bar{W}$ (= average waiting time + average service time). It can be reasonably expected that in the long run (system in steady state and in equilibrium—as far as the distribution of customers between the two facilities is concerned) the customers will "distribute" themselves among the two facilities, in the sense that *p will stabilize around a specific value*. What is that value of p?

b. Suppose now that your objective is to minimize the total amount of time (waiting and being serviced) that all customers spend in either of the two facilities per unit of time (with the system in the steady state). This is equivalent to minimizing, say, the "cost" suffered by the community each day, where cost is measured in terms of total commuter hours spent in traffic. You can thus set the value of p yourself (and thus force each arriving customer, independently, to choose between the two facilities on the basis of *your* p). What would you choose as the value of p?

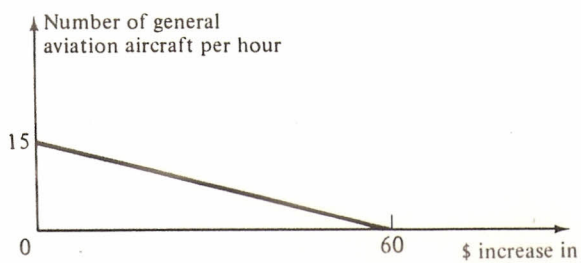
Note: Feel free to use trial and error (rather than an "elegant" approach) in determining p.

c. Can you explain intuitively why the answers to parts (a) and (b) are different? Can you suggest a situation parallel to the one above in highway traffic and how it might be possible to achieve the value of p that you found in part (b) in this case?

4.12 *Imposing fees at a congested facility* It is often necessary to impose congestion tolls to assure that expensive transportation facilities (tunnels, bridges, terminal space at bus terminals, etc.) are used efficiently. Consider a city-owned airport and assume that this airport has a runway that is used only for landings during the peak traffic hours. Under peak conditions, the arrivals of airplanes at the vicinity of the airport are assumed to be Poisson with a rate $\lambda = 55$ aircraft/hour. Of these airplanes, 40 on the average are commercial jets and 15 are small private airplanes ("general aviation"). The pdf for the duration of the service time to a random aircraft landing on the runway is given by the uniform pdf of Figure P4.12(a).



(a)

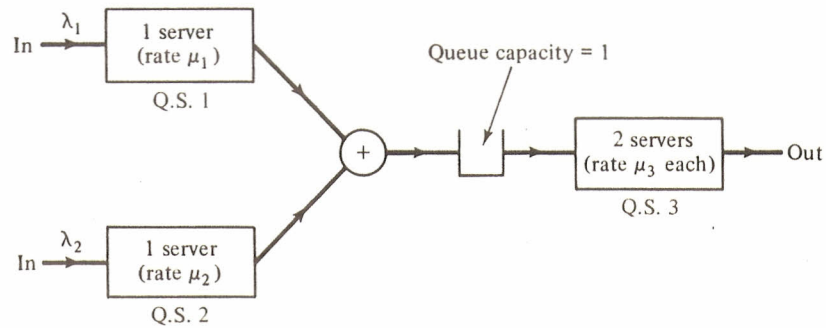


(b)

a. Peak-hour conditions occur during 1,000 hours/year and the average cost of 1 minute's waiting (i.e., "going into a holding pattern" near the airport) is \$12 for commercial jets (this accounts primarily for the extra fuel spent, the extra flight crew time and extra aircraft maintenance costs). Estimate the yearly costs to the airlines of peak-hour delays at this runway.

b. To alleviate peak-hour congestion, the airport's managers are considering an across-the-board increase in the fees that aircraft pay for using the runway ("landing fees"). While airline demand is insensitive to moderate increases in landing fees, general aviation demand is expected to drop drastically as these fees increase (since there are some good small airports near the city in question which can be used by small aircraft). In fact, a study of the general aviation runway users revealed that the relationship between demand per hour by general aviation and increase in landing fee is as shown in Figure P4.12(b). What is the most desirable amount of increase in landing fees from the point of view of the airlines? (Remember that the airlines will also be paying the higher fees.)

4.13 *Queueing network* In the queueing network of Figure P4.13, all service times are negative exponential and the two arrival processes shown are independent and Poisson.



There is space for one waiting user at the indicated point in front of queueing system 3. Whenever that queueing space is occupied, this "blocks" the departure of users from system 1 *and* from system 2. No space for waiting exists in front of either system 1 or system 2. Prospective users of either system are turned away if that system is either busy or blocked at the time of user arrival.

Write the steady-state balance equations for this queueing network. You will find it useful to define carefully the states of the network and to draw a state-transition diagram for it. Assume that if *both* systems 1 and 2 are "blocked" at any time, system 1 users receive priority whenever the single waiting space in front of system 3 becomes free.

4.14 *Police helicopters and police cars* Assume that in a circular city with a radius of R miles, calls for police service are generated in a Poisson manner at the rate of λ per hour per city square mile. Calls are uniformly distributed over the city.

The police department is contemplating the purchase of k helicopters to respond to certain types of police calls. The helicopters will be flying straight to the location of each incident, at an effective speed of v_H miles/hr.

The special police cars that are presently used for these calls have an effective speed of v_c miles/hr and travel in right-angle distances. The city is rather large and its streets have no particular orientation with respect to the coordinate axes.

The dispatching strategy that the police controller uses, whether operating a helicopter-based or a car-based system, is the following. Whenever all service units are busy, calls for police assistance are placed in a first-come, first-served, infinite-

capacity queue and the first service unit which becomes available is immediately dispatched to the first call in the queue. If, on the other hand, more than one service unit is available, the dispatcher selects one of the available units *randomly* (with no consideration to service unit locations) and dispatches it to the next call.

It is also known that the durations of service for incidents serviced by helicopters or by cars, *after* arrival on the scene, are random variables with negative exponential distributions with average service time equal to $1/\mu_H$ and $1/\mu_c$, respectively. The durations of service to successive incidents are statistically independent.

Finally, it is known that a unit that completes service to a call remains stationary until it is dispatched to a new incident.

a. Make the assumption that *travel times* to successive incidents for any given service unit are statistically independent. Also assume steady-state conditions.

   Let the criterion for comparison between the k helicopter system and the m police car system be that fraction of time that a randomly selected service unit is busy (a server is considered busy if it is either traveling to the scene of an incident or servicing a call at the scene). Find the value of m for which the car-based system will most closely match the helicopter system for any given number k of helicopters.

   Your answer should be an expression for m containing only variables defined above and constants. Please explain your work.

b. For $R = 4$, $v_H = 80$ mph, $v_c = 25$ mph, and $\mu_H = \mu_c = 4$ calls per hour, compute the ratio m/k for your answer in part (a).

c. In part (a), we assume that travel times to successive incidents for any given service unit are statistically independent.

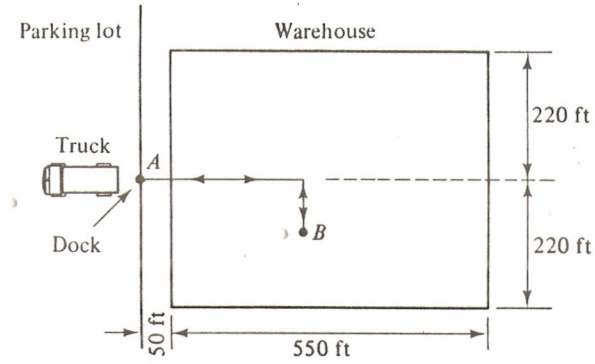
   Is this assumption a correct one? Please explain briefly and clearly in intuitive terms.

d. Assume now that there are two helicopters (and no police cars in the district). Also assume that $R/v_H \ll 1/\mu_H$. Find a good approximation for the probability that, with the system in the steady state, an observer arriving at a random instant will find both helicopters busy and exactly one call waiting in the queue. For what condition is your answer true?

e. Under the assumptions of part (d), find the probability that with the system in the steady state, the two helicopters will complete service to exactly l calls during a time interval T.

4.15 *Unloading at a warehouse* A forklift transfers cargo from an intercity truck to a warehouse with the following operational characteristics (see Figure P4.15):

1. The time spent on loading the forklift (at the truck location, A) is approximately constant and equal to 15 seconds.
2. All cargo from each intercity truck is transferred to the same designated storage location (B) in the warehouse.



3. For a given truck, any storage location in the warehouse is equally likely.
4. The average unloading time at B, including turning and reversing, is 10 seconds (negligible variance).
5. The forklift travels at a constant speed of 5 feet per second, and, because of the warehouse layout, in the east-west and north-south directions only.
6. Exactly 10 forklift loads are required to empty each truck.
7. Trucks are served one at a time and leave the dock as soon as they are emptied; the docking/undocking times for each truck last considerably less than a roundtrip by a forklift, so that no forklift time is wasted because of truck maneuvers.

In the following we shall refer to the truck parking lot and the dock as "the system."

a. Given that one truck arrives at the system every 50 minutes on the average and that the arrivals follow a Poisson process, find the average number of trucks in the system at any given time and the average time each truck remains in the system.

b. If two forklifts are used instead of one, answer approximately the questions in part (a) with an appropriately defined $M/G/1$ model. Assume that the forklifts do not interfere with one another throughout the operation.

   Hint: What happens to the mean and variance of the service time?

c. Use approximate bounds for $M/G/2$ systems (Section 5.8) to answer the questions in part (b) and compare these answers with your answers in (b).

4.16 *$M/H_k/m$ queueing systems* The hyperexponential pdf of order k is the pdf

$$f_X(x) = \sum_{i=1}^{k} \alpha_i \lambda_i e^{-\lambda_i x} \qquad x \geq 0,\ \alpha_i \geq 0,\ \lambda_i \geq 0$$

where

$$\sum_{i=1}^{k} \alpha_i = 1$$

In other words, a hyperexponential pdf can be viewed as the *weighted sum* of k distinct negative exponential pdf's.

a. Show that, for the random variable X with hyperexponential pdf,

$$E[X] = \sum_{i=1}^{k} \frac{\alpha_i}{\lambda_i} \quad \text{and} \quad E[X^2] = 2\sum_{i=1}^{k} \frac{\alpha_i}{\lambda_i^2}$$

b. Show that the coefficient of variation of X,

$$C_X^2 \triangleq \frac{\sigma_X^2}{E^2[X]} = \frac{E[X^2]}{E^2[X]} - 1 \geq 1$$

Hint: Use the Cauchy–Schwarz inequality,

$$(\sum_i a_i b_i)^2 \leq (\sum_i a_i^2)(\sum_i b_i^2) \qquad \text{for } a_i, b_i \text{ real.}$$

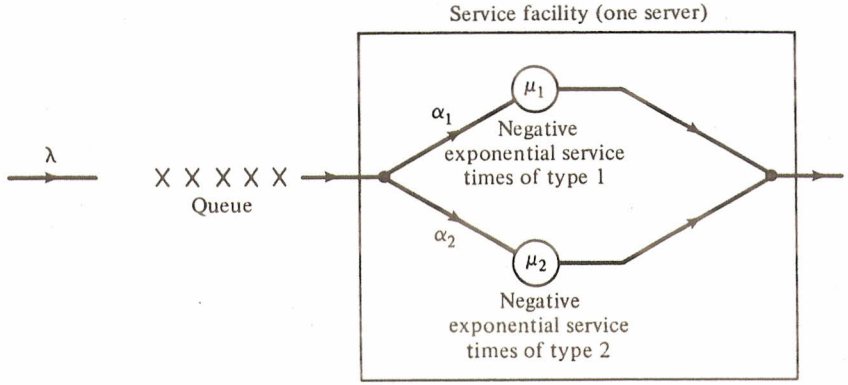
Because of the fact that $C_x^2 \geq 1$, hyperexponential random variables are said to be "more random" than negative exponential random variables (for which $C^2 = 1$).

c. Consider now a $M/H_2/1$ queueing system with infinite queue capacity

(H_2 indicates that service times are second-order hyperexponential random variables.) Let $\lambda = 3$ be the arrival rate at the system and let the service time pdf be given by

$$f_S(t) = \frac{2}{3}e^{-2t} + \frac{16}{3}e^{-8t} \qquad t \geq 0$$

A schematic representation of this system is shown in Figure P4.16. Each user, upon entrance to the service facility, will receive type 1 service with probability α_1 *or* type 2 service with probability α_2. Whenever *either one* of the two types of services is being offered, no other user can obtain access to the facility. What are the values of α_1, α_2, μ_1, and μ_2 in this case?



d. Find $\bar{L}$, $\bar{W}$, $\bar{L}_q$, and $\bar{W}_q$ in this case. Compare with the equivalent quantities for a $M/M/1$ system with service rate (when busy) equal to the service rate of this $M/H_2/1$ facility.

e. Carefully draw a state-transition diagram for this $M/H_2/1$ system.

Hint 1: Define states: "0" = the system is empty; "i, j" = i users are present ($i = 1, 2, \ldots$) *and* the user currently occupying the service facility is receiving type j service ($j = 1, 2$).

Hint 2: The rate of transitions from state 0 to state (1, 1) is equal to $\alpha_1\lambda$; the rate of transitions from state (2, 2) to state (1, 1) is equal to $\alpha_1\mu_2$.

f. Describe a possible situation in an urban service system context where $M/H_k/1$ (or $M/H_k/m$) models could be applicable.

4.17 *Probability that a server is busy* We have shown in the text that for a $M/M/1$ and a $M/G/1$ queueing system with infinite queueing capacity, the probability that the server is busy is equal to the utilization ρ ($= \lambda/\mu$) in steady state.

Show that in the case of a $M/M/m$ system with infinite queueing capacity, the steady-state probability of any server being busy is still equal to ρ ($= \lambda/m\mu$).