3

Birth–Death Queueing Systems in Equilibrium

In the previous chapter we studied a variety of stochastic processes. We indicated that Markov processes play a fundamental role in the study of queueing systems, and after presenting the main results from that theory, we then considered a special form of Markov process known as the birth–death process. We also showed that birth–death processes enjoy a most convenient property, namely, that the time between births and the time between deaths (when the system is nonempty) are each exponentially distributed.* We then developed Eq. (2.127), which gives the basic equations of motion for the general birth–death process with stationary birth and death rates.† The solution of this set of equations gives the transient behavior of the queueing process and some important special cases were discussed earlier. In this chapter we study the limiting form of these equations to obtain the equilibrium behavior of birth–death queueing systems.

The importance of elementary queueing theory comes from its historical influence as well as its ability to describe behavior that is to be found in more complex queueing systems. The methods of analysis to be used in this chapter in large part do *not* carry over to the more involved queueing situations; nevertheless, the obtained results *do* provide insight into the basic behavior of many of these other queueing systems.

It is necessary to keep in mind how the birth–death process describes queueing systems. As an example, consider a doctor's office made up of a waiting room (in which a queue is allowed to form, unfortunately) and a service facility consisting of the doctor's examination room. Each time a patient enters the waiting room from outside the office we consider this to be an *arrival* to the queueing system; on the other hand, this arrival may well be considered to be a *birth* of a new member of a population, where the population consists of all patients present. In a similar fashion, when a patient leaves

* This comes directly from the fact that they are Markov processes.

† In addition to these equations, one requires the conservation relation given in Eq. (2.122) and a set of initial conditions $\{P_k(0)\}$.

the office after being treated, he is considered to be a *departure* from the queueing system; in terms of a birth–death process this is considered to be a *death* of a member of the population.

We have considerable freedom in constructing a large number of queueing systems through the choice of the birth coefficients λ_k and death coefficients μ_k, as we shall see shortly. First, let us establish the general solution for the equilibrium behavior.

3.1. GENERAL EQUILIBRIUM SOLUTION

As we saw in Chapter 2 the time-dependent solution of the birth–death system quickly becomes unmanageable when we consider any sophisticated set of birth–death coefficients. Furthermore, were we always capable of solving for $P_k(t)$ it is not clear how useful that set of functions would be in aiding our understanding of the behavior of these queueing systems (too much information is sometimes a curse!). Consequently, it is natural for us to ask whether the probabilities $P_k(t)$ eventually settle down as t gets large and display no more "transient" behavior. This inquiry on our part is analogous to the questions we asked regarding the existence of π_k in the limit of $\pi_k(t)$ as $t \to \infty$. For our queueing studies here we choose to denote the limiting probability as p_k rather than π_k, purely for convenience. Accordingly, let

$$p_k \overset{\Delta}{=} \lim_{t\to\infty} P_k(t) \tag{3.1}$$

where p_k is interpreted as the limiting probability that the system contains k members (or equivalently is in state E_k) at some arbitrary time in the distant future. The question regarding the existence of these limiting probabilities is of concern to us, but will be deferred at this point until we obtain the general steady-state or limiting solution. It is important to understand that whereas p_k (assuming it exists) is no longer a function of t, we are not claiming that the process does not move from state to state in this limiting case; certainly, the number of members in the population will change with time, but the *long-run* probability of finding the system with k members will be properly described by p_k.

Accepting the existence of the limit in Eq. (3.1), we may then set lim $dP_k(t)/dt$ as $t \to \infty$ equal to zero in the Kolmogorov forward equations (of motion) for the birth–death system [given in Eqs. (2.127)] and immediately obtain the result

$$0 = -(\lambda_k + \mu_k)p_k + \lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k+1} \qquad k \geq 1 \tag{3.2}$$

$$0 = -\lambda_0 p_0 + \mu_1 p_1 \qquad k = 0 \tag{3.3}$$

The annoying task of providing a separate equation for $k = 0$ may be overcome by agreeing once and for all that the following birth and death coefficients are identically equal to 0:

$$\lambda_{-1} = \lambda_{-2} = \lambda_{-3} = \cdots = 0$$

$$\mu_0 = \mu_{-1} = \mu_{-2} = \cdots = 0$$

Furthermore, since it is perfectly clear that we cannot have a negative number of members in our population, we will, in most cases, adopt the convention that

$$p_{-1} = p_{-2} = p_{-3} = \cdots = 0$$

Thus, for all values of k, we may reformulate Eqs. (3.2) and (3.3) into the following set of difference equations for $k = \ldots, -2, -1, 0, 1, 2, \ldots$

$$0 = -(\lambda_k + \mu_k)p_k + \lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k+1} \tag{3.4}$$

We also require the conservation relation

$$\sum_{k=0}^{\infty} p_k = 1 \tag{3.5}$$

Recall from the previous chapter that the limit given in the Eq. (3.1) is independent of the initial conditions.

Just as we used the state-transition-rate diagram as an inspection technique for writing down the equations of motion in Chapter 2, so may we use the same concept in writing down the *equilibrium* equations [Eqs. (3.2) and (3.3)] directly from that diagram. In this equilibrium case it is clear that flow must be *conserved* in the sense that the input flow must equal the output flow from a given state. For example, if we look at Figure 2.9 once again and concentrate on state E_k in equilibrium, we observe that

$$\text{Flow rate into } E_k = \lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k+1}$$

and

$$\text{Flow rate out of } E_k = (\lambda_k + \mu_k)p_k$$

In equilibrium these two must be the same and so we have immediately

$$\lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k+1} = (\lambda_k + \mu_k)p_k \tag{3.6}$$

But this last is just Eq. (3.4) again! *By inspection we have established the equilibrium difference equations for our system.* The same comments apply here as applied earlier regarding the conservation of flow across *any* closed boundary; for example, rather than surrounding each state and writing down its equation we could choose a sequence of boundaries the first of which surrounds E_0, the second of which surrounds E_0 and E_1, and so on, each time adding the next higher-numbered state to get a new boundary. In such an example the kth boundary (which surrounds states $E_0, E_1, \ldots, E_{k-1}$) would lead to the following simple conservation of flow relationship:

$$\lambda_{k-1}p_{k-1} = \mu_k p_k \tag{3.7}$$

This last set of equations is equivalent to drawing a vertical line separating adjacent states and equating flows across this boundary; this set of difference equations is equivalent to our earlier set.

The solution for p_k in Eq. (3.4) may be obtained by at least two methods. One way is first to solve for p_1 in terms of p_0 by considering the case $k = 0$, that is,

$$p_1 = \frac{\lambda_0}{\mu_1} p_0 \tag{3.8}$$

We may then consider Eq. (3.4) for the case $k = 1$ and using Eq. (3.8) obtain

$$0 = -(\lambda_1 + \mu_1)p_1 + \lambda_0 p_0 + \mu_2 p_2$$

$$0 = -(\lambda_1 + \mu_1)\frac{\lambda_0}{\mu_1} p_0 + \lambda_0 p_0 + \mu_2 p_2$$

$$0 = -\frac{\lambda_1\lambda_0}{\mu_1} p_0 - \lambda_0 p_0 + \lambda_0 p_0 + \mu_2 p_2$$

and so

$$p_2 = \frac{\lambda_0\lambda_1}{\mu_1\mu_2} p_0 \tag{3.9}$$

If we examine Eqs. (3.8) and (3.9) we may justifiably guess that the general solution to Eq. (3.4) must be

$$p_k = \frac{\lambda_0\lambda_1 \cdots \lambda_{k-1}}{\mu_1\mu_2 \cdots \mu_k} p_0 \tag{3.10}$$

To validate this assertion we need merely use the inductive argument and apply Eq. (3.10) to Eq. (3.4) solving for p_{k+1}. Carrying out this operation we do, in fact, find that (3.10) is the solution to the general birth–death process in this steady-state or limiting case. We have thus expressed all equilibrium probabilities p_k in terms of a single unknown constant p_0:

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \qquad k = 0, 1, 2, \ldots \qquad ■(3.11)$$

(Recall the usual convention that an empty product is unity by definition.) Equation (3.5) provides the additional condition that allows us to determine p_0; thus, summing over all k, we obtain

$$p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} \qquad ■(3.12)$$

This "product" sol
principal equation in
departure for all of o
A second easy wa
equation as follows:

Defining

we have from Eq. (3.

Clearly Eq. (3.15) im

However, since λ_{-1} =

and so the constant i
iately obtain from Ec

Solving Eq. (3.17) su
solution, namely, Eq
We now address o
p_k given by Eqs. (3.1
sions to represent a p
This clearly places a
equations. Essentially
empties; that this is a
interprets it in terms
the possibilities by fi

* It is easy to construct
ments which follow.

This "product" solution for p_k $(k = 0, 1, 2, \ldots)$ simply obtained, is a *principal* equation in elementary queueing theory and, in fact, is the point of departure for all of our further solutions in this chapter.

A second easy way to obtain the solution to Eq. (3.4) is to rewrite that equation as follows:

$$\lambda_{k-1}p_{k-1} - \mu_k p_k = \lambda_k p_k - \mu_{k+1}p_{k+1} \tag{3.13}$$

Defining

$$g_k = \lambda_k p_k - \mu_{k+1}p_{k+1} \tag{3.14}$$

we have from Eq. (3.13) that

$$g_{k-1} = g_k \tag{3.15}$$

Clearly Eq. (3.15) implies that

$$g_k = \text{constant with respect to } k \tag{3.16}$$

However, since $\lambda_{-1} = \mu_0 = 0$, Eq. (3.14) gives

$$g_{-1} = 0$$

and so the constant in Eq. (3.16) must be 0. Setting g_k equal to 0, we immediately obtain from Eq. (3.14)

$$p_{k+1} = \frac{\lambda_k}{\mu_{k+1}}\, p_k \tag{3.17}$$

Solving Eq. (3.17) successively beginning with $k = 0$ we obtain the earlier solution, namely, Eqs. (3.11) and (3.12).

We now address ourselves to the *existence* of the steady-state probabilities p_k given by Eqs. (3.11) and (3.12). Simply stated, in order for those expressions to represent a probability distribution, we usually require that $p_0 > 0$. This clearly places a condition upon the birth and death coefficients in those equations. Essentially, what we are requiring is that the system occasionally empties; that this is a condition for stability seems quite reasonable when one interprets it in terms of real life situations.* More precisely, we may classify the possibilities by first defining the two sums

$$S_1 \triangleq \sum_{k=0}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \tag{3.18}$$

$$S_2 \triangleq \sum_{k=0}^{\infty} \left(1 \Bigg/ \left(\lambda_k \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) \right) \tag{3.19}$$

* It is easy to construct counterexamples to this case, and so we require the precise arguments which follow.

All states E_k of our birth–death process will be *ergodic* if and only if

$$\text{Ergodic:}\quad \begin{aligned} S_1 &< \infty \\ S_2 &= \infty \end{aligned}$$

On the other hand, all states will be *recurrent null* if and only if

$$\text{Recurrent null:}\quad \begin{aligned} S_1 &= \infty \\ S_2 &= \infty \end{aligned}$$

Also, all states will be *transient* if and only if

$$\text{Transient:}\quad \begin{aligned} S_1 &= \infty \\ S_2 &< \infty \end{aligned}$$

It is the ergodic case that gives rise to the equilibrium probabilities $\{p_k\}$ and that is of most interest to our studies. We note that the condition for ergodicity is met whenever the sequence $\{\lambda_k/\mu_k\}$ remains below unity from some k onwards, that is, if there exists some k_0 such that for all $k \geq k_0$ we have

$$\frac{\lambda_k}{\mu_k} < 1 \tag{3.20}$$

We will find this to be true in most of the queueing systems we study.

We are now ready to apply our general solution as given in Eqs. (3.11) and (3.12) to some very important special cases. Before we launch headlong into that discussion, let us put at ease those readers who feel that the birth–death constraints of permitting only nearest-neighbor transitions are too confining. It is true that the solution given in Eqs. (3.11) and (3.12) applies only to nearest-neighbor birth–death processes. However, rest assured that the equilibrium methods we have described can be extended to more general than nearest-neighbor systems; these generalizations are considered in Chapter 4.

3.2. M/M/1: THE CLASSICAL QUEUEING SYSTEM

As mentioned in Chapter 2, the celebrated M/M/1 queue is the simplest nontrivial interesting system and may be described by selecting the birth–death coefficients as follows:

$$\begin{aligned} \lambda_k &= \lambda \qquad k = 0, 1, 2, \ldots \\ \mu_k &= \mu \qquad k = 1, 2, 3, \ldots \end{aligned}$$

That is, we set all birth* coefficients equal to a constant λ and all death*

* In this case, the average interarrival time is $\bar{t} = 1/\lambda$ and the average service time is $\bar{x} = 1/\mu$; this follows since both $\tilde{t}$ and $\tilde{x}$ are both exponentially distributed.

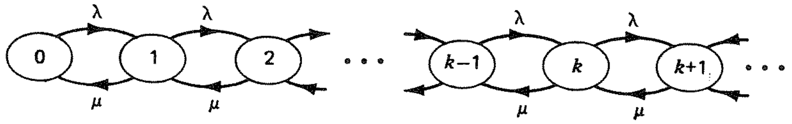


Figure 3.1 State-transition-rate diagram for M/M/1.

coefficients equal to a constant μ. We further assume that infinite queueing space is provided and that customers are served in a first-come-first-served fashion (although this last is not necessary for many of our results). For this important example the state-transition-rate diagram is as given in Figure 3.1.

Applying these coefficients to Eq. (3.11) we have

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu}$$

or

$$p_k = p_0\left(\frac{\lambda}{\mu}\right)^k \qquad k \geq 0 \tag{3.21}$$

The result is immediate. The conditions for our system to be ergodic (and, therefore, to have an equilibrium solution $p_k > 0$) are that $S_1 < \infty$ and $S_2 = \infty$; in this case the first condition becomes

$$S_1 = \sum_{k=0}^{\infty} \frac{p_k}{p_0} = \sum_{k=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^k < \infty$$

The series on the left-hand side of the inequality will converge if and only if $\lambda/\mu < 1$. The second condition for ergodicity becomes

$$S_2 = \sum_{k=0}^{\infty} \frac{1}{\lambda(p_k/p_0)} = \sum_{k=0}^{\infty} \frac{1}{\lambda}\left(\frac{\mu}{\lambda}\right)^k = \infty$$

This last condition will be satisfied if $\lambda/\mu \leq 1$; thus the necessary and sufficient condition for ergodicity in the M/M/1 queue is simply $\lambda < \mu$. In order to solve for p_0 we use Eq. (3.12) [or Eq. (3.5) as suits the reader] and obtain

$$p_0 = 1\Big/\left[1 + \sum_{k=1}^{\infty}\left(\frac{\lambda}{\mu}\right)^k\right]$$

The sum converges since $\lambda < \mu$ and so

$$p_0 = \cfrac{1}{1 + \cfrac{\lambda/\mu}{1 - \lambda/\mu}}$$

Thus

$$p_0 = 1 - \frac{\lambda}{\mu} \tag{3.22}$$

From Eq. (2.29) we have $\rho = \lambda/\mu$. From our stability conditions, we therefore require that $0 \leq \rho < 1$; note that this insures that $p_0 > 0$. From Eq. (3.21) we have, finally,

$$p_k = (1 - \rho)\rho^k \qquad k = 0, 1, 2, \ldots \qquad \blacksquare\,(3.23)$$

Equation (3.23) is indeed the solution for the steady-state probability of finding k customers in the system.* We make the important observation that p_k depends upon λ and μ only through their ratio ρ.

The solution given by Eq. (3.23) for this fundamental system is graphed in Figure 3.2 for the case of $\rho = 1/2$. Clearly, this is the geometric distribution (which shares the fundamental memoryless property with the exponential distribution). As we develop the behavior of the M/M/1 queue, we shall continue to see that almost all of its important probability distributions are of the memoryless type.

An important measure of a queueing system is the average number of customers in the system $\bar{N}$. This is clearly given by

$$\begin{aligned} \bar{N} &= \sum_{k=0}^{\infty} k p_k \\ &= (1 - \rho) \sum_{k=0}^{\infty} k \rho^k \end{aligned}$$

Using the trick similar to the one used in deriving Eq. (2.142) we have

$$\begin{aligned} \bar{N} &= (1 - \rho)\rho \frac{\partial}{\partial \rho} \sum_{k=0}^{\infty} \rho^k \\ &= (1 - \rho)\rho \frac{\partial}{\partial \rho} \frac{1}{1 - \rho} \end{aligned}$$

$$\bar{N} = \frac{\rho}{1 - \rho} \qquad \blacksquare\,(3.24)$$

* If we inspect the transient solution for M/M/1 given in Eq. (2.163), we see the term $(1 - \rho)\rho^k$; the reader may verify that, for $\rho < 1$, the limit of the transient solution agrees with our solution here.

(3.22)

ility conditions, we there-
es that $p_0 > 0$. From Eq.

2, . . . ■(3.23)

teady-state probability of
mportant observation that
ρ.
ental system is graphed in
the geometric distribution
erty with the exponential
e M/M/1 queue, we shall
obability distributions are

is the average number of
y

g Eq. (2.142) we have

–
ρ

■(3.24)

n Eq. (2.163), we see the term
t of the transient solution agrees

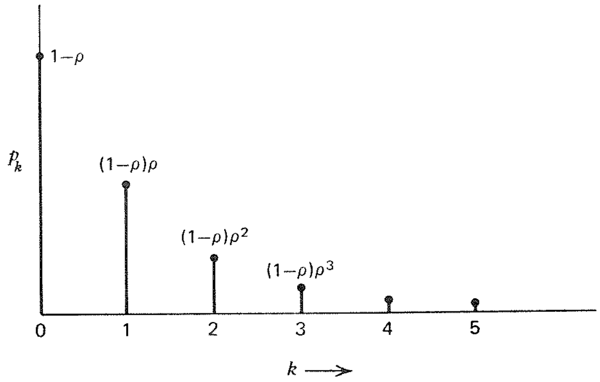


Figure 3.2 The solution for p_k in the system M/M/1.

The behavior of the expected number in the system is plotted in Figure 3.3. By similar methods we find that the variance of the number in the system is given by

$$\sigma_N^2 = \sum_{k=0}^{\infty}(k - \bar{N})^2 p_k$$

$$\sigma_N^2 = \frac{\rho}{(1-\rho)^2} \qquad ■(3.25)$$

We may now apply Little's result directly from Eq. (2.25) in order to obtain

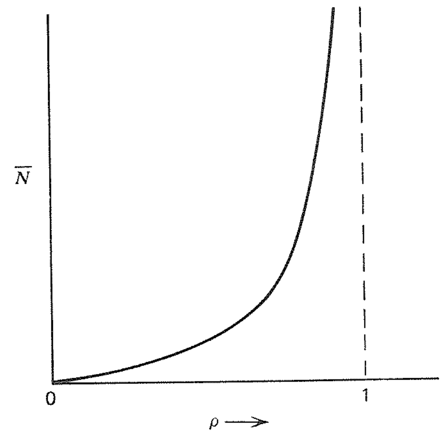


Figure 3.3 The average number in the system M/M/1.

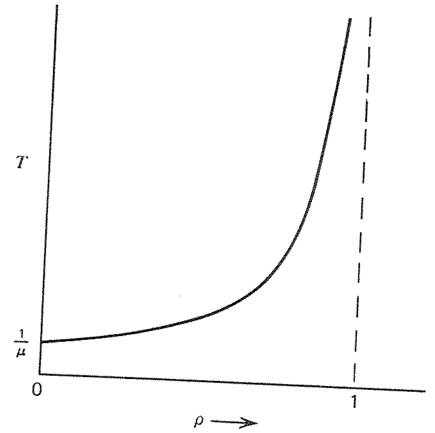


Figure 3.4 Average delay as a function of ρ for M/M/1.

T, the average *time* spent in the system as follows:

$$T = \frac{\bar{N}}{\lambda}$$

$$T = \left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{\lambda}\right)$$

$$T = \frac{1/\mu}{1-\rho} \qquad ■ (3.26)$$

This dependence of average time on the utilization factor ρ is shown in Figure 3.4. The value obtained by T when $\rho = 0$ is exactly the average service time expected by a customer; that is, he spends no time in queue and $1/\mu$ sec in service on the average.

The behavior given by Eqs. (3.24) and (3.26) is rather dramatic. As ρ approaches unity, both the average number in the system and the average time in the system grow in an unbounded fashion.* Both these quantities have a

* We observe at $\rho = 1$ that the system behavior is unstable; this is not surprising if one recalls that $\rho < 1$ was our condition for ergodicity. What is perhaps surprising is that the behavior of the average number $\bar{N}$ and of the average system time T deteriorates so badly as $\rho \to 1$ from below; we had seen for steady flow systems in Chapter 1 that so long as $R < C$ (which corresponds to the case $\rho < 1$) no queue formed and smooth, rapid flow proceeded through the system. Here in the M/M/1 queue we find this is no longer true and that we pay an extreme penalty when we attempt to run the system near (but below) its capacity. The

simple pole at $\rho = 1$. *This*
1 is characteristic of almost
see it again in M/G/1 in C
of G/G/1 (and also in the
Chapter 2.

Another interesting qua
least k customers in the sy

$P[\geq$

$P[\geq$

Thus we see that the proba
customers in the system is a
and decays very rapidly.

With the tools at hand w
density function for the ti
development until we treat
[see Eq. (5.118)]. Meanwhi
death queues in equilibrium

3.3. DISCOURAGED A

This next example consid
when more and more peopl
model this effect is to choos

$\lambda_k =$

$\mu_k =$

We are here assuming an har
the number present in the s

intuitive explanation here is that w
traffic which temporarily overwhe
idle on the average $1 - \rho = p_0$ c
uniformly within small time inter
hand, in the steady flow case (whi
will be distributed quite uniforml
secs) there will be an idle time of e
the interarrival time and in the ser
$\rho = 1$; any reduction in the variat
the average waiting time, as we sh

simple pole at $\rho = 1$. *This type of behavior with respect to ρ as ρ approaches 1 is characteristic of almost every queueing system one can encounter.* We will see it again in M/G/1 in Chapter 5 as well as in the heavy traffic behavior of G/G/1 (and also in the tight bounds on G/G/1 behavior) in Volume II, Chapter 2.

Another interesting quantity to calculate is the probability of finding at least k customers in the system:

$$P[\geq k \text{ in system}] = \sum_{i=k}^{\infty} p_i$$

$$= \sum_{i=k}^{\infty} (1 - \rho)\rho^i$$

$$P[\geq k \text{ in system}] = \rho^k \qquad ■ (3.27)$$

Thus we see that the probability of exceeding some limit on the number of customers in the system is a geometrically decreasing function of that number and decays very rapidly.

With the tools at hand we are now in a position to develop the probability density function for the time spent in the system. However, we defer that development until we treat the more general case of M/G/1 in Chapter 5 [see Eq. (5.118)]. Meanwhile, we proceed to discuss numerous other birth–death queues in equilibrium.

3.3. DISCOURAGED ARRIVALS

This next example considers a case where arrivals tend to get discouraged when more and more people are present in the system. One possible way to model this effect is to choose the birth and death coefficients as follows:

$$\lambda_k = \frac{\alpha}{k+1} \qquad k = 0, 1, 2, \ldots$$

$$\mu_k = \mu \qquad k = 1, 2, 3, \ldots$$

We are here assuming an harmonic discouragement of arrivals with respect to the number present in the system. The state-transition-rate diagram in this

intuitive explanation here is that with random flow (e.g., M/M/1) we get occasional bursts of traffic which temporarily overwhelm the server; while it is still true that the server will be idle on the average $1 - \rho = p_0$ of the time this average idle time will not be distributed uniformly within small time intervals but will only be true in the long run. On the other hand, in the steady flow case (which corresponds to our system D/D/1) the system idle time will be distributed quite uniformly in the sense that after every service time (of exactly $1/\mu$ secs) there will be an idle time of exactly $(1/\lambda) - (1/\mu)$ sec. Thus it is the *variability* in both the interarrival time and in the service time which gives rise to the disastrous behavior near $\rho = 1$; any reduction in the variation of either random variable will lead to a reduction in the average waiting time, as we shall see again and again.

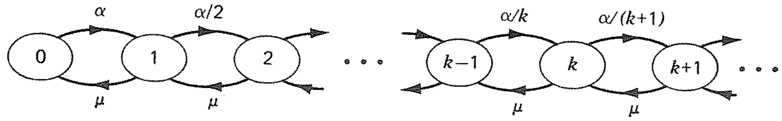


Figure 3.5 State-transition-rate diagram for discouraged arrivals.

case is as shown in Figure 3.5. We apply Eq. (3.11) immediately to obtain for p_k

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\alpha/(i+1)}{\mu} \qquad (3.28)$$

$$p_k = p_0 \left(\frac{\alpha}{\mu}\right)^k \frac{1}{k!} \qquad (3.29)$$

Solving for p_0 from Eq. (3.12) we have

$$p_0 = 1 \Big/ \left[1 + \sum_{k=1}^{\infty} \left(\frac{\alpha}{\mu}\right)^k \frac{1}{k!}\right]$$

$$p_0 = e^{-\alpha/\mu}$$

From Eq. (2.32) we have therefore,

$$\rho = 1 - e^{-\alpha/\mu} \qquad (3.30)$$

Note that the ergodic condition here is merely $\alpha/\mu < \infty$. Going back to Eq. (3.29) we have the final solution

$$p_k = \frac{(\alpha/\mu)^k}{k!} e^{-\alpha/\mu} \qquad k = 0, 1, 2, \ldots \qquad (3.31)$$

We thus have a Poisson distribution for the number of customers in the system of discouraged arrivals! From Eqs. (2.131) and (2.132) we have that the expected number in the system is

$$\bar{N} = \frac{\alpha}{\mu}$$

In order to calculate T, the average time spent in the system, we may use Little's result again. For this we require λ, which is directly calculated from $\rho = \lambda\bar{x} = \lambda/\mu$; thus from Eq. (3.30)

$$\lambda = \mu\rho = \mu(1 - e^{-\alpha/\mu})$$

Using this* and Little's result we then obtain

$$T = \frac{\alpha}{\mu^2(1 - e^{-\alpha/\mu})} \qquad (3.32)$$

[*] Note that this result could have been obtained from $\lambda = \sum_k \lambda_k p_k$. The reader should verify this last calculation.

3.4. M/M/∞: RESPONS OF SERVERS)

Here we consider the ca responsive server who ac customers are waiting or always a new clerk or serve ular, we set

$$\lambda_k =$$

$$\mu_k =$$

Here the state-transition-ra directly to Eq. (3.11) for th

Need we go any further? (3.28). These two are in fa solutions for p_k and $\bar{N}$,

$$p_k = \frac{(\lambda}{}$$

$$\bar{N} = \frac{\lambda}{\mu}$$

Here, too, the ergodic cond system of discouraged arri includes a responsive serve (and simpler) form for T he

This answer is, of course, each arriving customer is gr be merely his service time v

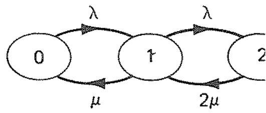
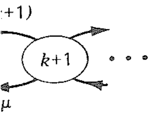


Figure 3.6 State-transitio



ırrivals.

ediately to obtain

(3.28)

(3.29)

(3.30)

∞. Going back to

(3.31)

ɔf customers in the
2.132) we have that

system, we may use
ctly calculated from

(3.32)

$\lambda_k p_k$. The reader should

3.4. M/M/∞: RESPONSIVE SERVERS (INFINITE NUMBER OF SERVERS)

Here we consider the case that may be interpreted either as that of a responsive server who accelerates her service rate linearly when more customers are waiting or may be interpreted as the case where there is always a new clerk or server available for each arriving customer. In particular, we set

$$\lambda_k = \lambda \qquad k = 0, 1, 2, \ldots$$
$$\mu_k = k\mu \qquad k = 1, 2, 3, \ldots$$

Here the state-transition-rate diagram is that shown in Figure 3.6. Going directly to Eq. (3.11) for the solution we obtain

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} \tag{3.33}$$

Need we go any further? The reader should compare Eq. (3.33) with Eq. (3.28). These two are in fact equivalent, and so we immediately have the solutions for p_k and $\bar{N}$,

$$p_k = \frac{(\lambda/\mu)^k}{k!} e^{-\lambda/\mu} \qquad k = 0, 1, 2, \ldots \tag{3.34}$$

$$\bar{N} = \frac{\lambda}{\mu}$$

Here, too, the ergodic condition is simply $\lambda/\mu < \infty$. It appears then that a system of discouraged arrivals behaves exactly the same as a system that includes a responsive server. However, Little's result provides a different (and simpler) form for T here than that given in Eq. (3.32); thus

$$T = \frac{1}{\mu}$$

This answer is, of course, obvious since if we use the interpretation where each arriving customer is granted his own server, then his time in system will be merely his service time which clearly equals $1/\mu$ on the average.

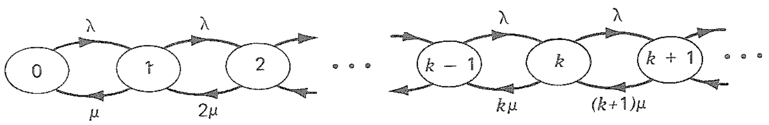


Figure 3.6 State-transition-rate diagram for the infinite-server case M/M/∞.

3.5. M/M/m: THE *m*-SERVER CASE

Here again we consider a system with an unlimited waiting room and with a constant arrival rate λ. The system provides for a *maximum* of m servers. This is within the reach of our birth–death formulation and leads to

$$\lambda_k = \lambda \qquad k = 0, 1, 2, \ldots$$

$$\begin{aligned}\mu_k &= \min\,[k\mu, m\mu] \\ &= \begin{cases} k\mu & 0 \le k \le m \\ m\mu & m \le k \end{cases}\end{aligned}$$

From Eq. (3.20) it is easily seen that the condition for ergodicity is $\lambda/m\mu < 1$. The state-transition-rate diagram is shown in Figure 3.7. When we go to solve for p_k from Eq. (3.11) we find that we must separate the solution into two parts, since the dependence of μ_k upon k is also in two parts. Accordingly, for $k \le m$,

$$\begin{aligned} p_k &= p_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} \\ &= p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \end{aligned} \qquad (3.35)$$

Similarly, for $k \ge m$,

$$\begin{aligned} p_k &= p_0 \prod_{i=0}^{m-1} \frac{\lambda}{(i+1)\mu} \prod_{j=m}^{k-1} \frac{\lambda}{m\mu} \\ &= p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{m!\, m^{k-m}} \end{aligned} \qquad (3.36)$$

Collecting together the results from Eqs. (3.35) and (3.36) we have

$$p_k = \begin{cases} p_0 \dfrac{(m\rho)^k}{k!} & k \le m \\[2ex] p_0 \dfrac{(\rho)^k m^m}{m!} & k \ge m \end{cases} \qquad ■\,(3.37)$$

where

$$\rho = \frac{\lambda}{m\mu} < 1 \qquad (3.38)$$

This expression for ρ follows that in Eq. (2.30) and is consistent with our

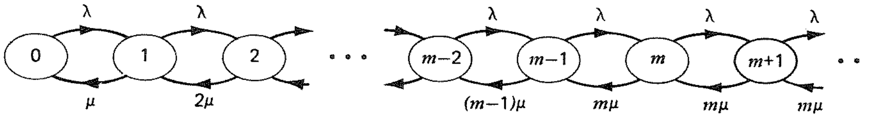


Figure 3.7 State-transition-rate diagram for M/M/m.

definition in terms of
solve for p_0 from Eq.

$p_0 =$

and so

$p_0 =$

The probability that
given by

Thus

$P[$queue

This probability is of
trunk (i.e., server) is
trunks; it is referred
$C(m, \lambda/\mu)$.

3.6. M/M/1/K: FI

We now consider f
there is a maximum
we assume the system
customer in service)
refused entry to the
Newly arriving cust
Poisson process but
customers will be a
considered to be "l
room at all) this is
single server.

* Europeans use the syn

definition in terms of the expected fraction of busy servers. We may now solve for p_0 from Eq. (3.12), which gives us

$$p_0 = \left[1 + \sum_{k=1}^{m-1} \frac{(m\rho)^k}{k!} + \sum_{k=m}^{\infty} \frac{(m\rho)^k}{m!} \frac{1}{m^{k-m}}\right]^{-1}$$

and so

$$p_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left(\frac{(m\rho)^m}{m!}\right)\left(\frac{1}{1-\rho}\right)\right]^{-1} \qquad \blacksquare\ (3.39)$$

The probability that an arriving customer is forced to join the queue is given by

$$P[\text{queueing}] = \sum_{k=m}^{\infty} p_k$$

$$= \sum_{k=m}^{\infty} p_0 \frac{(m\rho)^k}{m!} \frac{1}{m^{k-m}}$$

Thus

$$P[\text{queueing}] = \frac{\left(\frac{(m\rho)^m}{m!}\right)\left(\frac{1}{1-\rho}\right)}{\left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left(\frac{(m\rho)^m}{m!}\right)\left(\frac{1}{1-\rho}\right)\right]} \qquad \blacksquare\ (3.40)$$

This probability is of wide use in telephony and gives the probability that no trunk (i.e., server) is available for an arriving call (customer) in a system of m trunks; it is referred to as *Erlang's C formula* and is often denoted* by $C(m, \lambda/\mu)$.

3.6. M/M/1/K: FINITE STORAGE

We now consider for the first time the case of a queueing system in which there is a maximum number of customers that may be stored; in particular, we assume the system can hold at most a total of K customers (including the customer in service) and that any further arriving customers will in fact be refused entry to the system and will depart immediately without service. Newly arriving customers will continue to be generated according to a Poisson process but only those who find the system with strictly less than K customers will be allowed entry. In telephony the refused customers are considered to be "lost"; for the system in which $K = 1$ (i.e., no waiting room at all) this is referred to as a "blocked calls cleared" system with a single server.

* Europeans use the symbol $E_{2,m}(\lambda/\mu)$.

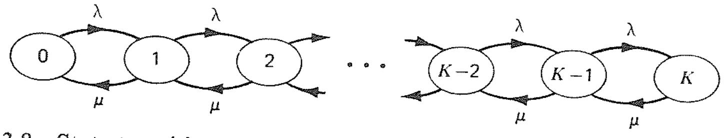


Figure 3.8 State-transition-rate diagram for the case of finite storage room M/M/1/K.

It is interesting that we are capable of accommodating this seemingly complex system description with our birth–death model. In particular, we accomplish this by effectively "turning off" the Poisson input as soon as the systems fills up, as follows:

$$\lambda_k = \begin{cases} \lambda & k < K \\ 0 & k \geq K \end{cases}$$

$$\mu_k = \mu \qquad k = 1, 2, \ldots, K$$

From Eq. (3.20), we see that this system is always ergodic. The state-transition-rate diagram for this finite Markov chain is shown in Figure 3.8. Proceeding directly with Eq. (3.11) we obtain

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu} \qquad k \leq K$$

or

$$p_k = p_0 \left(\frac{\lambda}{\mu}\right)^k \qquad k \leq K \tag{3.41}$$

Of course, we also have

$$p_k = 0 \qquad k > K \tag{3.42}$$

In order to solve for p_0 we use Eqs. (3.41) and (3.42) in Eq. (3.12) to obtain

$$p_0 = \left[1 + \sum_{k=1}^{K} \left(\frac{\lambda}{\mu}\right)^k\right]^{-1}$$

$$= \left[1 + \frac{(\lambda/\mu)(1 - (\lambda/\mu)^K)}{1 - \lambda/\mu}\right]^{-1}$$

and so

$$p_0 = \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}}$$

Thus, finally,

$$p_k = \begin{cases} \dfrac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}} \left(\dfrac{\lambda}{\mu}\right)^k & 0 \leq k \leq K \\ 0 & \text{otherwise} \end{cases} \qquad ■ (3.43)$$

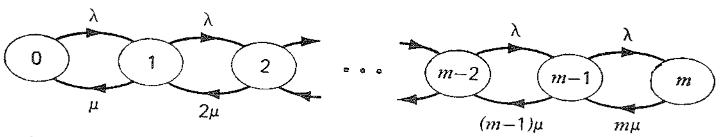


Figure 3.9 State-transition-rate diagram for *m*-server loss system M/M/m/m.

For the case of blocked calls cleared ($K = 1$) we have

$$p_k = \begin{cases} \dfrac{1}{1 + \lambda/\mu} & k = 0 \\ \dfrac{\lambda/\mu}{1 + \lambda/\mu} & k = 1 = K \\ 0 & \text{otherwise} \end{cases} \qquad (3.44)$$

3.7. M/M/m/m: *m*-SERVER LOSS SYSTEMS

Here we have again a blocked calls cleared situation in which there are available m servers. Each newly arriving customer is given his private server; however, if a customer arrives when all servers are occupied, that customer is lost. We create this artifact as above by choosing the following birth and death coefficients:

$$\lambda_k = \begin{cases} \lambda & k < m \\ 0 & k \geq m \end{cases}$$

$$\mu_k = k\mu \qquad k = 1, 2, \ldots, m$$

Here again, ergodicity is always assured. This finite state-transition-rate diagram is shown in Figure 3.9.

Applying Eq. (3.11) we obtain

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i + 1)\mu} \qquad k \leq m$$

or

$$p_k = \begin{cases} p_0 \left(\dfrac{\lambda}{\mu}\right)^k \dfrac{1}{k!} & k \leq m \\ 0 & k > m \end{cases} \qquad \blacksquare\ (3.45)$$

Solving for p_0 we have

$$p_0 = \left[\sum_{k=0}^{m} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}\right]^{-1} \qquad \blacksquare$$

This particular system is of great interest to those in telephony [so much so that a special case of Eq. (3.45) has been tabulated and graphed in many books on telephony]. Specifically, p_m describes the fraction of time that all m servers are busy. The name given to this probability expression is *Erlang's loss formula* and it is given by

$$p_m = \frac{(\lambda/\mu)^m/m!}{\sum_{k=0}^{m} (\lambda/\mu)^k/k!} \tag{3.46}$$

This equation is also referred to as *Erlang's B formula* and is commonly denoted* by $B(m, \lambda/\mu)$. Formula (3.46) was first derived by Erlang in 1917!

3.8. M/M/1//M†: FINITE CUSTOMER POPULATION—SINGLE SERVER

Here we consider the case where we no longer have a Poisson input process with an infinite user population, but rather have a *finite* population of possible users. The system structure is such that we have a total of M users; a customer is either in the system (consisting of a queue and a single server) or outside the system and in some sense "arriving." In particular, when a customer is in the "arriving" condition then the time it takes him to arrive is a random variable with an exponential distribution whose mean is $1/\lambda$ sec. All customers act independently of each other. As a result, when there are k customers in the system (queue plus service) then there are $M - k$ customers in the arriving state and, therefore, the total average arrival rate in this state is $\lambda(M - k)$. We see that this system is in a strong sense self-regulating. By this we mean that when the system gets busy, with many of these customers in the queue, then the rate at which additional customers arrive is in fact reduced, thus lowering the further congestion of the system. We model this quite appropriately with our birth–death process choosing for parameters

$$\lambda_k = \begin{cases} \lambda(M - k) & 0 \le k \le M \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_k = \mu \qquad k = 1, 2, \ldots$$

The system is ergodic. We assume that we have sufficient room to contain M customers in the system. The finite state-transition-rate diagram is shown in Figure 3.10. Using Eq. (3.11) we solve for p_k as follows:

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda(M - i)}{\mu} \qquad 0 \le k \le M$$

[*] Europeans use the notation $E_{1,m}(\lambda/\mu)$.

[†] Recall that a blank entry in either of the last two optional positions in this notation means an entry of ∞; thus here we have the system M/M/1/∞/M.

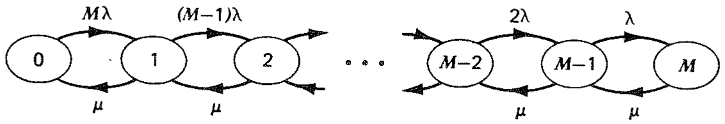


Figure 3.10 State-transition-rate diagram for single-server finite population system M/M/1//M.

Thus

$$p_k = \begin{cases} p_0\left(\dfrac{\lambda}{\mu}\right)^k \dfrac{M!}{(M-k)!} & 0 \le k \le M \\ 0 & k > M \end{cases} \tag*{■ (3.47)}$$

In addition, we obtain for p_0

$$p_0 = \left[\sum_{k=0}^{M}\left(\frac{\lambda}{\mu}\right)^k \frac{M!}{(M-k)!}\right]^{-1} \tag*{■ (3.48)}$$

3.9. M/M/∞//M: FINITE CUSTOMER POPULATION—"INFINITE" NUMBER OF SERVERS

We again consider the finite population case, but now provide a separate server for each customer in the system. We model this as follows:

$$\lambda_k = \begin{cases} \lambda(M-k) & 0 \le k \le M \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_k = k\mu \qquad k = 1, 2, \ldots$$

Clearly, this too is an ergodic system. The finite state-transition-rate diagram is shown in Figure 3.11. Solving this system, we have from Eq. (3.11)

$$\begin{aligned} p_k &= p_0 \prod_{i=0}^{k-1} \frac{\lambda(M-i)}{(i+1)\mu} \\ &= p_0\left(\frac{\lambda}{\mu}\right)^k \binom{M}{k} \qquad 0 \le k \le M \end{aligned} \tag{3.49}$$

where the binomial coefficient is defined in the usual way,

$$\binom{M}{k} \triangleq \frac{M!}{k!\,(M-k)!}$$

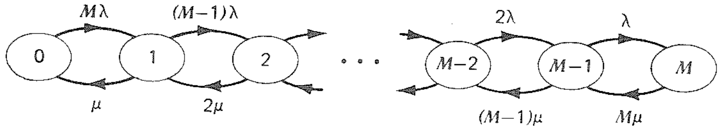


Figure 3.11 State-transition-rate diagram for "infinite"-server finite population system M/M/∞//M.

Solving for p_0 we have

$$p_0 = \left[\sum_{k=0}^{M} \left(\frac{\lambda}{\mu}\right)^k \binom{M}{k}\right]^{-1}$$

and so

$$p_0 = \frac{1}{(1 + \lambda/\mu)^M}$$

Thus

$$p_k = \begin{cases} \dfrac{\left(\dfrac{\lambda}{\mu}\right)^k \dbinom{M}{k}}{(1 + \lambda/\mu)^M} & 0 \le k \le M \\ 0 & \text{otherwise} \end{cases} \tag{3.50}$$

We may easily calculate the expected number of people in the system from

$$\begin{aligned} \bar{N} &= \sum_{k=0}^{M} k p_k \\ &= \frac{\displaystyle\sum_{k=0}^{M} k \left(\frac{\lambda}{\mu}\right)^k \binom{M}{k}}{(1 + \lambda/\mu)^M} \end{aligned}$$

Using the partial-differentiation trick such as for obtaining Eq. (3.24) we then have

$$\bar{N} = \frac{M\lambda/\mu}{1 + \lambda/\mu}$$

3.10. M/M/m/K/M: FINITE POPULATION, *m*-SERVER CASE, FINITE STORAGE

This rather general system is the most complicated we have so far considered and will reduce to all of the previous cases (except the example of discouraged arrivals) as we permit the parameters of this system to vary. We assume we have a finite population of M customers, each with an "arriving" parameter λ. In addition, the system has m servers, each with parameter μ. The system also has finite storage room such that the total number of customers in the system (queueing plus those in service) is no more than K. We assume $M \ge K \ge m$; customers arriving to find K already in the system are "lost" and return immediately to the arriving state as if they had just completed service. This leads to the following set of birth–death coefficients:

$$\lambda_k = \begin{cases} \lambda(M - k) & 0 \le k \le K - 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_k = \begin{cases} k\mu & 0 \le k \le m \\ m\mu & k \ge m \end{cases}$$

(3.50)

ystem from

(3.24) we

ASE,

far consid-
iple of dis-
vary. We
"arriving"
rameter μ.
r of custo-
e than K.
the system
y had just
oefficients:

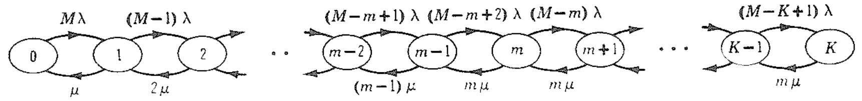


Figure 3.12 State-transition-rate diagram for m-server, finite storage, finite population system M/M/m/K/M.

In Figure 3.12 we see the most complicated of our finite state-transition-rate diagrams. In order to apply Eq. (3.11) we must consider two regions. First, for the range $0 \le k \le m-1$ we have

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda(M-i)}{(i+1)\mu}$$

$$= p_0 \left(\frac{\lambda}{\mu}\right)^k \binom{M}{k} \qquad 0 \le k \le m-1 \tag{3.51}$$

For the region $m \le k \le K$ we have

$$p_k = p_0 \prod_{i=0}^{m-1} \frac{\lambda(M-i)}{(i+1)\mu} \prod_{i=m}^{k-1} \frac{\lambda(M-i)}{m\mu}$$

$$= p_0 \left(\frac{\lambda}{\mu}\right)^k \binom{M}{k} \frac{k!}{m!} m^{m-k} \qquad m \le k \le K \tag{3.52}$$

The expression for p_0 is rather complex and will not be given here, although it may be computed in a straightforward manner. In the case of a pure loss system (i.e., $M \ge K = m$), the stationary state probabilities are given by

$$p_k = \frac{\binom{M}{k}\left(\frac{\lambda}{\mu}\right)^k}{\sum_{i=0}^{m} \binom{M}{i}\left(\frac{\lambda}{\mu}\right)^i} \qquad k = 0, 1, \ldots, m \tag{3.53}$$

This is known as the *Engset* distribution.

We could continue these examples ad nauseam but we will instead take a benevolent approach and terminate the set of examples here. Additional examples are given in the exercises. It should be clear to the reader by now that a large number of interesting queueing structures can be modeled with the birth–death process. In particular, we have demonstrated the ability to model the multiple-server case, the finite-population case, the finite-storage case and combinations thereof. The common element in all of these is that the solution for the equilibrium probabilities $\{p_k\}$ is given in Eqs. (3.11) and (3.12). Only systems whose solutions are given by these equations have been considered in this chapter. However, there are many other Markovian systems that lend themselves to simple solution and which are important in queueing theory. In the next chapter (4) we consider the equilibrium solution for Markovian queues; in Chapter 5 we will generalize to semi-Markov processes in which the service time distribution $B(x)$ is permitted to be general, and in Chapter 6 we revert back to the exponential service time case, but permit the interarrival time distribution $A(t)$ to be general; in both of these cases an imbedded Markov chain will be identified and solved. Only when both $A(t)$ and $B(x)$ are nonexponential do we require the methods of advanced queueing theory discussed in Chapter 8. (There are some special nonexponential distributions that may be described with the theory of Markov processes and these too are discussed in Chapter 4.)

EXERCISES

3.1. Consider a pure Markovian queueing system in which

$$\lambda_k = \begin{cases} \lambda & 0 \le k \le K \\ 2\lambda & K < k \end{cases}$$

$$\mu_k = \mu \qquad k = 1, 2, \ldots$$

(a) Find the equilibrium probabilities p_k for the number in the system.

(b) What relationship must exist among the parameters of the problem in order that the system be stable and, therefore, that this equilibrium solution in fact be reached? Interpret this answer in terms of the possible dynamics of the system.

3.2. Consider a Markovian queueing system in which

$$\lambda_k = \alpha^k \lambda \qquad k \ge 0, 0 \le \alpha < 1$$

$$\mu_k = \mu \qquad k \ge 1$$

(a) Find the equilibrium probability p_k of having k customers in the system. Express your answer in terms of p_0.

(b) Give an expression for p_0.

3.3. Consider an M/M/2 queueing system where the average arrival rate is λ customers per second and the average service time is $1/\mu$ sec, where $\lambda < 2\mu$.

(a) Find the differential equations that govern the time-dependent probabilities $P_k(t)$.

(b) Find the equilibrium probabilities

$$p_k = \lim_{t \to \infty} P_k(t)$$

(Continued on back flap)

3.4. Consider an M/M/1
are impatient. Spec
queueing time w an
leave with probabili
new arrival finds k i

(a) In terms of p_0
k in the system
parameters.

(b) For $0 < \alpha$, 0

solution hold?

(c) For $\alpha \to \infty$, fin
system.

3.5. Consider a birth–de
coefficients:

λ_k

μ_k

All other coefficient

(a) Solve for p_k. B
of λ, k, and μ

(b) Find the avera

3.6. Consider a birth–de

$\lambda_k = \alpha k($

$\mu_k = \beta k($

where $K_1 \le K_2$ and
$K_1 \le k \le K_2$. Solve
$K_1 \le k \le K_2$ custon

3.7. Consider an M/M/n
Poisson arrival strea
given by λ_i and ex
$1/\mu_i$ ($i = 1, 2$). The
arrival requires exac
then any newly arriv
second class each rec
occupy them all simu
amount of time who
finds less than m_0 id
the fraction of type
that are lost.