Chapter 8

Continuous-Parameter Markov Chains

8.1 INTRODUCTION

The analysis of continuous-parameter Markov chains is similar to that for the discrete-parameter case, except that the transitions from a given state to another state can take place at any instant of time. As in the last chapter, we confine our attention to **discrete-state** processes. This implies that, although the parameter t has a continuous range of values, the set of values of $X(t)$ is discrete. We let $I = \{0, 1, 2, \ldots\}$ denote the state space of the process and we let $T = [0, \infty)$ be its parameter space. As we recall from Chapter 6, a discrete-state continuous-parameter stochastic process $\{X(t),\ t \geqslant 0\}$ is called a Markov chain if for $t_0 < t_1 < t_2 \cdots < t_n < t$, with t and $t_r \geqslant 0$ $(r = 0, 1, \ldots, n)$, its conditional pmf satisfies the relation

$$\begin{aligned} P[X(t) = x | X(t_n) = x_n,\ X(t_{n-1}) = x_{n-1},\ \ldots,\ X(t_0) = x_0] \\ = P[X(t) = x |\ X(t_n) = x_n]. \end{aligned} \tag{8.1}$$

The behavior of the process is characterized by (1) the distribution of the initial state of the system given by the pmf of $X(t_0)$, $P[X(t_0) = k]$, $k = 0, 1, 2, \ldots$, and (2) the transition probabilities:

$$p_{ij}(v, t) = P[X(t) = j \mid X(v) = i] \tag{8.2}$$

for $0 \leqslant v \leqslant t$ and $i, j = 0, 1, 2, \ldots$, where we define:

$$p_{ij}(t, t) = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

The Markov chain $\{X(t), t \geqslant 0\}$ is said to be **(time)-homogeneous** (or is said to have **stationary transition probabilities**) if $p_{ij}(v, t)$ depends only on the time difference $(t - v)$. In this case, we abbreviate the notation for the transition probabilities:

$$p_{ij}(t) = P[X(t + u) = j \mid X(u) = i] \quad \text{for any } u \geqslant 0. \tag{8.2'}$$

Since (8.2) is a conditional pmf, it satisfies the relation:

$$\sum_j p_{ij}(v, t) = 1 \quad \text{for all } i;\ 0 \leqslant v \leqslant t. \tag{8.3}$$

Let us denote the pmf of $X(t)$ (or the **state probabilities** at time t) by:

$$P_j(t) = P[X(t) = j], \quad j = 0, 1, 2, \ldots, \ t \geqslant 0. \tag{8.4}$$

It is clear that:

$$\sum_j P_j(t) = 1$$

for each $t \geqslant 0$, since at any given time the process must be in *some* state.

By using the theorem of total probability, for given $t > v$, we can express the pmf of $X(t)$ in terms of the transition probabilities $p_{ij}(v, t)$ and the pmf of $X(v)$:

$$\begin{aligned} P_j(t) &= P[X(t) = j] \\ &= \sum_i P[X(t) = j | X(v) = i] P[X(v) = i] \\ &= \sum_i p_{ij}(v, t)\ P_i(v). \end{aligned} \tag{8.5}$$

If we let $v = 0$ in (8.5), then:

$$P_j(t) = \sum_i p_{ij}(0, t) P_i(0). \tag{8.5'}$$

Hence, the probabilistic behavior of a Markov chain is completely determined once the transition probabilities $p_{ij}(v, t)$ and the initial probability vector $\mathbf{P}(0) = [P_0(0),\ P_1(0), \ldots]$ are specified.

The transition probabilities of a Markov chain $\{X(t),\ t \geqslant 0\}$ satisfy the **Chapman-Kolmogorov equation**: for all $i, j \in I$,

$$p_{ij}(v, t) = \sum_{k \in I} p_{ik}(v, u) p_{kj}(u, t) \qquad 0 \leqslant v < u < t. \tag{8.6}$$

To prove (8.6), we use the theorem of total probability:

$$\begin{aligned} P[X(t) = j \mid X(v) = i] = \sum_{k \in I} & P[X(t) = j \mid X(u) = k,\ X(v) = i] \\ & \cdot P[X(u) = k \mid X(v) = i]. \end{aligned}$$

The subsequent application of the Markov property (8.1) yields (8.6).

The direct use of (8.6) is difficult. Usually we obtain the transition probabilities by solving a system of differential equations that we derive next. For this purpose, under certain regularity conditions, we can show that for each j there is a nonnegative continuous function $q_j(t)$ defined by:

$$q_j(t) = -\frac{\partial}{\partial t} p_{jj}(v, t)\Big|_{v=t}$$

$$= \lim_{h \to 0} \frac{p_{jj}(t, t) - p_{jj}(t, t+h)}{h} = \lim_{h \to 0} \frac{1 - p_{jj}(t, t+h)}{h}. \quad (8.7)$$

Similarly for each i and j ($\neq i$) there is a nonnegative continuous function $q_{ij}(t)$ defined by:

$$q_{ij}(t) = \frac{\partial}{\partial t} p_{ij}(v, t)\Big|_{v=t}$$

$$= \lim_{h \to 0} \frac{p_{ij}(t, t+h) - p_{ij}(t, t)}{h} = \lim_{h \to 0} \frac{p_{ij}(t, t+h)}{h}. \quad (8.8)$$

Then the transition probabilities and the transition rates are related by:[1]

$$p_{ij}(t, t+h) = q_{ij}(t) \cdot h + o(h), \qquad i \neq j,$$

and

$$p_{jj}(t, t+h) = 1 - q_j(t) \cdot h + o(h), \qquad i = j.$$

Substituting $t + h$ for t in equation (8.6), we get:

$$p_{ij}(v, t+h) = \sum_k p_{ik}(v, u) p_{kj}(u, t+h)$$

which implies

$$p_{ij}(v, t+h) - p_{ij}(v, t) = \sum_k p_{ik}(v, u)[p_{kj}(u, t+h) - p_{kj}(u, t)].$$

Dividing both sides by h and taking the limit $h \to 0$ and $u \to t$, we get the differential equation known as **Kolmogorov's forward equation:** for $0 \leqslant v < t$ and $i, j \in I$:

$$\frac{\partial p_{ij}(v, t)}{\partial t} = [\sum_{k \neq j} p_{ik}(v, t) q_{kj}(t)] - p_{ij}(v, t) q_j(t). \quad (8.9)$$

In a similar fashion we can also derive **Kolmogorov's backward equation:**

$$\frac{\partial p_{ij}(v, t)}{\partial v} = [\sum_{k \neq i} p_{kj}(v, t) q_{ik}(v)] - p_{ij}(v, t) q_i(v).$$

[1] $o(h)$ is any function of h that approaches zero faster than h:

$$\lim_{h \to 0} \frac{o(h)}{h} = 0$$

Using (8.5) and (8.9) we can also derive a differential equation for the unconditional probability $P_j(t)$ as:

$$\frac{dP_j}{dt} = [\sum_{i \neq j} P_i(t) q_{ij}(t)] - P_j(t) q_j(t). \quad (8.9')$$

We use (8.9) when we want specifically to show the initial state, (8.9') when the initial state (or initial distribution) is implied.

In many important applications the transition probabilities $p_{ij}(t, t+h)$ do not depend on the initial time t but only on the elapsed time h (that is, the resulting Markov chain is **time-homogeneous**). This implies that the transition rates $q_{ij}(t)$ and $q_j(t)$ are independent of t. Unless otherwise stated, we will be concerned only with time-homogeneous situations. In this case the transition rates are denoted by q_{ij} and the transition probabilities $p_{ij}(t, t+h)$ by $p_{ij}(h)$. Equations (8.9) and (8.9') are rewritten as:

$$\frac{dp_{ij}(t)}{dt} = [\sum_{k \neq j} p_{ik}(t) q_{kj}] - p_{ij}(t) q_j, \quad (8.10)$$

$$\frac{dP_j}{dt} = \sum_{i \neq j} P_i(t) q_{ij} - P_j(t) q_j. \quad (8.10')$$

Even in this simpler case of a time-homogeneous Markov chain, solution of equation (8.10) to obtain the time-dependent probabilities $P_j(t)$ is quite difficult. Nevertheless, in many interesting situations a further reduction is possible in that the probabilities $P_j(t)$ approach a limit p_j as t approaches infinity. We wish to explore the conditions under which such a limiting distribution exists.

A classification of states for a continuous-parameter Markov chain is similar to the discrete-parameter case. A state i is said to be an **absorbing state** provided that $q_{ij} = 0$ for all $j \neq i$, so that, once entered, the process is destined to remain in that state. For a Markov chain with two or more absorbing states, the limiting probabilities $\lim_{t \to \infty} p_{ij}(t)$ may well depend upon the initial state.

A state j is said to be **reachable** from state i if for some $t > 0$, $p_{ij}(t) > 0$. A continuous-parameter Markov chain is said to be **irreducible** if every state is reachable from every other state.

THEOREM 8.1.

For an irreducible continuous-parameter Markov chain, the limits:

$$p_j = \lim_{t \to \infty} p_{ij}(t) = \lim_{t \to \infty} P_j(t), \qquad i, j \in I, \quad (8.11)$$

always exist and are independent of the initial state i.

If the limiting probabilities p_j exist, then:

$$\lim_{t\to\infty} \frac{dP_j(t)}{dt} = 0, \tag{8.12}$$

and, substituting into equation (8.10'), we get the following system of linear homogeneous equations (one for each state j):

$$0 = \sum_{i\neq j} p_i q_{ij} - p_j q_j. \tag{8.13}$$

This is the continuous analog of equation (7.14).

For the homogeneous system of equations, one possible solution is that $p_j = 0$ for all j. If another solution exists, then an infinite number of solutions can be obtained by multiplying by scalars. To determine a nonzero unique solution, we use the condition:

$$\sum_j p_j = 1. \tag{8.14}$$

Irreducible Markov chains that yield positive limiting probabilities $\{p_j\}$ in this way are called **recurrent non-null** or **positive recurrent** and the probabilities $\{p_j\}$, satisfying (8.13) and (8.14) are also known as steady-state probabilities. It is clear that a finite irreducible Markov chain must be positive recurrent, hence we can obtain its unique limiting probabilities by solving the finite system of equations (8.13) under the condition (8.14).

We have seen in Chapter 6 that the distribution of times that a continuous-parameter homogeneous Markov chain spends in a given state must be memoryless. This implies that holding times in a state of a continuous-parameter Markov chain of the homogeneous type are exponentially distributed. In the next section we study the limiting distribution of a special type of Markov chain, called the birth-death process. In Section 8.4, we study limiting distributions of several non-birth-death processes.

The study of transient behavior $[P_j(t),\ t \geq 0]$ is quite complex for a general Markov chain. In Sections 8.3 and 8.5 we consider special cases where it is possible to obtain an explicit solution for $P_j(t)$.

Problems

*1. For a homogeneous Markov chain, define the **transition-rate matrix** Q so that its diagonal elements are given by $-q_j$ and the (i, j) element is given by q_{ij} $(i \neq j)$. Define the matrix $P(t) = [p_{ij}(t)]$ and show that the forward and backward Kolmogorov equations can be rewritten as:

$$\frac{dP}{dt} = P(t)Q \quad \text{and} \quad \frac{dP}{dt} = QP(t)$$

with the initial condition $P(0) = I$, the identity matrix. Show that the solution to these matrix equations can be written as:

$$P(t) = e^{Qt} = I + \sum_{n=1}^{\infty} Q^n \frac{t^n}{n!},$$

assuming that matrix series converges. Generalize this result to the case of a nonhomogeneous Markov chain.

*2. For a homogeneous Markov chain show that the Laplace transform of the transition probability matrix $P(t)$, denoted by $\bar{P}(s)$, is given by:

$$\bar{P}(s) = (sI - Q)^{-1}.$$

*3. Show that the integral (convolution) form of the Kolmogorov forward equation is given by:

$$p_{ij}(\nu, t) = \delta_{ij}\, e^{\int_\nu^t q_i(\tau)\,d\tau} + \int_\nu^t \sum_k p_{ik}(\nu, x) q_{kj}(x) e^{-\int_x^t q_j(\tau)\,d\tau}\, dx,$$

where δ_{ij} is the Kronecker delta function defined by $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Specialize this result to the case of a homogeneous Markov chain.

8.2 THE BIRTH AND DEATH PROCESS

A continuous-parameter homogeneous Markov chain $\{X(t),\ t \geq 0\}$ with the state space $\{0, 1, 2, \ldots\}$ is known as a **birth-death process** if there exist constants λ_i $(i = 0, 1, \ldots)$ and μ_i $(i = 1, 2, \ldots)$ such that the transition rates are given by:

$$\begin{aligned} q_{i,i+1} &= \lambda_i, \\ q_{i,i-1} &= \mu_i, \\ q_i &= \lambda_i + \mu_i, \\ q_{ij} &= 0 \qquad \text{for } |i - j| > 1. \end{aligned}$$

The **birth rate** λ_i (≥ 0) is the rate at which births occur in state i, and the **death rate** μ_i (≥ 0) is the rate at which deaths occur in state i. These rates are assumed to depend only on state i and are independent of time. Note that only "nearest-neighbor" transitions are allowed. In a given state, births and deaths occur independently of each other. Such a process is a useful model of many situations in queuing theory and reliability theory.

The process will be in state k at time $t + h$ if one of the following mutually exclusive and collectively exhaustive events occurs:

1. The system is in state k at time t, and no changes of state occur in the interval $(t, t + h]$; the associated conditional probability is:

$$p_{k,k}(t, t + h) = 1 - q_k(t) \cdot h + o(h) = 1 - (\lambda_k + \mu_k) \cdot h + o(h).$$

2. The system is in state $k - 1$ at time t, and one birth occurs in the interval $(t, t + h]$; the associated conditional probability is:

$$p_{k-1,k}(t, t + h) = q_{k-1,k}(t) \cdot h + o(h) = \lambda_{k-1} \cdot h + o(h).$$

3. The system is in state $k+1$ and one death occurs in the interval $(t, t+h]$; the associated conditional probability is

$$p_{k+1,k}(t, t+h) = q_{k+1,k}(t) \cdot h + o(h) = \mu_{k+1} \cdot h + o(h).$$

4. Two or more transitions occur in the interval $(t, t+h]$, resulting in $X(t+h) = k$, with associated conditional probability $o(h)$.

Then by the theorem of total probability we have:

$$\begin{aligned} P[X(t+h) = k] &= P_k(t+h) \\ &= P_k(t)p_{k,k}(t, t+h) + P_{k-1}(t)p_{k-1,k}(t, t+h) \\ &\quad + P_{k+1}(t)p_{k+1,k}(t, t+h) + o(h). \end{aligned}$$

After rearranging, dividing by h, and taking the limit as $h \to 0$, we get:

$$\frac{dP_k(t)}{dt} = -(\lambda_k + \mu_k)P_k(t) + \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t), \qquad k \geq 1, \tag{8.15}$$

$$\frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t), \qquad k = 0,$$

where the special equation for $k = 0$ is required because the state space of the process is assumed to be $\{0, 1, 2, \ldots\}$. Equation (8.15) is a special case of equation (8.10'), with $q_{k-1,k} = \lambda_{k-1}$, $q_{k+1,k} = \mu_{k+1}$, and $q_k = (\lambda_k + \mu_k)$.

The solution of this system of differential-difference equations is a formidable task. However, if we are not interested in the transient behavior, then we can set the derivative $dP_k(t)/dt$ equal to zero, and the resulting set of difference equations provide the steady-state solution of the Markov chain. Let p_k denote the steady-state probability that the chain is in state k; that is, $p_k = \lim_{t \to \infty} P_k(t)$ (assuming it exists). Then the above differential-difference equations reduce to [a special case of equation (8.13)]:

$$0 = -(\lambda_k + \mu_k)p_k + \lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k+1}, \qquad k \geq 1, \tag{8.16}$$

$$0 = -\lambda_0 p_0 + \mu_1 p_1. \tag{8.17}$$

These are known as the **balance equations** and we can obtain them directly from the state diagram, shown in Figure 8.1, by equating the rates of flow

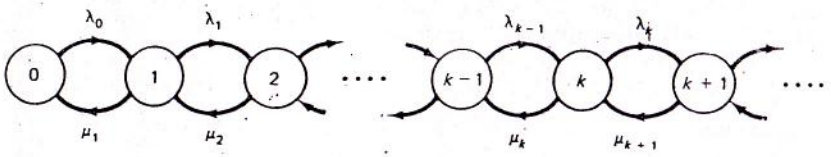

Figure 8.1 The state diagram of the birth-death process

into and out of each state. From the state diagram we have the rate of transition into state k as $\lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k+1}$ and the rate of transition out of state k as $(\lambda_k + \mu_k)p_k$. In the steady state no build-up occurs in state k, hence these two rates must be equal.

We should note the difference between this state diagram (of a continuous-parameter Markov chain) and the state diagram of a discrete-parameter Markov chain (Chapter 7). In the latter, the arcs are labeled with conditional proabilities; in the former they are labeled with state transition rates (hence, the name transition-rate diagram is sometimes used).

By rearranging equation (8.16), we get:

$$\lambda_k p_k - \mu_{k+1}p_{k+1} = \lambda_{k-1}p_{k-1} - \mu_k p_k = \cdots = \lambda_0 p_0 - \mu_1 p_1.$$

But from (8.17) we have $\lambda_0 p_0 - \mu_1 p_1 = 0$. It follows that:

$$\lambda_{k-1}p_{k-1} - \mu_k p_k = 0$$

and hence

$$p_k = \frac{\lambda_{k-1}}{\mu_k} p_{k-1}, \qquad k \geq 1.$$

Therefore:

$$p_k = \frac{\lambda_0\lambda_1 \cdots \lambda_{k-1}}{\mu_1\mu_2 \cdots \mu_k} p_0 = p_0 \prod_{i=0}^{k-1} \left(\frac{\lambda_i}{\mu_{i+1}}\right), \qquad k \geq 1. \tag{8.18}$$

Since $\sum_{k \geq 0} p_k = 1$, we have:

$$p_0 = \frac{1}{1 + \sum_{k \geq 1} \prod_{i=0}^{k-1} \left(\frac{\lambda_i}{\mu_{i+1}}\right)}. \tag{8.19}$$

Thus, the limiting distribution $(p_0, p_1, \ldots)$ is now completely determined. Note that the limiting probabilities are nonzero, provided that the series:

$$\sum_{k \geq 1} \prod_{i=0}^{k-1} \left(\frac{\lambda_i}{\mu_{i+1}}\right)$$

converges (in which case, all the states of the Markov chain are recurrent non-null).

Next we consider several special cases of the birth-death process.

8.2.1 The *M/M/1* Queue

We consider a single-server Markovian queue shown in Figure 8.2. Customer arrivals form a Poisson process with rate λ. Equivalently the customer interarrival times are exponentially distributed with mean $1/\lambda$. Service times of

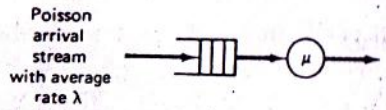

Figure 8.2 The $M/M/1$ queuing system

customers are independent identically distributed random variables, the common distribution being exponential with mean $1/\mu$. Assume that customers are served in their order of arrival (FCFS scheduling). If the "customer" denotes a job arriving into a computer system, then the server represents the computer system. [Since most computer systems consist of a set of interacting resources (and hence a network of queues), such a simple representation may be acceptable for "small" systems with little concurrency.] In another interpretation of the $M/M/1$ queue, the customer may represent a message and the server a communication channel.

Let $N(t)$ denote the number of customers in the system (those queued plus the one in service) at time t. [We change the notation from $X(t)$ to $N(t)$ to conform to standard practice.] Then $\{N(t),\ t \geq 0\}$ is a birth-death process with:

$$\lambda_k = \lambda, \quad k \geq 0; \qquad \mu_k = \mu, \quad k \geq 1.$$

The ratio $\rho = \lambda/\mu$ = mean service time /mean interarrival time, is an important parameter, called the **traffic intensity** of the system. Equations (8.18) and (8.19) in this case reduce to:

$$p_k = \left(\frac{\lambda}{\mu}\right)^k p_0 = \rho^k p_0$$

and

$$p_0 = \frac{1}{\sum_{k \geq 0} \rho^k} = 1 - \rho,$$

provided $\rho < 1$—that is, when the traffic intensity is less than unity. In the case that the arrival rate λ exceeds the service rate μ (i.e., $\rho \geq 1$), the geometric series in the denominator of the expression for p_0 diverges. In this case all the states of the Markov chain are either recurrent null or transient, hence the number of customers in the system tends to increase without bound. Such a system is called **unstable.** For a stable system ($\rho < 1$), the steady-state probabilities have a modified geometric distribution with parameter $1 - \rho$; that is,

$$p_k = (1 - \rho)\rho^k, \qquad k \geq 0. \tag{8.20}$$

The server utilization, $U_0 = 1 - p_0 = \rho$, is interpreted as the proportion of time the server is busy.

The mean and variance of the number of customers in the system are obtained using the properties of the modified geometric distribution as:

$$E[N] = \frac{\rho}{1 - \rho} \tag{8.21}$$

and

$$\text{Var}\,[N] = \frac{\rho}{(1 - \rho)^2}. \tag{8.22}$$

Let the random variable R denote the response time in the steady state. In order to compute the average response time $E[R]$ we use the well-known **Little's formula**, which states that the mean number of jobs in a queuing system in the steady-state is equal to the product of the arrival rate and the mean response time. When applied to the present case, Little's formula gives us:

$$E[N] = \lambda E[R].$$

Little's formula holds for a broad variety of queuing systems. For a proof see [STID 1974], and for its limitations see [BEUT 1980].

Using (8.21) and applying Little's formula to the present case, we have:

$$E[R] = \lambda^{-1} \frac{\rho}{1 - \rho} = \frac{1/\mu}{1 - \rho} = \frac{\text{average service time}}{\text{probability that the server is idle}}. \tag{8.23}$$

Note that the congestion in the system and hence the delay build rapidly as the traffic intensity increases (see Figures 8.3 and 8.4).

We may often employ a scheduling discipline other than FCFS. We distinguish between preemptive and nonpreemptive scheduling disciplines. A **nonpreemptive discipline** such as FCFS allows a job to complete execution once scheduled, whereas a **preemptive discipline** may interrupt the currently executing job in order to give preferential service to another job. A common example of a preemptive discipline is RR (round robin), which permits a job to remain in service for an interval of time referred to as its **quantum**. If the job does not finish execution within the quantum, it has to return to the end of the queue, awaiting further service. This gives preferential treatment to

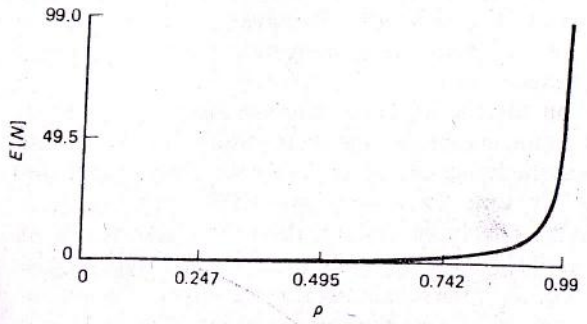

Figure 8.3 Expected number of jobs in system versus traffic intensity

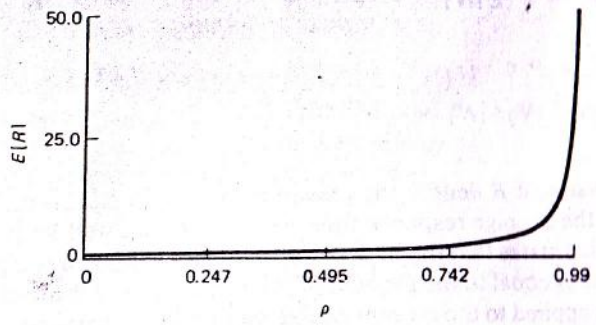

Figure 8.4 Average response time versus traffic intensity

short jobs at the expense of long jobs. When the time quantum approaches zero, the RR discipline is known as the PS (processor sharing) discipline.

Although we assumed FCFS scheduling discipline in deriving formula (8.20) for the queue-length distribution and formula (8.23) for the average response time, they hold for any scheduling discipline that satisfies the following conditions [COFF 1973, KOBA 1978]:

1. The server is not idle when there are jobs waiting for service.
2. The scheduler is not allowed to use any deterministic a priori information about job service times. Thus, for instance, if all job service times are known in advance, the use of a discipline known as SRPT (shortest remaining processing time first) is known to reduce $E[R]$ below that given by (8.23).
3. The service time distribution is not affected by the scheduling discipline.

Formulas (8.20) and (8.23) also apply for preemptive scheduling disciplines such as RR and PS, provided the overhead of preemption can be neglected (otherwise condition 3 above will be violated). We have also assumed in the above that a job is not allowed to leave the system before completion (see problems 3 and 4 below for exceptions).

Although the expression for the average response time (8.23) holds under a large class of scheduling disciplines, the distribution of the response time *does* depend upon the scheduling discipline. We shall derive the distribution function of the response time R assuming the FCFS scheduling discipline. If an arriving job finds n jobs in the system, then the response time is the sum of $n + 1$ random variables, $S + S'_1 + S_2 + \cdots + S_n$. Here S is the service time of the tagged job, S'_1 is the remaining service time of the job undergoing service, and $S_2, \ldots, S_n$ denote the service times of $(n - 1)$ jobs waiting in the queue. By our assumptions and the memoryless property of exponential distribution, these $(n + 1)$ random variables are independent and

exponentially distributed with parameter μ. Thus, the conditional Laplace transform of R given $N = n$ is the convolution:

$$L_{R|N}(s|n) = \left(\frac{\mu}{s+\mu}\right)^{n+1}. \tag{8.24}$$

Kleinrock [KLEI 1975] shows that the distribution of the number of jobs in the system as seen by an arriving job is the same as that given by (8.20). Then, applying the theorem of total Laplace transform, we obtain:

$$\begin{aligned} L_R(s) &= \sum_{n=0}^{\infty} \left(\frac{\mu}{s+\mu}\right)^{n+1}(1-\rho)\rho^n \\ &= \frac{\mu(1-\rho)}{s+\mu} \frac{1}{1-\frac{\mu\rho}{s+\mu}} \\ &= \frac{\mu(1-\rho)}{s+\mu(1-\rho)}. \end{aligned} \tag{8.25}$$

It follows that the response time R is exponentially distributed with parameter $\mu(1-\rho)$.

Other measures of system performance are easily obtained. Let the random variable W denote the waiting time; that is, let:

$$W = R - S. \tag{8.26}$$

Then:

$$E[R] = E[S] + E[W] = \frac{1}{\mu} + E[W].$$

It follows that the average waiting time is given by:

$$E[W] = \frac{1}{\mu(1-\rho)} - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)}. \tag{8.27}$$

If we now let the random variable Q denote the number of jobs waiting in the queue (excluding those if any, in service), then to determine the average number of jobs $E[Q]$ in the queue, we apply Little's formula to the queue excluding the server to obtain:

$$E[Q] = \lambda E[W] = \frac{\rho^2}{1-\rho}. \tag{8.28}$$

Note that the average number of jobs found in the server is:

$$E[N] - E[Q] = \rho. \tag{8.29}$$

Example 8.1

The capacity of a communication line is 2,000 bits per second. This line is used to transmit eight-bit characters, so the maximum rate is 250 characters per second. The

application calls for traffic from many devices to be sent on the line with a total volume of 12,000 characters per minute. In this case:

$$\lambda = \frac{12,000}{60} = 200 \text{ cps}, \qquad \mu = 250 \text{ cps},$$

and line utilization $\rho = \lambda/\mu = 4/5 = 0.8$.

The average number of characters waiting to be transmitted is $E[Q] = 0.8 \cdot 0.8/(1 - 0.8) = 3.2$, and the average transmission (including queuing delay) time per character is $E[N]/\lambda = 4/200$ seconds = 20 ms.

Example 8.2

We wish to determine the maximum call rate that can be supported by one telephone booth. Assume that the mean duration of a telephone conversation is three minutes, and that no more than a three-minute (average) wait for the phone may be tolerated; what is the largest amount of incoming traffic that can be supported?

1. $\mu = 1/3$ calls per minute; therefore, λ must be less than $1/3$ calls per minute, for the line to be stable.
2. The average waiting time $E[W]$ is given as three minutes; that is:

$$E[W] = \frac{\rho}{\mu(1 - \rho)} = 3,$$

and since $\mu = 1/3$, we get:

$$1 - \rho = \rho$$

or

$$\rho = 1/2.$$

Therefore, the call arrival rate is given by

$$\lambda = 1/6 \text{ calls per minute.} \qquad \#$$

Problems

1. Consider an $M/M/1$ queue with an average arrival rate λ and the average service rate μ. We have derived the distribution function of the response time R. Now we are interested in deriving the distribution function of the waiting time W. The waiting time W is the response time minus the service time. To get started, first compute the conditional distribution of W conditioned upon the number of jobs in the system, and later compute the unconditional distribution function. Note that W is a mixed random variable since its distribution function has a jump equal to $P(W = 0)$ at the origin.
2. A group of telephone subscribers is observed continuously during a 80-minute busy-hour period. During this time they make 30 calls, with the total conversation time being 4,200 seconds. Compute the call arrival rate and the traffic intensity.
3. Consider an $M/M/1$ queuing system in which the total number of jobs is limited to n owing to a limitation on queue size.
   (a) Find the steady-state probability that an arriving request is rejected because the queue is full.

   (b) Find the steady-state probability that the processor is idle.
   (c) Given that a request has been accepted, find its average response time.
4. The arrival of large jobs at a computing center forms a Poisson process with rate two per hour. The service times of such jobs are exponentially distributed with mean 20 minutes. Only four large jobs can be accommodated in the system at a time. Assuming that the fraction of computing power utilized by smaller jobs is negligible, determine the probability that a large job will be turned away because of lack of storage space.
5. Let the random variable T_k denote the holding time in state k of the $M/M/1$ queue. Starting from the given assumptions on the interarrival time and the service time distributions show that the distribution of T_k is exponential (for each k).
6. Derive an expression for the frequency of entering state 0 (server idle) in an $M/M/1$ queue. This quantity is useful in estimating the overhead of scheduling. Plot this probability as a function of ρ for a fixed μ.

8.2.2 The *M/M/m* Queue

Consider a queuing system with arrival rate λ as before, but where $m \geq 1$ servers, with rate μ each, share a common queue (see Figure 8.5). This gives rise to a birth-death model with the rates:

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

$$\mu_k = \begin{cases} k\mu, & 0 \leq k < m, \\ m\mu, & m \leq k. \end{cases}$$

The state diagram of this system is shown in Figure 8.6. The steady-state probabilities are given by [using equation (8.18)]:

$$\begin{aligned} p_k &= p_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} \\ &= p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}, \qquad k < m, \\ p_k &= p_0 \prod_{i=0}^{m-1} \frac{\lambda}{(i+1)\mu} \prod_{j=m}^{k-1} \frac{\lambda}{m\mu} \qquad (8.30) \\ &= p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{m!\, m^{k-m}}, \qquad k \geq m. \end{aligned}$$

Defining $\rho = \lambda/(m\mu)$, the condition for stability is given by $\rho < 1$. The expression for p_0 is obtained using (8.30) and the fact that $\sum_{k=0}^{\infty} p_k = 1$:

$$p_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!} \frac{1}{1-\rho}\right]^{-1}. \qquad (8.31)$$

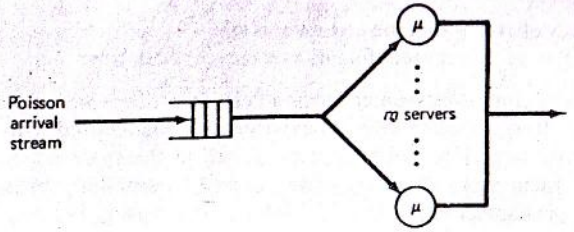

Figure 8.5 The $M/M/m$ queuing system

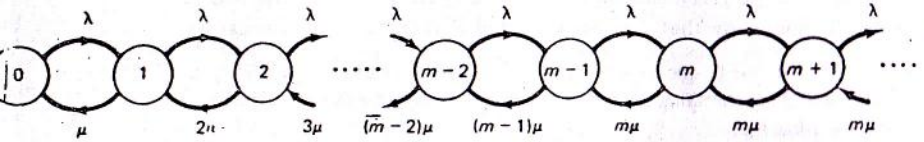

Figure 8.6 The state diagram of the $M/M/m$ queue

The expression for the average number of jobs in the system is (see problem 4 at the end of this section):

$$E[N] = \sum_{k \geq 0} kp_k = m\rho + \rho \frac{(m\rho)^m}{m!} \frac{p_0}{(1-\rho)^2}. \tag{8.32}$$

Let the random variable M denote the number of busy servers; then:

$$P(M = k) = \begin{cases} P(N = k) = p_k, & 0 \leq k \leq m-1, \\ P(N \geq m) = \sum_{k=m}^{\infty} p_k = \dfrac{p_m}{1-\rho}, & k = m. \end{cases}$$

The average number of busy servers is then:

$$E[M] = \sum_{k=0}^{m-1} kp_k + \frac{mp_m}{1-\rho},$$

which can be simplified (see problem 4 at the end of this section):

$$E[M] = m\rho = \frac{\lambda}{\mu}. \tag{8.33}$$

Thus, the utilization of any individual server is $\rho = \lambda/(m\mu)$, while the average number of busy servers is equal to the traffic intensity λ/μ.

The probability that an arriving customer is required to join the queue (or the probability of congestion) is derived as:

$$P[\text{queuing}] = \sum_{k=m}^{\infty} p_k = \frac{p_m}{1-\rho} = \frac{(m\rho)^m}{m!} \cdot \frac{p_0}{1-\rho}, \tag{8.34}$$

where p_0 is given in (8.31). Formula (8.34) finds wide application in telephone traffic theory and gives the probability that no trunk is available for an arriving call in an exchange with m trunks. This formula is referred to as Erlang's C formula (or Erlang's delayed-call formula).

Example 8.3

While designing a multiprocessor operating system, we wish to compare two different queuing schemes shown in Figure 8.7. The criterion for comparison will be the average response times $E[R_s]$ and $E[R_c]$. It is clear that the first organization corresponds to two independent $M/M/1$ queues, with $\rho = \lambda/(2\mu)$. Therefore, using equation (8.23), we have:

$$E[R_s] = \frac{\frac{1}{\mu}}{1 - \frac{\lambda}{2\mu}} = \frac{2}{2\mu - \lambda}.$$

On the other hand, the common queue organization corresponds to an $M/M/2$ system. To obtain $E[R_c]$, we first obtain $E[N_c]$ [using equation (8.32)] as:

$$E[N_c] = 2\rho + \frac{\rho(2\rho)^2}{2!} \frac{p_0}{(1-\rho)^2} \quad \text{where} \quad \rho = \frac{\lambda}{2\mu},$$

and using equation (8.31), we have:

$$p_0 = [1 + 2\rho + \frac{(2\rho)^2}{2!} \frac{1}{1-\rho}]^{-1} = \frac{1-\rho}{(1-\rho)(1+2\rho) + 2\rho^2} = \frac{1-\rho}{1+\rho}.$$

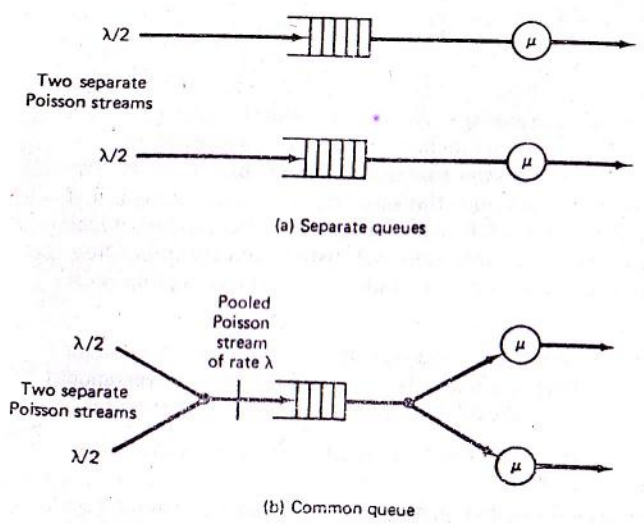

Figure 8.7 Queuing schemes

$$E[N_c] = 2\rho + 2\rho^3 \frac{1-\rho}{(1+\rho)(1-\rho)^2} = \frac{2\rho(1-\rho^2+\rho^2)}{1-\rho^2} = \frac{2\rho}{1-\rho^2}$$

and, using Little's formula, we have:

$$E[R_c] = \frac{E[N_c]}{\lambda} = \frac{2\,\frac{1}{2\mu}}{1-(\frac{\lambda}{2\mu})^2}$$

$$= \frac{1}{\mu(1-\rho^2)} = \frac{4\mu}{4\mu^2-\lambda^2}. \qquad (8.35)$$

Now:

$$E[R_s] = \frac{2}{2\mu-\lambda} = \frac{4\mu+2\lambda}{4\mu^2-\lambda^2} > E[R_c].$$

This implies that a common queue organization is better than a separate queue organization. This result generalizes to the case of m servers [KLEI 1976]. #

Example 8.4

Once again consider the problem of designing a system with two identical processors. We have two independent job streams with respective average arrival rates $\lambda_1 = 20$ and $\lambda_2 = 15$ per hour. The average service time for both job types is $1/\mu = 2$ min $= 1/30$ hours. Should we dedicate a processor per job stream, or should we pool the job streams and processors together (see Figure 8.8)? Let $E[R_{s1}]$ and $E[R_{s2}]$ be

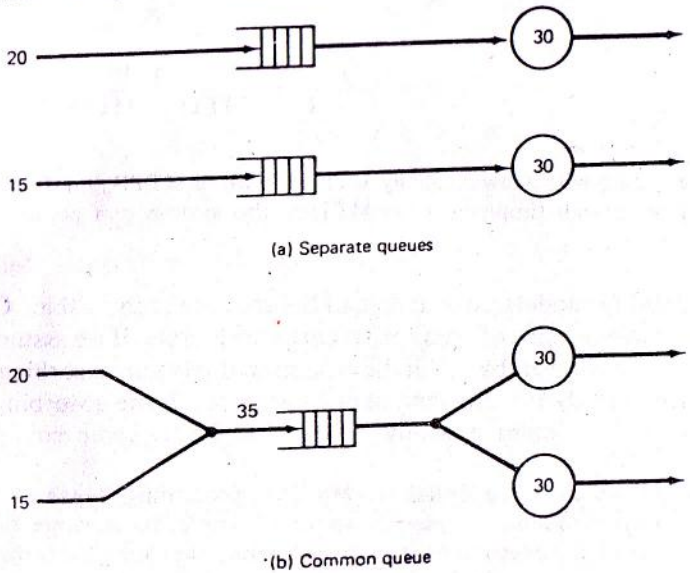

Figure 8.8 Queuing schemes for Example 8.4

the average response times of the two job streams in the separate queue organization and let $E[R_c]$ be the response time in the common-queue situation. Let $\rho_1 = \lambda_1/\mu = 20/30$, $\rho_2 = \lambda_2/\mu = 15/30$, $\rho = (\lambda_1+\lambda_2)/(2\mu) = 35/60$. Then:

$$E[R_{s1}] = \frac{\frac{1}{\mu}}{1-\rho_1} = \frac{\frac{1}{30}}{1-\frac{20}{30}}, \qquad \text{using formula (8.23)}$$

$$= \frac{1}{30-20} = \frac{1}{10} \text{ hour} = 6 \text{ minutes.}$$

$$E[R_{s2}] = \frac{\frac{1}{\mu}}{1-\rho_2} = \frac{\frac{1}{30}}{1-\frac{15}{30}}, \qquad \text{using formula (8.23)}$$

$$= \frac{1}{15} \text{ hour} = 4 \text{ minutes.}$$

$$E[R_c] = \frac{\frac{1}{\mu}}{1-\rho^2}, \qquad \text{using formula (8.35)}$$

$$= \frac{\frac{1}{30}}{1-(\frac{35}{60})^2} = 3.03 \text{ minutes.}$$

Clearly, it is much better to form a common pool of jobs. #

Problems

1. Consider a telephone switching system consisting of n trunks with an infinite caller population. The arrival stream is Poisson with rate λ, and call holding times are exponentially distributed with average $1/\mu$. The traffic offered, A (in Erlangs), is defined to be the average number of call arrivals per holding time. Thus, $A = \lambda/\mu = \rho$. We assume that an arriving call is lost if all trunks are busy. This is known as BCC (blocked calls cleared) scheduling discipline. Draw the state diagram and derive an expression for p_i, the steady-state probability that i trunks are busy. Show that this distribution approaches the Poisson distribution in the limit $n \to \infty$ (i.e., ample-trunks case). Therefore, for finite n, the above distribution is known as the "truncated Poisson distribution." Define the call congestion, B, as the proportion of lost calls in the long run. Then show that:

$$B = \frac{\frac{\rho^n}{n!}}{\sum_{i=0}^{n} \frac{\rho^i}{i!}}.$$

This is known as Erlang's B formula. Define traffic carried, C (in Erlangs), to be the average number of calls completed in a time interval $1/\mu$. Then:

$$C = \sum_{i=0}^{n} i p_i.$$

Verify that:

$$B = 1 - \frac{C}{A}.$$

2. Derive the steady-state distribution of the waiting time W for an $M/M/2$ queuing system as follows:
   (a) First show that:
   $$P(W = 0) = p_0 + p_1.$$
   (b) Now, conditioned on $n \geq 2$ jobs being present in the system at the time of arrival of the tagged job, argue that the distribution of W is $(n-1)$-stage Erlang with parameter 2μ.

   Compute the distribution function and hence compute the expected value of W.

3. [$M/M/\infty$ Queuing System] Suppose $P_n(t)$ is the probability that n telephone lines are busy at time t. Assume that infinitely many lines are available and that average call arrival rate is λ while average call duration is $1/\mu$. Derive the differential equation for $P_n(t)$. Solve the equation for $P_n(t)$ as $t \to \infty$. Let $E[N(t)]$ denote the average number of busy lines. Derive the differential equation for $E[N(t)]$. Obtain an expression for average length of queue $E[N]$ in the steady state. What is the average response time?

4. Show that the average number of busy servers for an $M/M/m$ queue in the steady state is given by:
   $$E[M] = \frac{\lambda}{\mu}.$$
   Also verify formula (8.32) for the average number in the system.

8.2.3 Finite State Space

We consider a special case of the birth-death process having a finite state space $\{0, 1, \ldots, n\}$, with constant birth rates $\lambda_i = \lambda$, $0 \leq i \leq n-1$, and constant death rates $\mu_i = \mu$, $1 \leq i \leq n$. Also let $\rho = \lambda/\mu$, as before. The state diagram is given by Figure 8.9, and the steady-state probabilities are:

$$p_i = \rho^i p_0, \qquad 0 \leq i \leq n,$$

$$p_0 = \frac{1}{\sum_{i=0}^{n} \rho^i} = \begin{cases} \dfrac{1-\rho}{1-\rho^{n+1}}, & \rho \neq 1 \\ \dfrac{1}{n+1}, & \rho = 1. \end{cases} \tag{8.36}$$

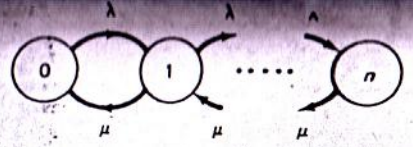

Figure 8.9 State diagram of a birth-death process with a finite state space

Note that such a system with a finite customer population will always be stable, no matter what value of ρ. Thus, (8.36) gives the steady-state probabilities for all finite values of ρ.

Example 8.5 (Machine Breakdown)

Consider a component with a constant failure rate λ. Upon a failure, it is repaired with an exponential repair-time distribution of parameter μ. Thus, the MTTF is $1/\lambda$ and the MTTR is $1/\mu$. This is an example of the Markov chain just discussed with $n = 1$ (a two-state system). Hence:

$$p_0 = \frac{1-\rho}{1-\rho^2} = \frac{1}{1+\rho}$$

and

$$p_1 = \frac{\rho}{1+\rho}.$$

The steady-state availability is the steady-state probability that the system is in state 0, the state with the system functioning properly. Thus:

$$\text{steady-state availability, } A = p_0 = \frac{1}{1+\rho} = \frac{1}{1+\frac{\lambda}{\mu}}$$

$$= \frac{\frac{1}{\lambda}}{\frac{1}{\lambda}+\frac{1}{\mu}} = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}}. \tag{8.37}$$

Note that a system with a low reliability will have a small MTTF, but if the repairs can be made fast enough (implying a low MTTR), the system may possess a high availability. #

Such availability models assume that all failures are recoverable. Consequently, the Markov chains of such systems are irreducible. If we assume that some failures are irrecoverable, then the system will have an absorbing state. In such cases, we study the distribution of time to reach the absorbing state (or failure state), and system reliability (see Section 8.5 for some examples).

Example 8.6 (Cyclic Queuing Model of a Multiprogramming System)

Consider the cyclic queuing model shown in Figure 8.10. Assume that the lengths of successive CPU execution bursts are independent exponentially distribut-

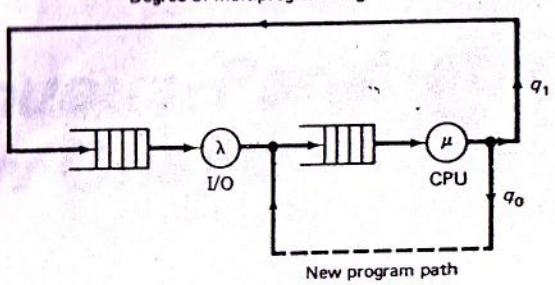

Figure 8.10 The cyclic queuing model of a multiprogramming system

ed random variables with mean $1/\mu$ and that successive I/O burst times are also independent exponentially distributed with mean $1/\lambda$. At the end of a CPU burst a program requests an I/O operation with probability $0 \leqslant q_1 \leqslant 1$, and it completes execution with probability q_0 $(q_1 + q_0 = 1)$. At the end of a program completion another statistically identical program enters the system, leaving the number of programs in the system at a constant level n (known as the **degree of multiprogramming**).

Let the number of programs in the CPU queue including any being served at the CPU denote the state of the system, i, where $0 \leqslant i \leqslant n$. Then the state diagram is given by Figure 8.11. Denoting $\lambda/(\mu q_1)$ by ρ, we see that the steady-state probabilities are given by:

$$p_i = \left(\frac{\lambda}{\mu q_1}\right)^i p_0 = \rho^i p_0, \quad \text{and} \quad p_0 = \frac{1}{\sum_{i=0}^{n} \rho^i},$$

so that:

$$p_0 = \begin{cases} \dfrac{1-\rho}{1-\rho^{n+1}}, & \rho \neq 1, \\ \dfrac{1}{n+1}, & \rho = 1. \end{cases}$$

The CPU utilization is given by:

$$U_0 = 1 - p_0 = \begin{cases} \dfrac{\rho - \rho^{n+1}}{1-\rho^{n+1}}, & \rho \neq 1, \\ \dfrac{n}{n+1}, & \rho = 1. \end{cases} \tag{8.38}$$

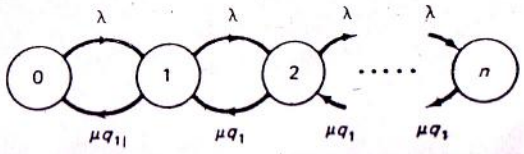

Figure 8.11 The state diagram for the cyclic queuing model

Let $C(t)$ denote the number of jobs completed by time t. Then the (time) average $C(t)/t$ converges, under appropriate conditions, to a limit as t approaches ∞ [ROSS 1970]. This limit is the average system throughput in the steady-state, and (with a slight abuse of notation) it is denoted here by $E[T]$. Whenever the CPU is busy, the rate at which CPU bursts are completed is μ, and a fraction q_0 of these will contribute to the throughput. Then:

$$E[T] = \mu q_0 U_0. \tag{8.39}$$

For fixed values of μ and q_0, $E[T]$ is proportional to the CPU utilization.

Let the random variable B_0 denote the total CPU time requirement of a tagged program. Then $B_0 \sim$ EXP (μq_0). This is true because B_0 is the random sum of K CPU service bursts, which are independent EXP (μ) random variables. Here the random variable K is the number of visits to the CPU per program and hence is geometrically distributed with parameter q_0. The required result is then obtained from our discussion on random sums in Chapter 5. Alternatively, the average number of visits V_0 to the CPU is $V_0 = 1/q_0$ (see Example 7.17), and thus $E[B_0] = V_0 E[S_0] = 1/(\mu q_0)$, where $E[S_0] = 1/\mu$ is the average CPU time per burst.

The average throughput can now be rewritten as.

$$E[T] = \frac{U_0}{E[B_0]} \tag{8.40}$$

If B_1 represents the total I/O service time per program, then as in the case of CPU:

$$E[B_1] = \frac{q_1}{q_0}\,\frac{1}{\lambda} = V_1\, E[S_1],$$

where the average number of visits V_1 to the I/O device is given by $V_1 = q_1/q_0$ (by Example 7.17), and $E[S_1] = 1/\lambda$ is the average time per I/O operation. [Note that if U_1 denotes the utilization of the I/O device then, similar to (8.40), we have $E[T] = U_1/E[B_1]$.] Now the parameter ρ can be rewritten:

$$\rho = \frac{\lambda}{\mu q_1} = \frac{q_0 \lambda}{q_1} \cdot \frac{1}{\mu q_0} = \frac{E[B_0]}{E[B_1]}. \tag{8.41}$$

Thus ρ indicates the relative measure of the CPU versus I/O requirements of a program. If the CPU requirement $E[B_0]$ is less than the I/O requirement $E[B_1]$ (that is, $\rho < 1$), the program is said to be **I/O-bound**; if $\rho > 1$, then program is said to be **CPU-bound**; and otherwise it is called **balanced**.

In Figure 8.12 we have plotted U_0 as a function of the balance factor ρ and of the degree of multiprogramming n. When $\rho << 1$ or $\rho >> 1$, U_0 is insensitive to n. Thus, multiprogramming is capable of improving throughput only when the workload is nearly balanced (that is, ρ is close to 1). #

Example 8.7

Let us return to the availability model of Example 8.5 and augment the system with $(n-1)$ identical copies of the component, which are to be used in a standby spare mode. Assume that an unpowered spare does not fail and that switching a spare is a fault-free process. Then the system is in state k provided that $n - k$ units are in