2905 Queueing Theory and Simulation

PART II: MARKOVIAN QUEUEING SYSTEMS

6 Introduction to Queueing Systems

A queueing situation is basically characterized by a flow of customers arriving at a service facility. On arrival at the facility the customer may be served immediately by a server or, if all the servers are busy, may have to wait in a queue until a server is available. The customer will then leave the system upon completion of service. The following are some typical examples of such queueing situations:

(i) Shoppers waiting for checkout in a supermarket [Customer: shoppers; servers: cashiers].

(ii) Diners waiting for tables in a restaurant [Customers: diners; servers: tables].

(iii) Patients waiting at an outpatient clinic [Customers: patients; servers: doctors].

(iv) Broken machines waiting to be serviced by a repairman [Customers: machines; server: repairman].

(v) People waiting to take lifts. [Customers: people; servers: lifts].

(vi) Parts waiting at a machine for further processing. [Customers: parts; servers: machine].

In general, the **arrival pattern** of the customers and the **service time** allocated to each customer can only be specified probabilistically. Such service facilities are difficult to schedule "**optimally**" because of the presence of randomness element in the arrival and service patterns.

A mathematical theory has thus evolved that provides means for analyzing such situations. This is known as **queueing theory** (waiting line theory, congestion theory, the theory of stochastic service system), which analyzes the operating characteristics of a queueing situation with the use of probability theory.

Examples of the characteristics that serve as a measure of the performance of a system are the "expected waiting time until the service of a customer is completed" or "the percentage of time that the service facility is not used" (the servers in the system are idle). Availability of such measures enables analysts to decide on an optimal way of operating such a system.

6.1 Basic Elements of Queueing Models

A queueing system is specified by the following elements.

(i) **Input Process**: How do customers arrive? Often, the input process is specified in terms of the distribution of the lengths of time between consecutive customer arrival instants (called the **inter-arrival times**). In some models, customers arrive and are served **individually** (e.g. supermarkets and clinic).

- In other models, customers may arrive and/or be served in groups (e.g. lifts) and is referred to as **bulk** queues.

- Customer arrival pattern also depends on the **source** from which calls for service (arrivals of customers) are generated. The calling source may be capable of generating a **finite number** of customers or (theoretically) **infinitely many** customers.

- In a machine shop with 4 machines (the machines are the customers and the repairman is the server), the calling source before any machine breaks down consists of 4 potential customers (i.e. anyone of the 4 machines may break down and therefore calls for the service of the repairman). Once a machine breaks down, it becomes a customer receiving the service of the repairman (until the time it is repaired), and only 3 other machines are capable generating new calls for service. This is a typical example of a **finite** source, where an arrival affects the rate of arrival of new customers.

- For shoppers in a supermarket, the arrival of a customer normally does not affect the source for generating new customer arrivals, and is therefore referred to as an input process with infinite source.

(ii) **Service Process**: The time allocated to serve a customer (service time) in a system (e.g. the time that a patient is served by a doctor in an outpatient clinic) varies and is assumed to follow some **probability distribution**.

- Some facility may include more than one server, thus allowing as many customers as the number of servers to be serviced simultaneously (e.g. supermarket cashiers). In this case, all servers offer the same type of service and the facility is said to have **parallel servers** .

- In some other models, a customer must pass through a series of servers one after the other before service is completed (e.g. processing a product on a sequence of machines). Such situations are known as **queues in series** or **tandem queues**.

(iii) **Queue Discipline**: The manner that a customer is chosen from the waiting line to start service is called the **queue discipline** . The most common discipline is the first-come-first-served rule (FCFS). Service in random order (SIRO), last-come-first-serve (LCFS) and service with **priority** are also used.

- If all servers are busy, in some models an arriving customer may leave immediately (**Blocked Customers Cleared: BCC**), or in some other models may wait until served (**Blocked Customers Delay: BCD**). In some facility, there is a restriction on the size of the queue. If the queue has reached a certain size, then all new arrivals will be cleared from the system.

6.2 Some Simple Examples

(i) (**Input process**) If the inter-arrival time of any two customers is a constant, let say one hour then at the end of the second hour there will be 2 arrived customers.

- Suppose that customers only arrive at the end of each hour and the probability that there is an arrival of customer is 0.5.

- Let x be the number of customers arrived at the end of the second hour. Then by the end of the second hour, we won't know the number of customers arrived.

- However, we know the probability that there are x arrived customers is given by (why?)

$$P(x = 0) = 0.25, \quad P(x = 1) = 0.5 \quad \text{and} \quad P(x = 2) = 0.25.$$

(ii) (**Service Process**) Suppose that there is a job to be processed by a machine. The job requires a one-hour machine time. A reliable machine will take one hour to finish the job.

- If the machine is unreliable and it may break down at the beginning of every hour with a probability of p. Once it breaks down it takes one hour to fix it. But it may break down immediately after the repair with the same probability p $(0 < p < 1)$. Clearly it takes at least one hour to finish the job but it may take much longer time.

- Let x be the number of hours to finish the job. Then the probability that the job can be finished at the end of the nth hour is given by the Geometric distribution

$$P(x = k) = p^{k-1}(1 - p), \quad k = 1, 2, \ldots.$$

(iii) (**Queueing Disciplines**) Suppose there are three customers A, B and C waiting at a counter for service and their service times are in the following order 10 minutes, 20 minutes and 30 minutes. It takes $10 + 20 + 30 = 60$ minutes to finish all the service. However, the average waiting time before service for the three customers can be quite different for different service disciplines.

- Case 1 (FCFS): The waiting time for the first customer is zero, the waiting time for the second customer is 10 minutes and the waiting time for the third customers is $10 + 20 = 30$ minutes. Therefore the average waiting time before service is

$$(0 + 10 + 30)/3 = 40/3.$$

- Case 2 (LCFS): The waiting time for the first customer is zero, the waiting time for the second customer is 30 minutes and the waiting time for the third customers is $30 + 20 = 50$ minutes. Therefore the average waiting time before service is

$$(0 + 30 + 50)/3 = 80/3$$

minutes which is twice of that in Case 1!

7 Some Definitions in Queueing Theory

To analyze a queueing system, normally we try to estimate quantities such as the average number of customers in the system, the fluctuation of the number of customers waiting, the proportion of time that the servers are idle, $\ldots$, etc.

• Let us now define formally some entities that are frequently used to measure the effectiveness of a queueing system (with s parallel servers).

(i) p_j = the probability that there are j customers in the system (waiting or in service) at an arbitrary epoch (given that the system is in statistical equilibrium or steady state). Equivalently p_j is defined as the proportion of time that there are j customers in the system (in steady state).

(ii) a = **offered load** = mean number of requests per service time. (In a system where blocked customers are cleared, requests that are lost are also counted.)

(iii) ρ = **traffic intensity** = offered load per server = a/s $(s < \infty)$.

(iv) a' = **carried load** = mean number of busy servers.

(v) $\rho' =$ **server occupancy** (or utilization factor) = carried load per server = a'/s.

(vi) $W_s =$ **the mean waiting time in the system**,i.e the mean length of time from the moment a customer arrives until the customer leaves the system (also called sojourn time).

(vii) $W_q =$ **the mean waiting time in the queue**, i.e. the mean length of time from the moment a customer arrives until the customer' service starts .

(viii) $L_s =$ **the mean number of customers in the system**, i.e. including all the customers waiting in the queue and all those being served.

(ix) $L_q =$ **the mean number of customers waiting in the queue**.

Remarks:

(i) If the mean arrival rate is λ and the mean service time is τ then the offered load $a = \lambda\tau$.

(ii) For an s server system, the carried load

$$\boxed{a' = \sum_{j=0}^{s-1} j p_j + s \sum_{j=s}^{\infty} p_j.}$$

Hence

$$a' \leq s \sum_{j=0}^{\infty} p_j = s \quad \text{and} \quad \rho' = \frac{a'}{s} \leq 1.$$

(iii) If $s = 1$ then $a' = \rho'$ and $a = \rho$.

(iv) The carried load can also be considered as the mean number of customers completing service per mean service time τ. Hence in a system where blocked customers are cleared, clearly the carried load is less than the offered load.

On the other hand, if all requests are handled, then we have

The carried load = The offered load.

In general $a' = a(1 - B)$ where $B =$ proportion of customers lost (or requests that are cleared).

7.1 Kendall's Notation

It is convenient to use a shorthand notation (introduced by D.G.Kendall) of the form $a/b/c/d$ to describe queueing models, where a specifies the arrival process, b specifies the service time, c is the number of servers and d is the number of waiting space. For example,

(i) GI/M/s/n : General Independent input, exponential (Markov) service time, s servers, n waiting space;

(ii) M/G/s/n : Poisson (Markov) input, arbitrary (General) service time, s servers, n waiting space;

(iii) M/D/s/n : Poisson (Markov) input, constant (Deterministic) service time, s servers, n waiting space;

(iv) E_k/M/s/n: k-phase Erlangian inter-arrival time, exponential (Markov) service time, s servers, n waiting space;

(v) M/M/s/n : Poisson input, exponential service time, s servers, n waiting space.

Here are some examples.

(i) M/M/2/10 represents a queueing system whose arrival and service process are random and there are 2 servers and 10 waiting space in the system.

(ii) M/M/1/∞ represents a queueing system whose arrival and service process are random and there is one server and no limit in waiting space.

8 Queueing Systems of One Server

In this section we will consider queueing systems having one server only.

8.1 One-server Queueing Systems Without Waiting Space (Re-visit)

Consider a one-server system of two states: 0 (idle) and 1 (busy).

The inter-arrival time of customers follows the exponential distribution with parameter λ.

The service time also follows the exponential distribution with parameter μ. There is no waiting space in the system.

An arrived customer will leave the system when he finds the server is busy (An M/M/1/0 queue). This queueing system resembles an one-line telephone system without call waiting.

8.2 Steady State Probability Distribution

We are interested in the long-run behavior of the system, i.e. when $t \to \infty$. Why?

Fact to note: Let $P_0(t)$ and $P_1(t)$ be the probability that there is 0 and 1 customer in the system.

If at $t = 0$ there is a customer in the system, then

$$P_0(t) = \frac{\mu}{\lambda+\mu}\left(1 - e^{-(\lambda+\mu)t}\right)$$

and

$$P_1(t) = \frac{1}{\lambda+\mu}\left(\mu e^{-(\lambda+\mu)t} + \lambda\right).$$

Here $P_0(t)$ and $P_1(t)$ are called the **transient probabilities**. We have

$$p_0 = \lim_{t\to\infty} P_0(t) = \frac{\mu}{\lambda+\mu} \quad \text{and} \quad p_1 = \lim_{t\to\infty} P_1(t) = \frac{\lambda}{\lambda+\mu}.$$

Here p_0 and p_1 are called the **steady state probabilities**.

- Moreover, we have

$$\left|P_0(t) - \frac{\mu}{\lambda+\mu}\right| = \frac{\mu e^{-(\lambda+\mu)t}}{\lambda+\mu} \to 0 \quad \text{as } t \to \infty$$

and

$$\left|P_1(t) - \frac{\lambda}{\lambda+\mu}\right| = \frac{\mu e^{-(\lambda+\mu)t}}{\lambda+\mu} \to 0 \quad \text{as } t \to \infty$$

very fast. This means that system will go into the steady state very fast. We will focus on the steady state probability instead of the transient probability.

8.3 The Meaning of the Steady State Probability

The meaning of the steady state probabilities p_0 and p_1 is as follows.

In the long run, the probability that there is no customer in the system is p_0 and there is one customer in the system is p_1.

For the server: In the other words, in the long run, the proportion of time that the server is idle is given by p_0 and the proportion of time that the server is busy is given by p_1.

For the customers: In the long run, the probability that an arrived customer can have his/her service is given by p_0 and the probability that an arrived customers will be rejected by the system is given by p_1.

How to find the steady state probability? We are going to develop a method based on Markov chain and generator matrix to solve the steady state probabilities.

Remarks:

(i) The system goes to its steady state very quickly.

(ii) In general it is much easier to obtain the steady state probabilities of a queueing system than the transient probabilities.

8.4 The Markov Chain and the Generator Matrix

A queueing system can be represented by a **Markov chain** (states and transition rates). We use the number of customers in the system to represent the state of the system. Therefore we have two states (0 and 1). The transition rate from State 1 to State 0 is μ and The transition rate from State 0 to State 1 is λ.

• In State 0, change of state occurs when there is an arrival of customers and the waiting time is exponentially distributed with parameter λ.

• In State 1, change of state occurs when the customer finishes his/her service and the waiting time is exponentially distributed with parameter μ.

• Recall that from the no-memory property, the waiting time distribution for change of state is the same independent of the past history (e.g. how long the customer has been in the system).

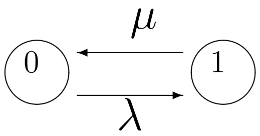

Figure 8.1 The Markov Chain of the Two-state System.

- From the **Markov chain**, one can construct the **generator matrix** as follows:

$$A_1 = \begin{pmatrix} -\lambda & \mu \\ \lambda & -\mu \end{pmatrix}.$$

What is the meaning of the generator matrix? The steady state probabilities will be the solution of the following linear system:

$$A_1\mathbf{p} = \begin{pmatrix} -\lambda & \mu \\ \lambda & -\mu \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad (1)$$

subject to $p_0 + p_1 = 1$.

Remarks:

(1) Given a Markov chain (generator matrix) one can construct the corresponding generator (Markov chain).

2) To interpret the system of linear equations. We note that in steady state, the expected incoming rate and the expected out going rate at any state must be equal. Therefore, we have the followings:

At State 0: expected out going rate $= \lambda p_0 = \mu p_1 =$ expected incoming rate;

At State 1: expected out going rate $= \mu p_1 = \lambda p_0 =$ expected incoming rate.

8.5 One-Server Queueing System with Waiting Space

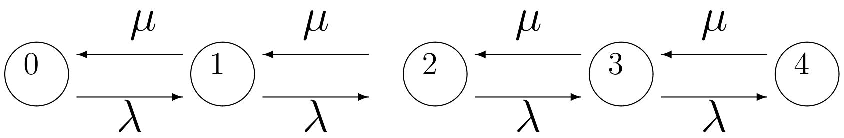

Figure 8.2 The Markov chain for the M/M/1/3 queue.

Consider an M/M/1/3 queue. The inter-arrival of customers and the service time follow the exponential distribution with parameters λ and μ respectively. Therefore there are 5 possible states. Why?

The generator matrix is a 5 × 5 matrix.

$$A_2 = \begin{pmatrix} -\lambda & \mu & & & 0 \\ \lambda & -\lambda-\mu & \mu & & \\ & \lambda & -\lambda-\mu & \mu & \\ & & \lambda & -\lambda-\mu & \mu \\ 0 & & & \lambda & -\mu \end{pmatrix}. \qquad (2)$$

Let the steady state probability distribution be

$$\mathbf{p} = (p_0, p_1, p_2, p_3, p_4)^t.$$

In steady state we have $A_2\mathbf{p} = \mathbf{0}$.

We can interpret the system of equations as follows:

At State 0: expected out going rate = $\lambda p_0 = \mu p_1$ = expected incoming rate;

At State 1: expected out going rate = $(\lambda + \mu)p_1 = \lambda p_0 + \mu p_2$ =expected incoming rate.

At State 2: expected out going rate = $(\lambda + \mu)p_2 = \lambda p_1 + \mu p_3$ = expected incoming rate;

At State 3: expected out going rate = $(\lambda + \mu)p_3 = \lambda p_2 + \mu p_4$ = expected incoming rate.

At State 4: expected out going rate = $\mu p_4 = \lambda p_3$ = expected incoming rate.

We are going to solve p_1, p_2, p_3, p_4 in terms of p_0.

From the first equation $-\lambda p_0 + \mu p_1 = 0$, we have

$$p_1 = \frac{\lambda}{\mu} p_0.$$

From the second equation $\lambda p_0 - (\lambda + \mu)p_1 + \mu p_2 = 0$, we have

$$p_2 = \frac{\lambda^2}{\mu^2} p_0.$$

From the third equation $\lambda p_1 - (\lambda+\mu)p_2 + \mu p_3 = 0$, we have

$$p_3 = \frac{\lambda^3}{\mu^3} p_0.$$

Finally from the fourth equation $\lambda p_2 - (\lambda+\mu)p_3 + \mu p_4 = 0$, we have

$$p_4 = \frac{\lambda^4}{\mu^4} p_0.$$

The last equation is not useful as A_2 is singular (Check!).

To determine p_0 we make use of the fact that

$$p_0 + p_1 + p_2 + p_3 + p_4 = 1.$$

Therefore

$$p_0 + \frac{\lambda}{\mu} p_0 + \frac{\lambda^2}{\mu^2} p_0 + \frac{\lambda^3}{\mu^3} p_0 + \frac{\lambda^4}{\mu^4} p_0 = 1.$$

Let $\rho = \lambda/\mu$, we have

$$p_0 = (1 + \rho + \rho^2 + \rho^3 + \rho^4)^{-1}, \quad p_i = p_0 \rho^i, i = 1, 2, 3, 4.$$

What will be the solution for a general one-server queueing system (M/M/1/n) ? We will discuss it in the next section.

9 General One-server Queueing System

Consider a one-server queueing system with waiting space. The inter-arrival of customers and the service time follows the Exponential distribution with parameters λ and μ respectively.

There is a waiting space of size $n-2$ in the system. An arrived customer will leave the system only when he finds no waiting space left. This is an M/M/1/$n-2$ queue.

We say that the system is in state i if there are i customers in the system. The minimum number of customers in the system is 0 and the maximum number of customers is $n-1$ (one at the server and $n-2$ waiting in the queue). Therefore there are n possible states in the system. The Markov chain of the system is shown in the figure.

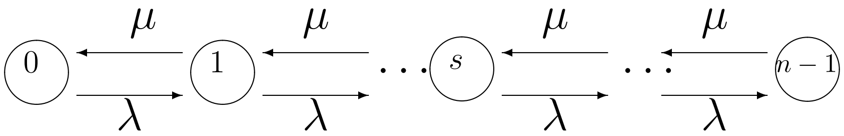

Figure 9.1 The Markov Chain for the M/M/1/n-2 Queue.

If we order the state from 0 up to $n-1$, then the generator matrix for the Markov chain is given by the following tridiagonal matrix A_2:

$$\begin{array}{cc}
 & \begin{array}{cccccccc} 0 & 1 & 2 & 3 & \cdots & n-3 & n-2 & n-1 \end{array} \\
\begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \\ \vdots \\ \vdots \\ n-2 \\ n-1 \end{array} &
\left(\begin{array}{cccccccc}
-\lambda & \mu & & & & & & 0 \\
\lambda & -\lambda-\mu & \mu & & & & & \\
 & \ddots & \ddots & \ddots & & & & \\
 & & \lambda & -\lambda-\mu & \mu & & & \\
 & & & \lambda & -\lambda-\mu & \mu & & \\
 & & & \ddots & \ddots & \ddots & & \\
 & & & & & \lambda & -\lambda-\mu & \mu \\
0 & & & & & & \lambda & -\mu
\end{array}\right)
\end{array}. \qquad (3)$$

We are going to solve the probability distribution $\mathbf{p}$. Let

$$\mathbf{p} = (p_0, p_1, \ldots, p_{n-2}, p_{n-1})^t$$

be the steady state probability vector. Here p_i is the steady state probability that there are i customers in the system and we have also

$$A_2\mathbf{p} = \mathbf{0} \quad \text{and} \quad \sum_{i=0}^{n-1} p_i = 1.$$

To solve p_i we begin with the first equation:

$$-\lambda p_0 + \mu p_1 = 0 \;\Rightarrow p_1 = \frac{\lambda}{\mu} p_0.$$

We then proceed to the second equation:

$$\lambda p_0 - (\lambda + \mu)p_1 + \mu p_2 = 0 \;\Rightarrow p_2 = -\frac{\lambda}{\mu}p_0 + (\frac{\lambda}{\mu} + 1)p_1 \;\Rightarrow p_2 = \frac{\lambda^2}{\mu^2}p_0.$$

Inductively we may get

$$p_3 = \frac{\lambda^3}{\mu^3}p_0,\; p_4 = \frac{\lambda^4}{\mu^4}p_0,\; \ldots,\; p_{n-1} = \frac{\lambda^{n-1}}{\mu^{n-1}}p_0.$$

Let $\rho = \lambda/\mu$ (the traffic intensity), we have

$$p_i = \rho^i p_0, \quad i = 0, 1, \ldots, n-1.$$

To solve for p_0 we need to make use of the condition

$$\sum_{i=0}^{n-1} p_i = 1.$$

Therefore we get

$$\sum_{i=0}^{n-1} p_i = \sum_{i=0}^{n-1} \rho^i p_0 = 1.$$

One may obtain

$$p_0 = \frac{1-\rho}{1-\rho^n}.$$

Hence the steady state probability vector $\mathbf{p}$ is given by

$$\frac{1-\rho}{1-\rho^n}(1, \rho, \rho^2, \ldots, \rho^{n-1})^t.$$

9.1 Performance of a Queueing System

Using the steady state probability distribution, one can compute

(a) the probability that a customer finds no more waiting space left when he arrives

$$p_{n-1} = \frac{1-\rho}{1-\rho^n}\rho^{n-1}.$$

(b) the probability that a customer finds the server is not busy (he can have the service immediately) when he arrives

$$p_0 = \frac{1-\rho}{1-\rho^n}.$$

(c) the expected number of customer at the server:

$$\begin{aligned} L_c &= 0 \cdot p_0 + 1 \cdot (p_1 + p_2 + \ldots + p_{n-1}) \\ &= \frac{1-\rho}{1-\rho^n}(\rho + \rho^2 + \ldots + \rho^{n-1}) \\ &= \frac{\rho(1-\rho^{n-1})}{1-\rho^n}. \end{aligned} \tag{4}$$

(d) the expected number of customers in the system is given by

$$\begin{aligned} L_s &= \sum_{i=0}^{n-1} i p_i = \sum_{i=1}^{n-1} i p_0 \rho^i \\ &= \frac{\rho - n\rho^n + (n-1)\rho^{n+1}}{(1-\rho)(1-\rho^n)}. \end{aligned} \tag{5}$$

(e) the expected number of customers in the queue

$$\begin{aligned} L_q &= \sum_{i=1}^{n-1}(i-1)p_i = \sum_{i=1}^{n-1}(i-1)p_0\rho^i = \sum_{i=1}^{n-1} ip_0\rho^i - \sum_{i=1}^{n-1} p_0\rho^i \\ &= \frac{\rho^2-(n-1)\rho^n+(n-2)\rho^{n+1}}{(1-\rho)(1-\rho^n)}. \end{aligned} \tag{6}$$

We note that $L_s = L_q + L_c$.

Remark: To obtain the results in (d) and(e) we need the following results.

$$\begin{aligned} \sum_{i=1}^{n-1} i\rho^i &= \frac{1}{1-\rho}\sum_{i=1}^{n-1}(1-\rho)i\rho^i \\ &= \frac{1}{1-\rho}(\sum_{i=1}^{n-1} i\rho^i - \sum_{i=1}^{n-1} i\rho^{i+1}) \\ &= \frac{1}{1-\rho}(\rho+\rho^2+\ldots+\rho^{n-1}-(n-1)\rho^n) \\ &= \frac{\rho+(n-1)\rho^{n+1}-n\rho^n}{(1-\rho)^2}. \end{aligned} \tag{7}$$

Moreover if $|\rho| < 1$ we have

$$\sum_{i=1}^{\infty} i\rho^i = \frac{\rho}{(1-\rho)^2}.$$

10 Queueing Systems with Multiple Servers

Now let us consider a more general queueing system with s parallel and identical exponential servers. The customer arrival rate is λ and the service rate of each server is μ. There are $n-s-1$ waiting space in the system.

• The queueing discipline is again **FCFS**. When a customer arrives and finds all the servers busy, the customer can still wait in the queue if there is waiting space available. Otherwise, the customer has to leave the system, this is an M/M/$s/n-s-1$ queue.

• Before we study the steady state probability of this system, let us discuss the following example (revisited).

• Suppose there are k identical independent busy exponential servers, let t be the waiting time for one of the servers to be free (change of state), i.e. one of the customers finishes his service.

• We let $t_1, t_2, \ldots, t_k$ be the service time of the k customers in the system. Then t_i follows the Exponential distribution $\lambda e^{-\lambda t}$ and

$$\boxed{t = \min\{t_1, t_2, \ldots, t_k\}}.$$

We will derive the probability density function of t.

We note that

$$\begin{aligned}\text{Prob}(t \geq x) &= \text{Prob}(t_1 \geq x) \cdot \text{Prob}(t_2 \geq x) \ldots \text{Prob}\,(t_k \geq x) \\ &= \left(\int_x^\infty \lambda e^{-\lambda t} dt \right)^k \\ &= \left(e^{-\lambda x} \right)^k = e^{-k\lambda x}. \end{aligned} \tag{8}$$

- Thus

$$\int_x^\infty f(t)dt = e^{-k\lambda x} \quad \text{and} \quad f(t) = k\lambda e^{-k\lambda t}.$$

Therefore the waiting time t is also exponentially distributed with parameter $k\lambda$.

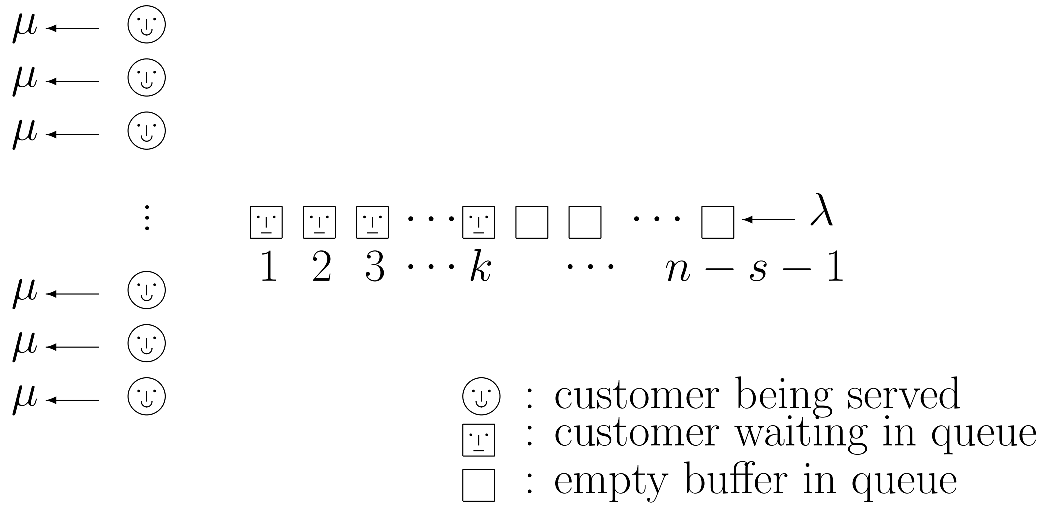

Figure 10.1 The multiple server queue.

- To describe the queueing system, we use the number of customers in the queueing system to represent the state of the system.

- There are n possible states (number of customers) , namely $0, 1, \ldots, n-1$.

- The Markov chain for the queueing system is given in Figure 11.2.

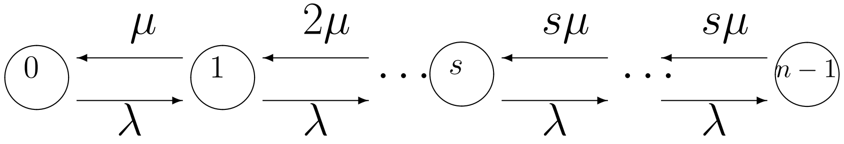

Figure 10.2 The Markov chain for the M/M/$s/n-s-1$ queue.

If we order the states of the system in increasing number of customers then it is not difficult to show that the generator matrix for this queueing system is given by the following $n \times n$ tri-diagonal matrix:

$$A_3 = \begin{pmatrix} -\lambda & \mu & & & & & 0 \\ \lambda & -\lambda-\mu & 2\mu & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \lambda & -\lambda-s\mu & s\mu & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \lambda & -\lambda-s\mu & s\mu \\ 0 & & & & & \lambda & -s\mu \end{pmatrix}. \qquad (9)$$

10.1 A Two-server Queueing System

Let us consider a small size example the M/M/2/2 queue.

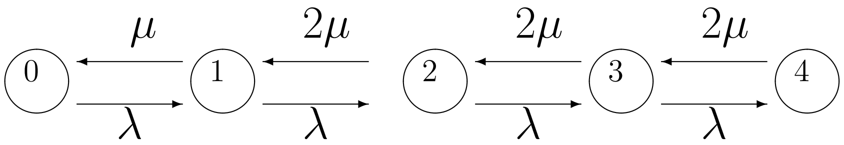

Figure 10.3 The Markov Chain for the M/M/2/2 Queue.

- The generator matrix is an 5×5 matrix.

$$A_4 = \begin{pmatrix} -\lambda & \mu & & & 0 \\ \lambda & -\lambda-\mu & 2\mu & & \\ & \lambda & -\lambda-2\mu & 2\mu & \\ & & \lambda & -\lambda-2\mu & 2\mu \\ 0 & & & \lambda & -2\mu \end{pmatrix}. \tag{10}$$

Let the steady state probability distribution be

$$\mathbf{p} = (p_0, p_1, p_2, p_3, p_4)^t.$$

In steady state we have $A_4\mathbf{p} = \mathbf{0}$.

From the first equation $-\lambda p_0 + \mu p_1 = 0$, we have

$$p_1 = \frac{\lambda}{\mu} p_0.$$

From the second equation $\lambda p_0 - (\lambda + \mu) p_1 + 2\mu p_2 = 0$, we have

$$p_2 = \frac{\lambda^2}{2!\mu^2} p_0.$$

From the third equation $\lambda p_1 - (\lambda + 2\mu) p_2 + 2\mu p_3 = 0$, we have

$$p_3 = \frac{\lambda^3}{2 \cdot 2!\mu^3} p_0.$$

Finally from the fourth equation $\lambda p_2 - (\lambda + 2\mu) p_3 + 2\mu p_4 = 0$, we have

$$p_4 = \frac{\lambda^4}{2^2 \cdot 2!\mu^4} p_0.$$

The last equation is not useful as A_4 is singular.

To determine p_0 we make use of the fact that

$$p_0 + p_1 + p_2 + p_3 + p_4 = 1.$$

Therefore

$$p_0 + \frac{\lambda}{\mu} p_0 + \frac{\lambda^2}{2!\mu^2} p_0 + \frac{\lambda^3}{2 \cdot 2!\mu^3} p_0 + \frac{\lambda^4}{2^2 2!\mu^4} p_0 = 1.$$

Let $\tau = \lambda/(2\mu)$, we have

$$p_0 = (1 + \frac{\lambda}{\mu} + (\frac{\lambda^2}{2!\mu^2})\frac{1-\tau^3}{1-\tau})^{-1}, \quad p_1 = \frac{\lambda}{\mu}p_0, \quad \text{and} \quad p_i = p_0(\frac{\lambda^2}{2!\mu^2})\tau^{i-2}, i = 2, 3, 4.$$

The result above can be further extended to the M/M/2/k queue as follows:

$$p_0 = (1 + \frac{\lambda}{\mu} + (\frac{\lambda^2}{2!\mu^2})\frac{1-\tau^{k+1}}{1-\tau})^{-1}, \quad p_1 = \frac{\lambda}{\mu}p_0, \quad \text{and} \quad p_i = p_0(\frac{\lambda^2}{2!\mu^2})\tau^{i-2}, i = 2, \ldots, k+2.$$

The queueing system has finite number of waiting space.

The result above can also be further extended to M/M/2/∞ queue when $\tau = \lambda/(2\mu) < 1$ as follows:

$$p_0 = \left(1 + \frac{\lambda}{\mu} + (\frac{\lambda^2}{2!\mu^2})\frac{1}{1-\tau}\right)^{-1}, \quad p_1 = \frac{\lambda}{\mu}p_0, \quad \text{and} \quad p_i = p_0(\frac{\lambda^2}{2!\mu^2})\tau^{i-2}, i = 2, 3, \ldots.$$

or

$$p_0 = \frac{1-\tau}{1+\tau}, \quad \text{and} \quad p_i = 2p_0\tau^i, i = 1, 2, \ldots.$$

The queueing system has infinite number of waiting space.

We then derive the expected number of customers in the system.

10.2 Expected Number of Customers in the M/M/2/∞ Queue

The expected number of customers in the M/M/2/∞ queue is given by

$$L_s = \sum_{k=1}^{\infty} k p_k = \frac{1-\tau}{1+\tau} \sum_{k=1}^{\infty} 2k\tau^k.$$

Now we let

$$S = \sum_{k=1}^{\infty} k\tau^k = \tau + 2\tau^2 + \ldots$$

and we have

$$\tau S = \tau^2 + 2\tau^3 + \ldots + .$$

Therefore by subtraction we get

$$(1-\tau)S = \tau + \tau^2 + \tau^3 + \ldots = \frac{\tau}{1-\tau}$$

and

$$S = \frac{\tau}{(1-\tau)^2}.$$

We have

$$L_s = \frac{2\tau}{1-\tau^2}. \tag{11}$$

11 Multiple-Server Queues and Birth-and-Death Process

In this section, we consider queueing models with Poisson input, independent, identically distributed, exponential service times and s parallel servers.

• Specifically, we shall consider two different queue disciplines, namely **Blocked Customers Cleared (BCC)** and **Block Customers Delayed (BCD)**. In the following two subsections, we assume that the Poisson input has rate λ and the exponential service times have mean μ^{-1} .

11.1 Blocked Customers Cleared (Erlang loss system)

The queueing system has s **servers** and there is **no waiting space** and we assume blocked customers are cleared. Total possible number of states is $s+1$ and the generator matrix for this system is given by

$$A_5 = \begin{pmatrix} -\lambda & \mu & & & & 0 \\ \lambda & -\lambda-\mu & 2\mu & & & \\ & \lambda & -\lambda-2\mu & & & \\ & & \ddots & \ddots & \ddots & \\ & & & \lambda & -\lambda-(s-1)\mu & s\mu \\ 0 & & & & \lambda & -s\mu \end{pmatrix}.$$

• Let p_i be the steady state probability that there are i customers in the queueing system. Then by solving $A_5\mathbf{p} = \mathbf{0}$ with $\displaystyle\sum_{i=0}^{s} p_i = 1$ one can get

$$\left\{\begin{aligned} p_j &= \frac{(\lambda/\mu)^j/j!}{\displaystyle\sum_{k=0}^{s}(\lambda/\mu)^k/k!} \qquad (j = 0, 1, \cdots, s) \\ &= \frac{a^j/j!}{\displaystyle\sum_{k=0}^{s} a^k/k!} \end{aligned}\right. \tag{12}$$

and $p_j = 0$ for $j > s$; where $a = \lambda/\mu$ is the offered load.

- This distribution is called the **truncated Poisson distribution** (also called **Erlang loss distribution**).

- On the other hand one can identify this system as a birth-and-death process, we proceed to find p_j. Since customers arrive at random with rate λ, but affect state changes only when $j < s$ (BCC), the arrival rates (the birth rates) are

$$\lambda_j = \begin{cases} \lambda & \text{when } j = 0, \cdots, s-1 \\ 0 & \text{when } j = s \end{cases}$$

Since service times are exponential, the service completion rates (the death rates) are

$$\mu_j = j\mu \qquad (j = 0, 1, 2, \cdots, s).$$

Remarks:

(1) The proportion of customers who have requested for service but are cleared from the system (when all servers are busy) is given by p_s which is also called the **Erlang loss formula** and is denoted by

$$\boxed{B(s,a) = \frac{a^s/s!}{\sum\limits_{k=0}^{s}(a^k/k!)}.}$$

(2) The mean number of busy servers, which is also equal to the mean number of customers completing service per mean service time, is given by the carried load

$$a' = \sum_{j=1}^{s} j p_j .$$

An interesting relation can be derived between the Erlang loss formula and the carried load:

$$\boxed{\begin{aligned} a' &= \sum_{j=1}^{s} j(a^j/j!)/\sum_{k=0}^{s}(a^k/k!) \\ &= a\left(\sum_{j=0}^{s-1}(a^j/j!)/\sum_{k=0}^{s}(a^k/k!)\right) = a\left(1 - B(s,a)\right). \end{aligned}}$$

This shows that the carried load is the **portion** of the offered load that is not lost (captured) from the system.

11.2 Blocked Customers Delayed (Erlang delay system)

The queueing system has s **servers** and there is **infinite many waiting space** and we assume blocked customers are delayed.

- In this case we have the arrival rates $\lambda_j = \lambda \quad (j = 0, 1, \cdots)$, and the service completion rates

$$\mu_j = \begin{cases} j\mu & (j = 0, 1, \cdots, s) \\ s\mu & (j = s, s+1, \cdots). \end{cases}$$

- Hence we have

$$\boxed{p_j = \begin{cases} \frac{a^j}{j!} p_0 & (j = 0, 1, \cdots, s) \\ \frac{a^j}{s! s^{j-s}} p_0 & (j = s+1, \cdots) \end{cases}}$$

where $a = \lambda/\mu$ and

$$\boxed{p_0 = \left(\sum_{k=0}^{s-1} \frac{a^k}{k!} + \sum_{k=s}^{\infty} \frac{a^k}{s! s^{k-s}} \right)^{-1}.}$$

If $a < s$, the infinite geometric sum on the right converges, and

$$\boxed{p_0 = \left(\sum_{k=0}^{s-1} \frac{a^k}{k!} + \frac{a^s}{(s-1)!(s-a)} \right)^{-1}.}$$

• If $a \geq s$, the infinite geometric sum **diverges to infinity**. Then $p_0 = 0$ and hence $p_j = 0$ for all finite j. For $a \geq s$, therefore the queue length tends to infinity with probability 1 as time increases. In this case we say that no statistical equilibrium distribution exists.

Remarks:

(i) The probability that all servers are occupied (as observed by an outside observer) is given by the **Erlang delay formula**

$$\boxed{C(s,a) = \sum_{j=s}^{\infty} p_j = \frac{a^s}{(s-1)!}\frac{1}{s-a}p_0 = \frac{a^s/[(s-1)!(s-a)]}{\left(\sum\limits_{k=0}^{s-1} a^k/k!\right) + a^s/[(s-1)!(s-a)]}.}$$

Since the arriving customer's distribution is equal to the outside observer's distribution, the probability that an arriving customer finds all servers busy (equivalently the probability that the waiting time in the queue $w > 0$) is also given by $C(s,a)$.

(ii) The carried load is equal to the offered load since no request for service has been cleared from the system without being served. In fact, this equality holds for BCD queues with arbitrary arrival and service time distributions.

(iii) Suppose that an arriving customer finds that all the servers are busy. What is the probability that he finds j customers waiting in the 'queue' ?

- This is equivalent to find the conditional probability $P\{Q = j|w > 0\}$ where Q denotes the number of customers waiting in the queue.

- By the definition of conditional probability,

$$\boxed{P\{Q = j|w > 0\} = \frac{P\{Q = j, w > 0\}}{P\{w > 0\}}.}$$

- Thus

$$\boxed{P\{Q = j \text{ and } w > 0\} = P_{s+j} = \frac{a^s}{s!}\left(\frac{a}{s}\right)^j p_0,}$$

we get the **Geometric distribution**

$$\boxed{P\{Q = j|w > 0\} = \frac{\frac{a^s}{s!}\left(\frac{a}{s}\right)^j p_0}{\frac{a^s}{s!}\left(\frac{s}{s-a}\right) p_0} = (1-\rho)\rho^j \quad (j = 0, 1 \ldots).}$$

where $\rho = a/s$ is the traffic intensity.

(iv) Suppose that an arriving customer (test customer) finds all servers are busy and there are j customers waiting in the queue. What is the probability distribution of his waiting time?

- Let X_1 be the elapsed time from $t = 0$ until the customer at the head of the queue enters service and X_i be the length of time that the ith customer in the queue spends at the head of the queue.

- Clearly the test customer's waiting time is equal to

$$X_1 + X_2 + \cdots + X_{j+1}.$$

- Each X_j is equal to the shortest of the s service times then in progress. Because the service times are independent, identical, exponential random variables with mean μ^{-1}, the duration of time from an arbitrary instant until the completion of the shortest remaining service time is also exponentially distributed with mean $(\mathbf{s}\mu)^{-\mathbf{1}}$. Hence

$$X_1 + X_2 + \cdots + X_{j+1}$$

has the $(j+1)$**-phase Erlangian distribution:**

$$\boxed{P\{w > t|w > 0, Q = j\} = P\{X_1 + \cdots + X_{j+1} > t\} = \sum_{i=0}^{j} \frac{(s\mu t)^i}{i!} e^{-s\mu t}.}$$

(v) If an arriving customer finds all servers are busy, the probability that his waiting time in the queue is greater than t is given by

$$\begin{aligned} P\{w > t | w > 0\} &= \sum_{j=0}^{\infty} P\{w > t | w > 0, Q = j\} P\{Q = j | w > 0\} \\ &= \sum_{j=0}^{\infty} \sum_{i=0}^{j} \frac{(s\mu t)^i}{i!} e^{-s\mu t} (1-\rho)\rho^j \\ &= \sum_{i=0}^{\infty} \sum_{j=i}^{\infty} \frac{(s\mu t)^i}{i!} e^{-s\mu t} (1-\rho)\rho^j = e^{-(1-\rho)s\mu t}. \end{aligned}$$

From this, we obtain

$$P\{w > t\} = P\{w > t | w > 0\} P\{w > 0\} = C(s,a) e^{-(1-\rho)s\mu t}.$$

(vi) It can be easily derived that the mean waiting time in the queue

$$\boxed{W_q = E(w) = \int_0^{\infty} P(w > t) dt = \int_0^{\infty} C(s,a) e^{-(1-\rho)s\mu t} dt = \frac{C(s,a)}{(1-\rho)s\mu}}$$

and

$$\boxed{E(w | w > 0) = \int_0^{\infty} P(w > t | w > 0) dt = \int_0^{\infty} e^{-(1-\rho)s\mu t} dt = \frac{1}{(1-\rho)s\mu},}$$

The mean waiting time in the system (including service time) is given by

$$\boxed{W_s = W_q + \mu^{-1} = \frac{C(s,a) + (1-\rho)s}{(1-\rho)s\mu}.}$$

Note:

$$\begin{aligned}
E(x) &= \lim_{k\to\infty} \int_0^k xf(x)dx = \lim_{k\to\infty} \int_0^k xdF(x) \\
&= \lim_{k\to\infty} (xF(x)|_0^k - \int_0^k F(x)dx) \\
&= \lim_{k\to\infty} (kF(k) - \int_0^k F(x)dx) \\
&= \lim_{k\to\infty} (kF(k) - k + k - \int_0^k F(x)dx) \\
&= \lim_{k\to\infty} (kF(k) - k + \int_0^k 1 - F(x)dx) \\
&= \lim_{k\to\infty} (0 + \int_0^k P(t > x)dx) = \int_0^\infty P(t > x)dx.
\end{aligned}$$

Here we assume that

$$\lim_{k\to\infty} k(F(k) - 1) = \lim_{k\to\infty} \frac{\int_k^\infty f(x)dx}{-1/k} = \lim_{k\to\infty} -k^2 f(k) = 0.$$

The assumption is reasonable because for $\int_0^\infty xf(x)dx$ to be finite, $f(x)$ should tend to zero 'faster' than x^{-2} as $x \to \infty$.

12 Little's Queueing Formula

If λ is the mean arrival rate, W is the mean time spent in the system (mean sojourn time) and L is the mean number of customers present, J.D.C. Little proved in 1961 that

$$\boxed{L = \lambda W.}$$

This result is one of the most general and useful results in queueing theory for a **blocked customer delay queue**. The formal proof of this theorem is too long for this course. Let us just formally state the theorem and then give a heuristic proof.

Proposition 1 (**Little's Theorem**) Let $L(x)$ be the number of customers present at time x, and define the mean number L of customers present throughout the time interval $[0, \infty)$ as

$$L = \lim_{t\to\infty} \frac{1}{t} \int_0^t L(x)dx;$$

let $N(t)$ be the number of customers who arrive in $[0, t]$, and define the arrival rate λ as

$$\lambda = \lim_{t\to\infty} \frac{N(t)}{t};$$

and let W_i be the sojourn time of the ith customer, and define the mean sojourn time W as

$$W = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} W_i.$$

If λ and W exist and are finite, then so does L, and they are related by $\lambda W = L$.

Proof: Let us follow the heuristic argument suggested by P. J. Burke.

• Assume that the mean values L and W exist, and consider a long time interval $(0, t)$ throughout which **statistical equilibrium (steady state) prevails**.

• The mean number of customers who enter the system during this interval is λt.

Imagine that a sojourn time is associated with each arriving customer; i.e., each arrival brings a sojourn time with him. Thus the average sojourn time brought into the system during $(0, t)$ is $\lambda t W$.

On the other hand, each customer present in the system uses up his sojourn time linearly with time. If L is the average number of customers present throughout $(0, t)$, then Lt is the average amount of time used up in $(0, t)$.

Now as $t \to \infty$ the accumulation of sojourn time must equal the amount of sojourn time used up; that is,

$$\boxed{\lim_{t\to\infty} \frac{\lambda t W}{Lt} = 1.}$$

$\square$

With the help of Little's formula, we get the following useful results:

(a) λ, the average number of arrivals entering the system,

(b) L_s, the average number of customers in the queueing system,

(c) L_q, the average number of customers waiting in the queue,

(d) L_c, the average number of customers in the server,

(e) W_s, the average time a customer spends in the queueing system,

(f) W_q, the average time a customer spends in waiting in the queue,

(g) W_c, the average time a customer spends in the server.

then the Little's formula states that if the steady state probability distribution exists, we have

$$L_s = \lambda W_s, \quad L_q = \lambda W_q, \quad \text{and} \quad L_c = \lambda W_c.$$

12.1 Little's queueing Formula for the M/M/1/∞ Queue

In the following, we are going to prove Little's queueing formula for the case of M/M/1/∞ queue. We recall that

$$L_s = \frac{\rho}{1-\rho}, \quad L_q = \frac{\rho^2}{1-\rho}, \quad L_c = \rho, \quad L_s = L_q + L_c, \quad \rho = \frac{\lambda}{\mu}.$$

• We first note that the expected waiting time W_c at the server is $1/\mu$.

Therefore we have

$$W_c = \frac{1}{\mu} = \frac{\lambda}{\lambda\mu} = \frac{L_c}{\lambda}.$$

• Secondly we note that when a customer arrived, there can be i customers already in the system. The expected waiting time before joining the server when there are already i customers in the system is of course i/μ. Because there is only server and the mean service time of each customer in front of him is $1/\mu$.

• Therefore the expected waiting time W_q before one joins the server will be

$$\sum_{i=1}^{\infty} p_i \left(\frac{i}{\mu}\right) = \frac{1}{\mu}\sum_{i=1}^{\infty} i p_i = \frac{L_s}{\mu} = \frac{\rho}{(1-\rho)\mu}.$$

- Since i can be $0, 1, 2, \ldots$, we have

$$W_q = \frac{\rho}{(1-\rho)\mu} = \frac{\rho^2}{(1-\rho)\mu\rho} = \frac{L_q}{\lambda}$$

- The expected waiting time at the server W_c will be of course $1/\mu$. Thus we have

$$\begin{aligned} W_s &= W_q + W_c \\ &= \frac{L_q}{\mu} + \frac{1}{\mu} \\ &= \frac{1}{\mu}(\frac{\rho}{1-\rho} + 1) \\ &= \frac{1}{\mu(1-\rho)} \\ &= \frac{\rho}{\lambda(1-\rho)} \\ &= \frac{L_s}{\lambda}. \end{aligned}$$

Here

$$\rho = \lambda/\mu$$

and

$$L_s = \rho/(1-\rho).$$

12.2 Applications of the Little's queueing Formula

Arrival rate	λ
Service rate	μ
Traffic intensity	$\rho = \lambda/\mu$
Probability that no customer in the queue	$p_0 = 1 - \rho$
Probability that i customers in the queue	$p_i = p_0 \rho^i$
Probability that an arrival has to wait for service	$1 - p_0 = \rho$
Expected number of customers in the system	$L_s = \rho/(1-\rho)$
Expected number of customers in the queue	$L_q = \rho^2/(1-\rho)$
Expected number of customers in the server	$L_c = \rho$
Expected waiting time in the system	$L_s/\lambda = 1/(1-\rho)\mu$
Expected waiting time in the queue	$L_q/\lambda = \rho/(1-\rho)\mu$
Expected waiting time in the server	$L_c/\lambda = 1/\mu$

Table 12.1: A summary of the M/M/1/∞ queue.

Example 1 Consider the M/M/2/∞ queue with arrival rate λ and service rate μ. What is the expected waiting time for a customer in the system?

We recall that the expected number of customers L_s in the system is given by

$$L_s = \frac{2\rho}{1-\rho^2}.$$

Here $\rho = \lambda/(2\mu)$. By applying the Little's queueing formula we have

$$W_s = \frac{L_s}{\lambda} = \frac{1}{\mu(1-\rho^2)}.$$

Example 2 On average 30 patients arrive each hour to the health centre. They are first seen by the receptionist, who takes an average of 1 min to see each patient. If we assume that the M/M/1 queueing model can be applied to this problem, then we can calculate the average measure of the system performance, see Table 2.

Arrival rate	$\lambda = 30$ (per hour)
Service rate	$\mu = 60$ (per hour)
Traffic intensity	$\rho = 0.5$
Probability that no customer in the queue	$p_0 = 0.5$
Probability that i customers in the queue	$p_i = 0.5^{i+1}$
Probability that an arrival has to wait for service	0.5
Expected number of customers in the system	$L_s = 1$
Expected number of customers in the queue	$L_q = 0.5$
Expected number of customers in the server	$L_c = 0.5$
Expected waiting time in the system	$L_s/\lambda = 1/30$
Expected waiting time in the queue	$L_q/\lambda = 1/60$
Expected waiting time in the server	$L_c/\lambda = 1/60$

Table 12.2: A summary of the system performance

13 Allocation of the Arrivals in a System of M/M/1 Queues

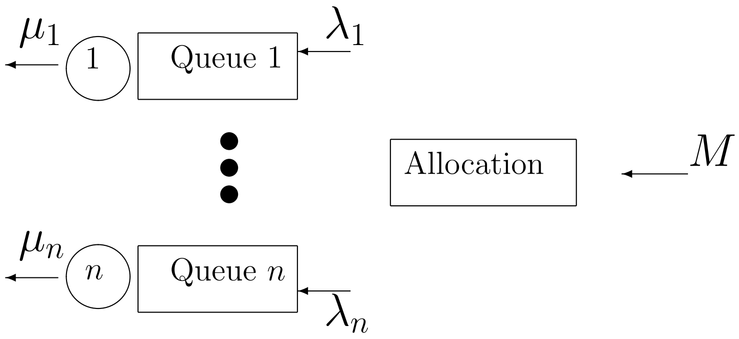

Figure 13.1 The Queueing System with Allocation of Arrivals.

- We consider a queueing system consisting of n independent M/M/1 queues. The service rate of the serve at the ith queue is μ_i.

- The arrival process is a Poisson process with rate M.

- An allocation process is implemented such that it diverts an arrived customers to queue i with probability

$$\boxed{\frac{\lambda_i}{\lambda_1 + \ldots + \lambda_n} = \frac{\lambda_i}{M}.}$$

Then the input process of queue i is a Poisson process with rate λ_i.

- The objective here is to find the parameters λ_i such that some system performance is optimized. We remark that we must have $\lambda_i < \mu_i$.

13.1 Minimizing Number of Customers in the System

The expected number of customers in queue i is

$$\frac{\lambda_i/\mu_i}{1-\lambda_i/\mu_i}.$$

The total expected number of customers in the system is

$$\sum_{i=1}^{n} \frac{\lambda_i/\mu_i}{1-\lambda_i/\mu_i}.$$

The optimization problem is then given as follows:

$$\min_{\lambda_i} \left\{ \sum_{i=1}^{n} \frac{\lambda_i/\mu_i}{1-\lambda_i/\mu_i} \right\}.$$

subject to

$$\sum_{i=1}^{m} \lambda_i = M$$

and

$$0 \le \lambda_i < \mu_i \quad \text{for} \quad i = 1, 2, \ldots, n.$$

By consider the Lagrangian function

$$L(\lambda_1, \ldots, \lambda_n, m) = \sum_{i=1}^{n} \frac{\lambda_i/\mu_i}{1 - \lambda_i/\mu_i} - m \left(\sum_{i=1}^{n} \lambda_i - M \right)$$

and solving

$$\frac{\partial L}{\partial \lambda_i} = 0 \quad \text{and} \quad \frac{\partial L}{\partial m} = 0$$

we have the optimal solution

$$\lambda_i = \mu_i \left(1 - \frac{1}{\sqrt{m\mu_i}} \right) < \mu_i$$

where

$$m = \left(\frac{\sum_{i=1}^{n} \sqrt{\mu_i}}{\sum_{i=1}^{n} \mu_i - M} \right)^2 .$$

13.2 Minimizing Number of Customers Waiting in the System

The expected number of customers waiting in queue i is

$$\frac{(\lambda_i/\mu_i)^2}{1-\lambda_i/\mu_i}.$$

The total expected number of customers waiting in the system is

$$\sum_{i=1}^{n}\frac{(\lambda_i/\mu_i)^2}{1-\lambda_i/\mu_i}.$$

The optimization problem is then given as follows:

$$\min_{\lambda_i}\left\{\sum_{i=1}^{n}\frac{(\lambda_i/\mu_i)^2}{1-\lambda_i/\mu_i}\right\}.$$

subject to

$$\sum_{i=1}^{m}\lambda_i = M$$

and

$$0 \le \lambda_i < \mu_i \quad \text{for} \quad i = 1, 2, \ldots, n.$$

By consider the Lagrangian function

$$L(\lambda_1, \ldots, \lambda_n, m) = \sum_{i=1}^{n} \frac{(\lambda_i/\mu_i)^2}{1 - \lambda_i/\mu_i} - m \left(\sum_{i=1}^{n} \lambda_i - M \right)$$

and solving

$$\frac{\partial L}{\partial \lambda_i} = 0 \quad \text{and} \quad \frac{\partial L}{\partial m} = 0$$

we have the optimal solution

$$\lambda_i = \mu_i \left(1 - \frac{1}{\sqrt{1 + m\mu_i}} \right) < \mu_i$$

where m is the solution of

$$\sum_{i=1}^{n} \mu_i \left(1 - \frac{1}{\sqrt{1 + m\mu_i}} \right) = M.$$

14 Applications of Queues

We are going to look at one application of queueing systems. In a large machine repairing company, workers must get their tools from the tool centre which is managed by an operator. Suppose the mean number of workers seeking for tools per hour is 5 and each worker is paid 8 dollars per hour.

14.1 Which Operator to Employ?

There are two possible operators (A and B) to employ. In average Operator A takes 10 minutes to handle one request for tools is paid 5 dollars per hour. While Operator B takes 11 minutes to handle one request for tools is paid 3 dollars per hour.

• Assume that the inter-arrival time of workers and the processing time of the operators are exponentially distributed. One may regard the request for tools as a queueing process (An $M/M/1/\infty$) where the arrival rate $\lambda = 5$ per hour.

• For **Operator A**, the service rate is $\mu = 60/10 = 6$ per hour. Thus we have

$$\rho = \lambda/\mu = 5/6.$$

• The expected number of workers waiting for tools at the tool centre will be

$$\frac{\rho}{1-\rho} = \frac{5/6}{1-5/6} = 5.$$

- The expected delay cost of the workers is

$$5 \times 8 = 40$$

dollars per hour and the operator cost is 5 dollars per hour. Therefore the total expected cost is

$$\mathbf{40 + 5 = 45}.$$

- For **Operator B**, the service rate is $\mu = 60/11$ per hour. Thus we have

$$\rho = \lambda/\mu = 11/12.$$

The expected number of workers waiting for tools at the tool centre will be

$$\frac{\rho}{1-\rho} = \frac{11/12}{1-11/12} = 11.$$

The expected delay cost of the workers is

$$11 \times 8 = 88$$

dollars per hour and the operator cost is 3 dollars per hour. Therefore the total expected cost is

$$88 + 3 = 91.$$

- **Conclusion:** Operator A should be employed.

14.2 Two M/M/1 Queues Or One M/M/2 Queue ?

If one more identical operator can be employed, then which of followings is better ? (In our analysis, we assume that $\lambda < \mu$).

(i) Put the two operators separately. Therefore we have two M/M/1/∞ queues. In this case, we assume that an arrived customer can either join the first queue or the second with equal chance.

(ii) Put the two operators together. Therefore we have an M/M/2/∞ queue.

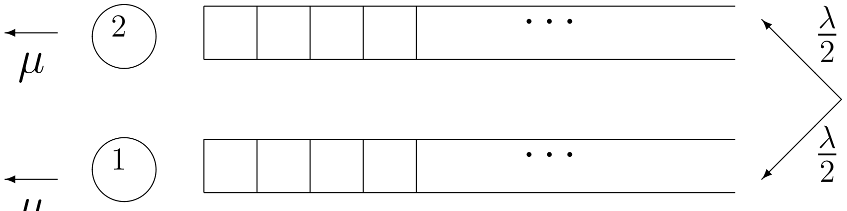

Figure 14.1 Case (i) Two M/M/1/∞ Queues.

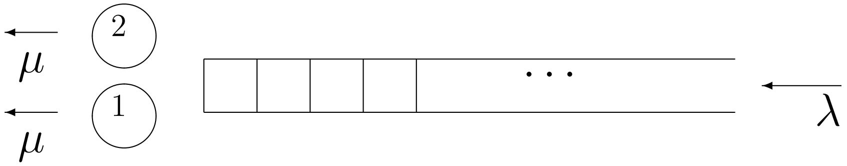

Figure 14.2 Case (ii) One M/M/2/∞ Queue.

To determine which case is better, we calculate the expected number of customers (workers) in both cases. Clearly in our consideration, the smaller the better (Why?).

In case (i), the expected number of customers in any one of the queues will be given by

$$\frac{(\frac{\lambda}{2\mu})}{1-(\frac{\lambda}{2\mu})}.$$

Hence the total expected number of customers (workers) in the system is

$$\boxed{S_1 = 2 \times \frac{(\frac{\lambda}{2\mu})}{1-(\frac{\lambda}{2\mu})} = \frac{(\frac{\lambda}{\mu})}{1-(\frac{\lambda}{2\mu})}.}$$

In case (ii), the expected number of customers in the system will be given by (see previous section)

$$\boxed{S_2 = \frac{(\frac{\lambda}{\mu})}{1-(\frac{\lambda}{2\mu})^2}.}$$

Clearly $S_2 < S_1$.

- **Conclusion:** Case (ii) is better. We should put all the servers (operators) together.

14.3 One More Operator?

Operator A later complains that he is overloaded and the workers have wasted their time in waiting for a tool. To improve this situation, the senior manager wonders if it is cost effective to employ one more identical operator at the tool centre. Assume that the inter-arrival time of workers and the processing time of the operators are exponentially distributed.

• For the present situation, one may regard the request for tools as a queueing process (An M/M/1/∞) where the arrival rate $\lambda = 5$ per hour and the service rate $\mu = 60/10 = 6$ per hour. Thus we have $\rho = \lambda/\mu = 5/6$.

• The expected number of workers waiting for tools at the tool centre will be

$$\frac{\rho}{1-\rho} = \frac{5/6}{1-5/6} = 5.$$

The expected delay cost of the workers is $5 \times 8 = 40$ dollars per hour and the operator cost is 5 dollars per hour. Therefore the total expected cost is $40 + 5 = 45$ dollars.

When one extra operator is added then there are 2 identical operators at the tool center and this will be an M/M/2/∞ queue.

- The expected number of workers in the system is given by (c.f. (11))

$$\frac{1-\rho}{1+\rho}\sum_{i=1}^{\infty} 2i\rho^{i} = \frac{2\rho}{1-\rho^2}$$

where

$$\rho = \frac{\lambda}{2\mu} = \frac{5}{12}.$$

- In this case the expected delay cost and the operator cost will be given respectively by

$$\frac{8 \times 2\rho}{1-\rho^2} = \frac{8 \times 120}{119} = 8.07 \quad \text{and} \quad 2 \times 5 = 10 \text{ dollars.}$$

- Thus the expected cost when there are 2 operators is given by 18.07 dollars.

- **Conclusion:** Hence the senior manager should employ one more operator.

How about employing **three** operators? (You may consider M/M/3/∞ queue). But it is clear that there is no need to employ **four** operators. Why?

15 More Applications of Queueing Systems

15.1 An Unreliable Machine System

Consider an unreliable machine system. The normal time of the machine is exponentially distributed with mean λ^{-1}. Once the machine is broken, it is subject to a **n-phase repairing process**. The repairing time at phase i is also exponentially distributed with mean $\mu_i^{-1}(i = 1, 2, \ldots, n)$. After the repairing process, the machine is back to normal. Let 0 be the state that the machine is normal and i be the state that the machine is in repairing phase i. Then the Markov chain of the model is given by

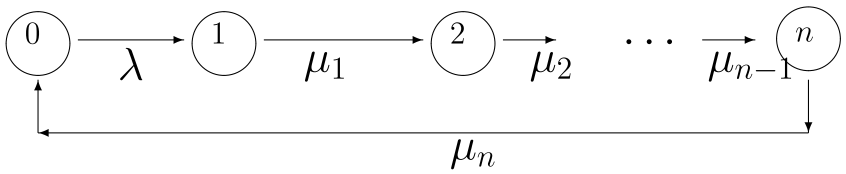

Figure 15.1 The Markov Chain for the Unreliable Machine System.

- Let the steady state probability vector be $\mathbf{p} = (p_0, p_1, \ldots, p_n)$ satisfies $A_6\mathbf{p} = \mathbf{0}$ where

$$A_6 = \begin{pmatrix} -\lambda & 0 & & & \mu_n \\ \lambda & -\mu_1 & & & \\ & \mu_1 & -\mu_2 & & \\ & & \ddots & \ddots & 0 \\ 0 & & & \mu_{n-1} & -\mu_n \end{pmatrix}.$$

From the first equation $-\lambda p_0 + \mu_n p_n$ we have

$$p_n = \frac{\lambda}{\mu_n} p_0.$$

From the second equation $\lambda p_0 - \mu_1 p_1$ we have

$$p_1 = \frac{\lambda}{\mu_1} p_0.$$

From the third equation $\mu_1 p_1 - \mu_2 p_2$ we have

$$p_2 = \frac{\lambda}{\mu_2} p_0.$$

We continue this process and therefore

$$p_i = \frac{\lambda}{\mu_i} p_0.$$

Since $p_0 + p_1 + p_2 + \ldots + p_n = 1$, we have

$$p_0 \left(1 + \sum_{i=1}^{n} \frac{\lambda}{\mu_i} \right) = 1.$$

Therefore

$$p_0 = \left(1 + \sum_{i=1}^{n} \frac{\lambda}{\mu_i} \right)^{-1}.$$

15.2 A Reliable One-machine Manufacturing System

Here we consider an Markovian model of reliable one-machine manufacturing system.

• The production time for one unit of product is exponentially distributed with a mean time of μ^{-1}.

• The inter-arrival time of a demand is also exponentially distributed with a mean time of λ^{-1}.

• The demand is served in a first come first serve manner. In order to retain the customers, there is no backlog limit in the system. However, there is an upper limit $n (n \geq 0)$ for the inventory level.

• The machine keeps on producing until this inventory level is reached and the production is stopped once this level is attained.

• We seek for the optimal value of n (**the hedging point or the safety stock**) which minimizes the expected running cost.

• The running cost consists of a deterministic inventory cost and a backlog cost. In fact, the optimal value of n is the best amount of inventory to be kept in the system so as to hedge against the fluctuation of the demand.

Let us summarized the notations as follows.

I, the unit inventory cost;

B, the unit backlog cost;

$n \geq 0$, the hedging point;

μ^{-1}, the mean production time for one unit of product;

λ^{-1}, the mean inter-arrival time of a demand.

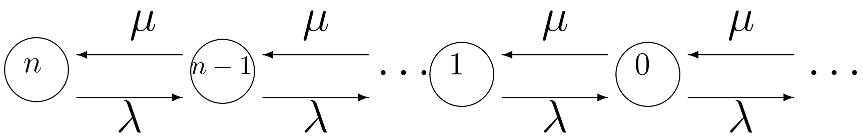

Figure 15.2 The Markov Chain (M/M/1 Queue) for the Manufacturing System.

If the inventory level (negative inventory level means backlog) is used to represent the state of the system, one may write down the Markov chain for the system as follows.

Here we assume that $\mu > \lambda$, so that the steady state probability distribution of the above M/M/1 queue exists and has analytic solution

$$q(i) = (1-p)p^{n-i}, \quad i = n, n-1, n-2, \cdots$$

where

$$p = \lambda/\mu$$

and $q(i)$ is the steady state probability that the inventory level is i.

Hence the expected running cost of the system (sum of the inventory cost and the backlog cost) can be written down as follows:

$$\boxed{E(n) = \underbrace{I\sum_{i=0}^{n}(n-i)(1-p)p^{i}}_{\text{inventory cost}} + \underbrace{B\sum_{i=n+1}^{\infty}(i-n)(1-p)p^{i}}_{\text{backlog cost}}.} \qquad (13)$$

Proposition 2 *The expected running cost $E(n)$ is minimized if the hedging point n is chosen such that*

$$p^{n+1} \le \frac{I}{I+B} \le p^n.$$

Proof: We note that

$$E(n-1) - E(n) = B - (I+B)(1-p)\sum_{i=0}^{n-1} p^i = -I + (I+B)p^n$$

and

$$E(n+1) - E(n) = -B + (I+B)(1-p)\sum_{i=0}^{n} p^i = I - (I+B)p^{n+1}.$$

Therefore we have

$$E(n-1) \ge E(n) \Leftrightarrow p^n \ge \frac{I}{I+B}$$

and

$$E(n+1) \ge E(n) \Leftrightarrow p^{n+1} \le \frac{I}{I+B}.$$

Thus the optimal value of n is the one such that

$$p^{n+1} \le \frac{I}{I+B} \le p^n.$$

□

A Summary on Markovian Queueing Systems

- The Kendall's notations.

- M/M/s/n queue:

- the Markov chain diagram, the generator matrix, the steady state probability.
- Erlang loss formula, Erlang delay formula.
- waiting time distribution in an M/M/s/∞ queue.

- The statement of the Little's queueing formula.

- System performance analysis:

- Expected number of customers.
- Expected number of busy servers.
- Mean waiting time.

- Applications:

- Employment of operators.
- Unreliable machine system.
- Manufacturing system.