PART II

(3) **Continuous Time Markov Chains : Theory and Examples**
- Pure Birth Process with Constant Rates
- Pure Death Process
- More on Birth-and-Death Process
- Statistical Equilibrium

(4) **Introduction to Queueing Systems**
- Basic Elements of Queueing Models
- Queueing Systems of One Server
- Queueing Systems with Multiple Servers
- Little’s Queueing Formula
- Applications of Queues
- An Inventory Model with Returns and Lateral Transshipments
- Queueing Systems with Two Types of Customers
- Queues in Tandem and in Parallel

“**All models are wrong / inaccurate, but some are useful.”**

George Box (Wikipedia).

http://hkumath.hku.hk/~wkc/course/part2.pdf
3 Continuous Time Markov Chains: Theory and Examples

We discuss the theory of birth-and-death processes, the analysis of which is relatively simple and has important applications in the context of queueing theory.

• Let us consider a system that can be represented by a family of random variables \( \{N(t)\} \) parameterized by the time variable \( t \). This is called a stochastic process.

• In particular, let us assume that for each \( t \), \( N(t) \) is a non-negative integral-valued random variable. Examples are the followings.

   (i) a telephone switchboard, where \( N(t) \) is the number of calls occurring in an interval of length \( t \).

   (ii) a queue, where \( N(t) \) is the number of customers waiting or in service at time \( t \).

We say that the system is in state \( E_j \) at time \( t \) if \( N(t) = j \). Our aim is to compute the state probabilities \( P\{N(t) = j\}, j = 0, 1, 2, \ldots \).
Definition 1 A process obeying the following three postulates is called a birth-and-death process:

(1) At any time $t$, $P\{E_j \to E_{j+1} \text{ during } (t, t+h) | E_j \text{ at } t\} = \lambda_j h + o(h)$ as $h \to 0$ ($j = 0, 1, 2, \cdots$). Here $\lambda_j$ is a constant depending on $j$ (State $E_j$).

(2) At any time $t$, $P\{E_j \to E_{j-1} \text{ during } (t, t+h) | E_j \text{ at } t\} = \mu_j h + o(h)$ as $h \to 0$ ($j = 1, 2, \cdots$). Here $\mu_j$ is a constant depending on $j$ (State $E_j$).

(3) At any time $t$, $P\{E_j \to E_{j+k} \text{ during } (t, t+h) | E_j \text{ at } t\} = o(h)$ as $h \to 0$ for $k \geq 2$ ($j = 0, 1, \cdots$).

Figure 2.1: The Birth and Death Process.

Notation: Let $P_j(t) = P\{N(t) = j\}$ and let $\lambda_{-1} = \mu_0 = P_{-1}(t) = 0$. 


• Then it follows from the above postulates that (where $h \to 0; j = 0, 1, \ldots$)

\[
P_j(t + h) = (\lambda_{j-1} h + o(h)) P_{j-1}(t) + (\mu_{j+1} h + o(h)) P_{j+1}(t)
\]

- an arrival

\[
+ [1 - ((\lambda_j + \mu_j) h + o(h))] P_j(t)
\]

- a departure

- no arrival or departure

• Therefore we have

\[
P_j(t + h) = (\lambda_{j-1}) P_{j-1}(t) + (\mu_{j+1}) P_{j+1}(t) + [1 - (\lambda_j + \mu_j) h] P_j(t) + o(h).
\]

• Re-arranging terms, we have

\[
\frac{P_j(t + h) - P_j(t)}{h} = \lambda_{j-1} P_{j-1}(t) + \mu_{j+1} P_{j+1}(t) - (\lambda_j + \mu_j) P_j(t) + \frac{o(h)}{h}.
\]
Letting $h \to 0$, we get the differential-difference equations
\[
\frac{d}{dt} P_j(t) = \lambda_{j-1} P_{j-1}(t) + \mu_{j+1} P_{j+1}(t) - (\lambda_j + \mu_j) P_j(t). \quad (3.1)
\]
At time $t = 0$ the system is in state $E_i$, the initial conditions are
\[
P_j(0) = \delta_{ij} \quad \text{where} \quad \delta_{ij} = \begin{cases} 
1 & \text{if } i = j \\
0 & \text{if } i \neq j.
\end{cases}
\]

**Definition 2** The coefficients $\{\lambda_j\}$ and $\{\mu_j\}$ are called the **birth** and **death rates** respectively.

- When $\mu_j = 0$ for all $j$, the process is called a **pure birth process**;
- and when $\lambda_j = 0$ for all $j$, the process is called a **pure death process**.
- In the case of either a pure birth process or a pure death process, the equations (3.1) can be solved by using recurrence relation.
3.1 Pure Birth Process with Constant Rates

We consider a Pure birth process \( \mu_i = 0 \) with constant \( \lambda_j = \lambda \) and initial state \( E_0 \).

- Equations in (3.1) become

\[
\frac{d}{dt} P_j(t) = \lambda P_{j-1}(t) - \lambda P_j(t) \quad (j = 0, 1, \cdots)
\]

where \( P_{-1}(t) = 0 \) and \( P_j(0) = \delta_{0j} \).

- Here

\[
j = 0, \quad P_0'(t) = -\lambda P_0(t),
\]

hence

\[
P_0(t) = a_0 e^{-\lambda t}.
\]

From the initial conditions, we get \( a_0 = 1 \).
• Inductively, we can prove that if

\[ P_{j-1}(t) = \frac{(\lambda t)^{j-1}}{(j-1)!} e^{-\lambda t} \]

then the equation

\[ P_j'(t) = \lambda \left( \frac{(\lambda t)^{j-1}}{(j-1)!} e^{-\lambda t} \right) - \lambda P_j(t) \]

gives the solution

\[ P_j(t) = \frac{(\lambda t)^j}{j!} e^{-\lambda t}. \quad (3.2) \]

**Remark 1** Probabilities (3.2) satisfy the normalization condition

\[ \sum_{j=0}^{\infty} P_j(t) = 1 \quad (t \geq 0). \]

**Remark 2** For each \( t \), \( N(t) \) is the Poisson distribution, given by \( \{P_j(t)\} \). We say that \( N(t) \) describes a **Poisson process**.

**Remark 3** Since the assumption \( \lambda_j = \lambda \) is often a realistic one, the simple formula (3.2) plays a central role in queueing theory.
3.1.1 Generating Function Approach

Here we demonstrate using the generating function approach for solving the pure birth problem.

- Let \( \{P_i\} \) be a discrete probability density distribution for a random variable \( X \), i.e.,
  \[
P(X = i) = P_i, \quad i = 0, 1, \ldots,
  \]
Recall that the probability generating function is defined as
  \[
g(z) = \sum_{n=0}^{\infty} P_n z^n.
  \]
Let the probability generating function for \( P_n(t) \) be
  \[
g(z, t) = \sum_{n=0}^{\infty} P_n(t) z^n.
  \]
- The idea is that if we can find \( g(z, t) \) and obtain its coefficients when expressed in a power series of \( z \) then one can solve \( P_n(t) \).
From the differential-difference equations, we have

\[
\sum_{n=0}^{\infty} \frac{dP_n(t)}{dt} z^n = z \sum_{n=0}^{\infty} \lambda P_n(t) z^n - \sum_{n=0}^{\infty} \lambda P_n(t) z^n.
\]

Assuming one can inter-change the operation between the summation and the differentiation, then we have

\[
\frac{dg(z, t)}{dt} = \lambda (z - 1) g(z, t)
\]

when \(z\) is regarded as a constant.

- Then we have

\[
g(z, t) = Ke^{\lambda(z-1)t}.
\]

Since

\[
g(z, 0) = \sum_{n=0}^{\infty} P_n(0) z^n = 1
\]

we have \(K = 1\). Hence we have

\[
g(z, t) = e^{-\lambda t} \left( 1 + \lambda z + \frac{(\lambda t z)^2}{2!} + \ldots + \right) = \left( e^{-\lambda t} + e^{-\lambda t} \lambda z + \frac{e^{-\lambda t} (\lambda t)^2}{2!} z^2 + \ldots + \right).
\]

Then the result follows.
3.2 Pure Death Process

We then consider a pure death process with $\mu_j = j\mu$ and initial state $E_n$.

- Equations in (3.1) become

$$\frac{d}{dt} P_j(t) = (j + 1)\mu P_{j+1}(t) - j\mu P_j(t) \quad j = n, n - 1, \cdots, 0$$

(3.3)

where

$$P_{n+1}(t) = 0 \quad \text{and} \quad P_j(0) = \delta_{nj}.$$  

- We solve these equations recursively, starting from the case $j = n$.

$$\frac{d}{dt} P_n(t) = -n\mu P_n(t) , \quad P_n(0) = 1$$

implies that

$$P_n(t) = e^{-n\mu t}.$$
• The equation with \( j = n - 1 \) is

\[
\frac{d}{dt} P_{n-1}(t) = n\mu P_n(t) - (n - 1)\mu P_{n-1}(t) = n\mu e^{-n\mu t} - (n - 1)\mu P_{n-1}(t).
\]

• Solving this differential equation and we get

\[
P_{n-1}(t) = n(e^{-\mu t})^{n-1}(1 - e^{-\mu t}).
\]

• Recursively, we get

\[
P_j(t) = \binom{n}{j} (e^{-\mu t})^j (1 - e^{-\mu t})^{n-j} \quad \text{for} \quad j = 0, 1, \cdots, n. \quad (3.4)
\]

Remark 4 For each \( t \), the probabilities (3.4) comprise a binomial distribution.

Remark 5 The number of equations in (3.3) is finite in number. For a pure birth process, the number of equations is infinite.
3.3 More on Birth-and-Death Process

A simple queueing example is given as follows (An illustration of birth-and-death process in queueing theory context).

- We consider a queueing system with one server and no waiting position, with

\[ P\{\text{one customer arriving during } (t, t+h)\} = \lambda h + o(h) \]

and

\[ P\{\text{service ends in } (t, t+h) | \text{ server busy at } t\} = \mu h + o(h) \text{ as } h \to 0. \]

- This corresponds to a two state birth-and-death process with \( j = 0, 1 \). The arrival rates are \( \lambda_0 = \lambda \) and \( \lambda_j = 0 \) for \( j \neq 0 \) (an arrival that occurs when the server is busy has no effect on the system since the customer leaves immediately); and the departure rates are \( \mu_j = 0 \) when \( j \neq 1 \) and \( \mu_1 = \mu \) (no customers can complete service when no customer is in the system).

![Figure 3.1. The Two-state Birth-and-Death Process.](image-url)
The equations for the birth-and-death process are given by

\[
\frac{d}{dt} P_0(t) = -\lambda P_0(t) + \mu P_1(t) \quad \text{and} \quad \frac{d}{dt} P_1(t) = \lambda P_0(t) - \mu P_1(t).
\] (3.5)

One convenient way of solving this set of simultaneous linear differential equations (not a standard method!) is as follows:

• Adding the two equations in (3.5), we get

\[
\frac{d}{dt} (P_0(t) + P_1(t)) = 0,
\]

hence \( P_0(t) + P_1(t) = \text{constant}. \)

• Initial conditions are \( P_0(0) + P_1(0) = 1; \) thus \( P_0(t) + P_1(t) = 1. \) Hence we get

\[
\frac{d}{dt} P_0(t) + (\lambda + \mu) P_0(t) = \mu.
\]

The solution (exercise) is given by

\[
P_0(t) = \frac{\mu}{\lambda + \mu} + \left( P_0(0) - \frac{\mu}{\lambda + \mu} \right) e^{-(\lambda+\mu)t}.
\]
Since $P_1(t) = 1 - P_0(t)$,

$$P_1(t) = \frac{\lambda}{\lambda + \mu} + \left( P_1(0) - \frac{\lambda}{\lambda + \mu} \right) e^{-(\lambda+\mu)t}. \quad (3.6)$$

- For the three examples of birth-and-death processes that we have considered, the system of differential-difference equations are much simplified and can therefore be solved very easily.

- In general, the solution of differential-difference equations is no easy matter. Here we merely state the properties of its solution without proof.

**Proposition 1** For arbitrarily prescribed coefficients $\lambda_n \geq 0, \mu_n \geq 0$ there always exists a positive solution $\{P_n(t)\}$ of differential-difference equations (3.1) such that

$$\sum P_n(t) \leq 1.$$  

If the coefficients are bounded, this solution is unique and satisfies the regularity condition $\sum P_n(t) = 1$.

**Remark 6** Fortunately in all cases of practical significance, the regularity condition $\sum P_n(t) = 1$ and uniqueness of solution are satisfied.
3.4 Statistical Equilibrium (Steady-State Probability Distribution)

Consider the state probabilities of the above example when \( t \to \infty \), from (3.6) we have

\[
\begin{align*}
    P_0 &= \lim_{t \to \infty} P_0(t) = \frac{\mu}{\lambda + \mu}, \\
    P_1 &= \lim_{t \to \infty} P_1(t) = \frac{\lambda}{\lambda + \mu}.
\end{align*}
\]

(3.7)

We note that \( P_0 + P_1 = 1 \) and they are called the \textit{steady-state probabilities} of the system.

\textbf{Remark 7} Both \( P_0 \) and \( P_1 \) are independent of the initial values \( P_0(0) \) and \( P_1(0) \). If at time \( t = 0 \),

\[
\begin{align*}
    P_0(0) &= \frac{\mu}{\lambda + \mu} = P_0, \\
    P_1(0) &= \frac{\lambda}{\lambda + \mu} = P_1,
\end{align*}
\]

(comes from Eq. (3.6)) clearly show that these initial values will persist for ever.
• This leads us to the important notion of statistical equilibrium. We say that a system is in **statistical equilibrium** (or the state distribution is **stationary**) if its state probabilities are constant in time.

• Note that the system still fluctuate from state to state, but there is no net trend in such fluctuations.

• In the above queueing example, we have shown that the system attains statistical equilibrium as $t \to \infty$.

• Practically speaking, this means the system is in statistical equilibrium after sufficiently long time (so that initial conditions have no more effect on the system). For the general birth-and-death processes, the following holds.
Proposition 2  (a) Let $P_j(t)$ be the state probabilities of a birth-and-death process. Then

$$\lim_{t \to \infty} P_j(t) = P_j$$

exist and are independent of the initial conditions; they satisfy the system of linear difference equations obtained from the difference-differential equations in previous chapter by replacing the derivative on the left by zero.

(b) If all $\mu_j > 0$ and the series

$$S = 1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} + \cdots + \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} + \cdots$$

(3.8)

converges, then

$$P_0 = S^{-1} \quad \text{and} \quad P_j = \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} S^{-1} \quad (j = 1, 2, \cdots).$$

If the series in Eq. (3.8) diverges, then

$$P_j = 0 \quad (j = 0, 1, \cdots).$$
Proof: We shall NOT attempt to prove Part (a) of the above proposition, but rather we assume the truth of (a) and use it to prove Part (b).

By using part (a) of the proposition, we obtain the following linear difference equations.

\[
(\lambda_j + \mu_j) P_j = \lambda_{j-1} P_{j-1} + \mu_{j+1} P_{j+1} \\
(\lambda_{-1} = \mu_0 = 0 ; \quad j = 0, 1, \cdots) \quad P_{-1} = 0.
\]  
\[ (3.9) \]

Re-arranging terms, we have

\[
\lambda_j P_j - \mu_{j+1} P_{j+1} = \lambda_{j-1} P_{j-1} - \mu_j P_j .
\]
\[ (3.10) \]

If we let

\[
[f(j) = \lambda_j P_j - \mu_{j+1} P_{j+1}]
\]
then Equation (3.10) simply becomes

\[
f(j) = f(j - 1) \quad \text{for} \quad j = 0, 1, \cdots
\]
as \( f(-1) = 0 \). Hence

\[
f(j) = 0 \ (j = 0, 1, \cdots).
\]
This implies
\[ \lambda_j P_j = \mu_{j+1} P_{j+1}. \]
By recurrence, we get (if \( \mu_1, \cdots, \mu_j > 0 \))
\[ P_j = \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j} P_0 \quad (j = 1, 2, \cdots). \] (3.11)

Finally, by using the normalization condition \( \sum P_j = 1 \) we have the result in part (b).

\[ \square \]

**Remark 8** Part (a) of the above proposition suggests that to find the statistical equilibrium distribution
\[ \lim_{t \to \infty} P_j(t) = P_j. \]
We set the derivatives on the left side of difference-differential equations to be zero and replace \( P_j(t) \) by \( P_j \) and then solve the linear difference equations for \( P_j \). In most cases, the latter method is much easier and shorter.

**Remark 9** If \( \mu_j = 0 \) for some \( j = k \) (\( \lambda_j > 0 \) for all \( j \)), then, as equation (3.11) shows,
\[ P_j = 0 \quad \text{for} \quad j = 0, 1, \cdots, k - 1. \]
In particular, for pure birth process, \( P_j = 0 \) for all \( j \).
**Example 1** Suppose that for all $i$ we have

$$\lambda_i = \lambda \quad \text{and} \quad \mu_j = j\mu$$

then

$$S = 1 + \frac{\lambda}{\mu} + \frac{1}{2!} \left(\frac{\lambda}{\mu}\right)^2 + \frac{1}{3!} \left(\frac{\lambda}{\mu}\right)^3 + \ldots + = e^{\frac{\lambda}{\mu}}.$$

Therefore we have the Poisson distribution

$$P_j = \frac{1}{j!} \left(\frac{\lambda}{\mu}\right)^j e^{-\frac{\lambda}{\mu}}.$$

**Example 2** Suppose that for all $i$ we have

$$\lambda_i = \lambda \quad \text{and} \quad \mu_j = \mu$$

such that $\lambda < \mu$ then

$$S = 1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + \left(\frac{\lambda}{\mu}\right)^3 + \ldots + = \frac{1}{1 - \frac{\lambda}{\mu}}.$$

Therefore we have the Geometric distribution

$$P_j = \left(\frac{\lambda}{\mu}\right)^j \left(1 - \frac{\lambda}{\mu}\right).$$
3.5 A Summary of Learning Outcomes

• Able to give the definition of a birth-and-death process.

• Able to derive the general solutions for the pure birth process, the pure death process and the two-state birth-and-death process.

• Able to compute and interpret the steady-state solution of a birth-and-death process.

• Able to give the condition for the existence of the steady-state probability of a birth-and-death process.
3.6  Exercises

1. For the pure death process with death rate $\mu_j = j\mu$, prove that

$$p_j(t) = \binom{n}{j}(e^{-\mu t})^j(1 - e^{-\mu t})^{n-j} \quad (j = 0, 1, \cdots, n)$$

where $n$ is the state of the system at $t = 0$ and $p_j(t)$ is the probability that the system is in State $j$ at time $t$.

2. In a birth and death process, if $\lambda_i = \lambda/(i + 1)$ and $\mu_i = \mu$

show that the equilibrium distribution is Poisson.

3. Consider a birth-death system with the following birth and death coefficients:

$$\left\{ \begin{array}{l}
\lambda_k = (k + 2)\lambda \quad k = 0, 1, 2, \ldots \\
\mu_k = k\mu \quad k = 0, 1, 2, \ldots
\end{array} \right.$$ 

All other coefficients are zero.

(a) Solve for $p_k$. Make sure you express your answer explicitly in terms of $\lambda$, $k$ and $\mu$ only.

(b) Find the average number of customers in the system.
3.7 Suggested Solutions

1. In the assumed pure death process, actually the lifetime of each of the individual follows the exponential distribution $\mu e^{-\mu x}$. The probability that one can survive over time $t$ will be given by

$$\int_t^{\infty} \mu e^{-\mu x} dx = e^{-\mu t}.$$ 

Therefore the probability that one can find $j$ still survive at time $t$ is given by

$$\binom{n}{j} (e^{-\mu t})^j (1 - e^{-\mu t})^{n-j} \quad (j = 0, 1, \cdots, n)$$

2. Let $P_i$ be the steady-state probability and

$$P_j = \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j} P_0 \quad (j = 1, 2, \cdots)$$

$$= \frac{\lambda^j P_0}{j! \mu^j}$$

and $P_0 = (\sum_{i=0}^{\infty} P_i)^{-1} = (e^{\lambda/\mu})^{-1}$. Hence we have a Poisson distribution

$$P_n = \frac{\lambda^j e^{-\lambda/\mu}}{j! \mu^j}.$$
3. (a) Because $\lambda_k = (k + 2)\lambda$ and $\mu_k = k\mu$, we can get that

$$s = 1 + 2\frac{\lambda}{\mu} + \cdots + (j + 1)(\frac{\lambda}{\mu})^j + \cdots + = \sum_{k=1}^{\infty} k\rho^{k-1} = \frac{1}{(1 - \rho)^2}$$

so

$$P_0 = 1/s = (1 - \rho)^2$$

and

$$P_i = (i + 1)\rho^i(1 - \rho)^2.$$ 

(b)

$$E = \sum_{k=0}^{\infty} iP_i = 0 \cdot P_0 + \sum_{k=1}^{\infty} k(k + 1)\rho^k(1 - \rho)^2 = \frac{2\rho}{1 - \rho}.$$ 

Note:

$$\sum_{k=1}^{\infty} k(k + 1)\rho^{k-1} = \frac{d}{d\rho} \sum_{k=1}^{\infty} (k + 1)\rho^k = \frac{d}{d\rho} \left[ \frac{d}{d\rho} \sum_{k=1}^{\infty} \rho^{k+1} \right]$$
4 Introduction to Queueing Systems

A queueing situation $^1$ 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 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].

$^1$Queueing theory is the mathematical study of waiting lines, or queues. In queueing theory a model is constructed so that queue lengths and waiting times can be predicted. Queueing theory is generally considered a branch of operations research because the results are often used when making business decisions about the resources needed to provide a service. Queueing theory has its origins in research by Agner Krarup Erlang when he created models to describe the Copenhagen telephone exchange. The ideas have since seen applications including telecommunications, traffic engineering, computing and the design of factories, shops, offices and hospitals. (Taken from From Wikipedia)
• 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**”.

• Availability of such measures enables analysts to decide on an optimal way of operating such a system.
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 *interarrival 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 four machines (the machines are the customers and the repairman is the server), the calling source before any machine breaks down consists of four potential customers (i.e. anyone of the four 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 three 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.
(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. For a reliable machine, it takes 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 $n$th 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. Clearly 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

$$\frac{(0 + 10 + 30)}{3} = \frac{40}{3}.$$ 

**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

$$\frac{(0 + 30 + 50)}{3} = \frac{80}{3}$$

minutes which is twice of that in Case 1!
4.1.2 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, ... 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) $\rho =$ traffic intensity = offered load per server = $a/s$ ($s < \infty$).
(iv) $a' = \textit{carried load} = \text{mean number of busy servers.}$

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

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

(vii) $W_q = \textit{mean waiting time in the queue}$, i.e. the mean length of time from the moment a customer arrives until the customer’ service starts.

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

(ix) $L_q = \textit{mean number of customers waiting in the queue}$. 
Remark 10 If the mean arrival rate is \( \lambda \) and the mean service time is \( \tau \) then the offered load \( a = \lambda \tau \).

Remark 11 For an \( s \) server system, the carried load

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

Hence

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

Remark 12 If \( s = 1 \) then \( a' = \rho' \) and \( a = \rho \).

Remark 13 The carried load can also be considered as the mean number of customers completing service per mean service time \( \tau \). 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 the carried load = the offered load. In general

\[
a' = a(1 - B)
\]

where \( B = \) proportion of customers lost (or requests that are cleared).
4.1.3 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.
4.2 Queueing Systems of One Server

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

4.2.1 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 $\lambda$.

• The service time also follows the exponential distribution with parameter $\mu$. 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.
4.2.2 Steady-state Probability Distribution

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

Remark 14 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

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

and

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

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

$$p_0 = \lim_{t \to \infty} P_0(t) = \frac{\mu}{\lambda + \mu}$$

and

$$p_1 = \lim_{t \to \infty} P_1(t) = \frac{\lambda}{\lambda + \mu}.$$

Here $p_0$ and $p_1$ are called the steady-state probabilities.
• Moreover, we have

\[ P_0(t) - \frac{\mu}{\lambda + \mu} = \frac{\mu e^{-(\lambda+\mu)t}}{\lambda + \mu} \rightarrow 0 \quad \text{as } t \rightarrow \infty \]

and

\[ P_1(t) - \frac{\lambda}{\lambda + \mu} = \frac{\mu e^{-(\lambda+\mu)t}}{\lambda + \mu} \rightarrow 0 \quad \text{as } t \rightarrow \infty \]

very fast.

• This means that the system will go into the steady state very fast.

• Therefore, it will be a good idea if we focus on the steady-state probability instead of the transient probability. The former is easier to be obtained.
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 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 $p_1$.

**Remark 15** The system goes to its steady state very quickly. In general it is much easier to obtain the steady-state probabilities of a queueing system than the transient probabilities. Moreover we are interested in the long-run performance of the system. Therefore we will focus on the steady-state probabilities of a queueing system.
4.2.4 The Markov Chain and the Generator Matrix

A queueing system can be represented by a continuous time 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 $\mu$ and The transition rate from State 0 to State 1 is $\lambda$.

- In State 0, change of state occurs when there is an arrival of customers and the waiting time is exponentially distributed with parameter $\lambda$.

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

- Recall that from the no-memory (Markov) 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 4.1. The Markov Chain of the Two-state System.](image)
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:

\[ A_1 p = \begin{pmatrix} -\lambda & \mu \\ \lambda & -\mu \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad (4.1) \]

subject to \( p_0 + p_1 = 1. \)

**Remark 16** Given a Markov chain (generator matrix) one can construct the corresponding generator (Markov chain). To interpret the system of linear equations. We note that in steady state, the expected incoming rate and the expected outgoing rate at any state must be equal. Therefore, we have the followings:

- **At State 0**: expected outgoing rate = \( \lambda p_0 = \mu p_1 \) = expected incoming rate;
- **At State 1**: expected outgoing rate = \( \mu p_1 = \lambda p_0 \) = expected incoming rate.
4.2.5 One-Server Queueing System with Waiting Space

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

- The generator matrix is a $5 \times 5$ matrix.

$$A_2 = \begin{pmatrix} -\lambda & \mu & 0 \\ \lambda & -\lambda - \mu & \mu \\ -\lambda - \mu & \mu & 0 \\ 0 & \lambda & -\mu \end{pmatrix}.$$  \hfill (4.2)
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:** the expected outgoing rate = \( \lambda p_0 = \mu p_1 \) = expected incoming rate;

**At State 1:** the expected outgoing rate = \( (\lambda + \mu)p_1 = \lambda p_0 + \mu p_2 \) = expected incoming rate.

**At State 2:** the expected outgoing rate = \( (\lambda + \mu)p_2 = \lambda p_1 + \mu p_3 \) = expected incoming rate;

**At State 3:** the expected outgoing rate = \( (\lambda + \mu)p_3 = \lambda p_2 + \mu p_4 \) = expected incoming rate.

**At State 4:** the expected outgoing 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

$$p_0 = (1 + \rho + \rho^2 + \rho^3 + \rho^4)^{-1}$$  

and

$$p_i = p_0 \rho^i, \quad i = 1, 2, 3, 4.$$  

• What is the solution of a general one-server queueing system (M/M/1/n)? We shall discuss it in the next section.
4.2.6 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 $\lambda$ and $\mu$ 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 Figure 4.3.

\[
\begin{align*}
0 \quad &\xrightarrow{\mu} 1 \quad &\xrightarrow{\mu} \cdots \xrightarrow{\mu} s \quad &\xrightarrow{\mu} \cdots \xrightarrow{\mu} (n-1) \\
&\xrightarrow{\lambda} &\xrightarrow{\lambda} &\xrightarrow{\lambda} &\xrightarrow{\lambda} \\
\end{align*}
\]

Figure 4.3. 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 \( A_2 \):

\[
\begin{pmatrix}
0 & 1 & 2 & 3 & \cdots & n - 3 & n - 2 & n - 1 \\
0 & -\lambda & \mu & & & & & \\
1 & \lambda & -\lambda - \mu & \mu & & & & \\
2 & \cdots & \cdots & \cdots & \cdots & \cdots & & \\
\vdots & \lambda & -\lambda - \mu & \mu & & & & \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\
n - 2 & \lambda & -\lambda - \mu & \mu & & & & \\
n - 1 & 0 & \lambda & -\lambda - \mu & \mu & & & \\
\end{pmatrix}
\]

\[ (4.3) \]

• We are going to solve the probability distribution \( p \). Let

\[ p = (p_0, p_1, \ldots, p_{n-2}, p_{n-1})^T \]

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

\[ A_2 p = 0 \quad \text{and} \quad \sum_{i=0}^{n-1} p_i = 1. \]
• 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:

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

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 distribution vector $p$ is given by

$$\frac{1 - \rho}{1 - \rho^n}(1, \rho, \rho^2, \ldots, \rho^{n-1})^T.$$
Remark 17 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:

\[ L_c = 0 \cdot p_0 + 1 \cdot (p_1 + p_2 + \ldots + p_{n-1}) \]

\[ = \frac{1 - \rho}{1 - \rho^n} (\rho + \rho^2 + \ldots + \rho^{n-1}) \]

\[ = \frac{\rho (1 - \rho^{n-1})}{1 - \rho^n}. \] (4.4)
(d) the expected number of customers in the system is given by

\[
L_s = \sum_{i=0}^{n-1} ip_i = \sum_{i=1}^{n-1} ip_0 \rho^i \\
= \frac{\rho - n \rho^n + (n - 1) \rho^{n+1}}{(1 - \rho)(1 - \rho^n)}.
\]

(e) the expected number of customers in the queue

\[
L_q = \sum_{i=1}^{n-1} (i - 1)p_i \\
= \sum_{i=1}^{n-1} (i - 1)p_0 \rho^i \\
= \sum_{i=1}^{n-1} ip_0 \rho^i - \sum_{i=1}^{n-1} p_0 \rho^i \\
= \frac{\rho^2 - (n - 1) \rho^n + (n - 2) \rho^{n+1}}{(1 - \rho)(1 - \rho^n)}.
\]

We note that \( L_s = L_q + L_c \).
Remark 18 To obtain the results in (d) and (e) we need the following results.

\[ \sum_{i=1}^{n-1} i\rho^i = \frac{1}{1 - \rho} \sum_{i=1}^{n-1} (1 - \rho)i\rho^i \]

\[ = \frac{1}{1 - \rho} \left( \sum_{i=1}^{n-1} i\rho^i - \sum_{i=1}^{n-1} i\rho^{i+1} \right) \]

\[ = \frac{1}{1 - \rho} \left( \rho + \rho^2 + \ldots + \rho^{n-1} - (n - 1)\rho^n \right) \]

\[ = \frac{\rho + (n - 1)\rho^{n+1} - n\rho^n}{(1 - \rho)^2}. \quad (4.7) \]

Moreover if \(|\rho| < 1\) we have

\[ \sum_{i=1}^{\infty} i\rho^i = \frac{\rho}{(1 - \rho)^2}. \]
4.3 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 $\lambda$ and the service rate of each server is $\mu$. 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 (re-visited).

- 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

\[ t = \min\{t_1, t_2, \ldots, t_k\}. \]

We will derive the probability density function of $t$.

We note that

\[
\text{Prob}(t \geq x) = \text{Prob}(t_1 \geq x) \cdot \text{Prob}(t_2 \geq x) \ldots \text{Prob}(t_k \geq x)
\]

\[
= \left( \int_{x}^{\infty} \lambda e^{-\lambda t} dt \right)^k
\]

\[
= (e^{-\lambda x})^k = e^{-k\lambda x}.
\]

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 still \textbf{exponentially distributed} with parameter $k\lambda$. 

57
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 the following figure.

\[
\begin{array}{cccccc}
0 & \mu & 2\mu & \ldots & s\mu & s\mu \\
\lambda & \lambda & \lambda & \ldots & \lambda & \lambda \\
\end{array}
\]

Figure 4.5. 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 the 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:

\[
A_3 = \begin{pmatrix}
-\lambda & \mu & 0 \\
\lambda & -\lambda - \mu & 2\mu \\
& \ddots & \ddots & \ddots \\
\lambda & -\lambda - s\mu & s\mu & \ddots & \ddots \\
0 & \lambda & -\lambda - s\mu & s\mu & \lambda & -s\mu
\end{pmatrix}.
\] (4.9)
4.3.1 A Two-server Queueing System

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

![Figure 4.6. The Markov Chain for the M/M/2/2 Queue.](image)

• The generator matrix is an $5 \times 5$ matrix.

\[
A_4 = \begin{pmatrix}
-\lambda & \mu & 0 \\
\lambda & -\lambda - \mu & 2\mu \\
\lambda & -\lambda - 2\mu & 2\mu \\
\lambda & -2\mu & 0 \\
0 & \lambda & -2\mu \\
\end{pmatrix}.
\]  

Let the steady-state probability distribution be

\[
p = (p_0, p_1, p_2, p_3, p_4)^T.
\]

In steady state we have $A_4p = 0$. 

• 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

$$p_0 = \left(1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2!\mu^2} \left(\frac{1 - \tau^3}{1 - \tau}\right)\right)^{-1}$$

and

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

and

$$p_i = p_0 \left(\frac{\lambda^2}{2!\mu^2}\right) \tau^{i-2}, \quad i = 2, 3, 4.$$
• The result above can be further extended to the \( M/M/2/k \) queue as follows:

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

The queueing system has finite number of waiting space.

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

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

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.
4.3.2 Expected Number of Customers in the M/M/2/∞ Queue

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

\[ L_s = \sum_{k=1}^{\infty} k \rho_k = \frac{1 - \tau}{1 + \tau} \sum_{k=1}^{\infty} 2k \tau^k. \]

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

\[ L_s = \frac{2\tau}{1 - \tau^2}. \]  \hspace{1cm} (4.11)
4.3.3 Multiple-Server Queues

Now we consider general 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, we assume that the Poisson input has rate $\lambda$ and the exponential service times have mean $\mu^{-1}$. 
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

\[
A_5 = \begin{pmatrix}
-\lambda & \mu & 0 \\
\lambda & -\lambda - \mu & 2\mu \\
\lambda & -\lambda - 2\mu & \\
\lambda & -\lambda - (s - 1)\mu & s\mu \\
0 & \lambda & -s\mu
\end{pmatrix}.
\]
Let $p_i$ be the steady-state probability that there are $i$ customers in the queueing system. Then by solving

$$A_5 \mathbf{p} = \mathbf{0} \quad \text{with} \quad \sum_{i=0}^{s} p_i = 1$$

one can get

$$p_j = \frac{(\lambda/\mu)^j / j!}{\sum_{k=0}^{s} (\lambda/\mu)^k / k!} \quad (j = 0, 1, \cdots, s)$$

$$= \frac{a^j / j!}{\sum_{k=0}^{s} a^k / k!}$$

(4.12)

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 $\lambda$, 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

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

**Remark 19** 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

$$
B(s, a) = \frac{a^s / s!}{\sum_{k=0}^{s} (a^k / k!)}
$$
Remark 20 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:

\[
\begin{align*}
    a' &= \sum_{j=1}^{s} j (a^j / j!) / \sum_{k=0}^{s} (a^k / k!)
    \\
    &= a \left( \sum_{j=0}^{s-1} (a^j / j!) / \sum_{k=0}^{s} (a^k / k!) \right)
    \\
    &= a \left( 1 - B(s, a) \right).
\end{align*}
\]

This shows that the carried load is the portion of the offered load that is not lost (captured) from the system.
4.3.5 Blocked Customers Delayed (Erlang delay system)

The queueing system has $s$ servers and there is no limit in waiting space and we assume blocked customers are delayed.

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

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

where $a = \lambda/\mu$ and

$$p_0 = \left( \sum_{k=0}^{s-1} \frac{a^k}{k!} + \sum_{k=s}^{\infty} \frac{a^k}{s!s^{k-s}} \right)^{-1}.$$
• If $a < s$, the infinite geometric sum on the right converges, and

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

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.
Remark 21 The probability that all servers are occupied (as observed by an outside observer) is given by the **Erlang delay formula**

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

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 \( C(s, a) \).

**Remark 22** 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.
**Remark 23** 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 \( P\{Q = j \mid w > 0\} \) where \( Q \) denotes the number of customers waiting in the queue.

- By the definition of conditional probability,

\[
P\{Q = j \mid w > 0\} = \frac{P\{Q = j, w > 0\}}{P\{w > 0\}}.
\]

Thus

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

we get the Geometric distribution

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

where \( \rho = a/s \) is the traffic intensity.
4.4 Little’s Queueing Formula

If $\lambda$ 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

$$L = \lambda W.$$  

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 3 (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 \( \lambda \) as

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

and let \( W_i \) be the sojourn time of the \( i \)th customer, and define the mean sojourn time \( W \) as

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

If \( \lambda \) 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 $\lambda t$. 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 $\lambda tW$.

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,

$$\lim_{t \to \infty} \frac{\lambda tW}{Lt} = 1.$$
With the help of Little’s formula, we get the following useful results:

(a) $\lambda$, 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

$$L_s = \lambda W_s, \quad L_q = \lambda W_q, \quad \text{and} \quad L_c = \lambda W_c.$$
4.4.1 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 \( i/\mu \). 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

\[ \sum_{i=1}^{\infty} \frac{i}{\mu} p_i = \frac{1}{\mu} \sum_{i=1}^{\infty} ip_i = \frac{L_s}{\mu} = \frac{\rho}{(1 - \rho)\mu}. \]
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

\[
W_s = W_q + W_c = \frac{L_q}{\mu} + \frac{1}{\mu} = \frac{1}{\mu(1 - \rho)} + 1
\]

\[
= \frac{\lambda(1 - \rho)}{\mu(1 - \rho)} = \frac{L_s}{\lambda}
\]

Here

\[
\rho = \frac{\lambda}{\mu}
\]

and

\[
L_s = \frac{\rho}{(1 - \rho)}.
\]
### 4.4.2 Applications of the Little’s Queueing Formula

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

Table 4.1. A summary of the M/M/1/∞ queue.
Example 3 Consider the M/M/2/∞ queue with arrival rate \( \lambda \) and service rate \( \mu \). 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 4 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 4.2.
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrival rate</td>
<td>$\lambda = 30$ (per hour)</td>
</tr>
<tr>
<td>Service rate</td>
<td>$\mu = 60$ (per hour)</td>
</tr>
<tr>
<td>Traffic intensity</td>
<td>$\rho = 0.5$</td>
</tr>
<tr>
<td>Probability that no customer in the queue</td>
<td>$p_0 = 0.5$</td>
</tr>
<tr>
<td>Probability that $i$ customers in the queue</td>
<td>$p_i = 0.5^{i+1}$</td>
</tr>
<tr>
<td>Probability that an arrival has to wait for service</td>
<td>0.5</td>
</tr>
<tr>
<td>Expected number of customers in the system</td>
<td>$L_s = 1$</td>
</tr>
<tr>
<td>Expected number of customers in the queue</td>
<td>$L_q = 0.5$</td>
</tr>
<tr>
<td>Expected number of customers in the server</td>
<td>$L_c = 0.5$</td>
</tr>
<tr>
<td>Expected waiting time in the system</td>
<td>$L_s/\lambda = 1/30$</td>
</tr>
<tr>
<td>Expected waiting time in the queue</td>
<td>$L_q/\lambda = 1/60$</td>
</tr>
<tr>
<td>Expected waiting time in the server</td>
<td>$L_c/\lambda = 1/60$</td>
</tr>
</tbody>
</table>

Table 4.2. A summary of the system performance
4.5 Applications of Queues

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

We consider a queueing system consisting of $n$ independent M/M/1 queues. The service rate of the serve at the $i$th queue is $\mu_i$.

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

Figure 4.7. The Queueing System with Allocation of Arrivals.
• An allocation process is implemented such that it diverts an arrived customers to queue $i$ with probability

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

• Then the input process of queue $i$ is a Poisson process with rate $\lambda_i$.

• The objective here is to find the parameters $\lambda_i$ such that some system performance is optimized.

• We remark that we must have $\lambda_i < \mu_i$.  

4.5.2 Minimizing Number of Customers in the System

The expected number of customers in queueing system $i$:

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

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

$$\sum_{i=1}^{m} \lambda_i = M \quad \text{and} \quad 0 \leq \lambda_i < \mu_i \quad \text{for} \quad i = 1, 2, \ldots, n.$$
• By considering 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

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

\[
0 \leq \lambda_i < \mu_i \quad \text{for} \quad i = 1, 2, \ldots, n.
\]
By considering 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

\[ \sum_{i=1}^{n} \mu_i \left( 1 - \frac{1}{\sqrt{1 + m\mu_i}} \right) = M. \]
4.5.4 Which Operator to Employ?

- 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.

- 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 the inter-arrival time of workers and the processing time of the operators are exponentially distributed. We may regard the request for tools as a queueing process (M/M/1/∞) with $\lambda = 5$. 
• 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

$$40 + 5 = 45.$$
• For **Operator B**, the service rate is $\mu = 60/11$ per hour. Thus we have

$$\rho = \frac{\lambda}{\mu} = \frac{11}{12}.$$  

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

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

(ii) **Put two operators together.** We have an M/M/2/$\infty$ queue.

![Diagram of two M/M/1/$\infty$ queues](image1.png)

Figure 4.8. Case (i) Two M/M/1/$\infty$ Queues.

![Diagram of one M/M/2/$\infty$ queue](image2.png)

Figure 4.9. Case (ii) One M/M/2/$\infty$ 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

\[ S_1 = 2 \times \frac{\frac{\lambda}{2\mu}}{1 - \frac{\lambda}{2\mu}} = \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{2\mu}}. \]

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

\[ S_1 = 2 \times \frac{\frac{\lambda}{2\mu}}{1 - \frac{\lambda}{2\mu}} = \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{2\mu}}. \]

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

\[ S_2 = \frac{\frac{\lambda}{\mu}}{1 - (\frac{\lambda}{2\mu})^2}. \]

Clearly \( S_2 < S_1 \).

Conclusion: Case (ii) is better. We should put all the servers (operators) together.
• 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/$\infty$) 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/$\infty$ queue.
• The expected number of workers in the system is given by (c.f. (4.11))

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

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?
Consider an unreliable machine system. The normal time of the machine is exponentially distributed with mean $\lambda^{-1}$. 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 $i$. The Markov chain of the model is given by

![Figure 4.10. The Markov Chain for the Unreliable Machine System.](image-url)
Let the steady-state probability vector be

\[ \mathbf{p} = (p_0, p_1, \ldots, p_n) \]

satisfies

\[ A_6 \mathbf{p} = \mathbf{0} \]

where

\[ A_6 = \begin{pmatrix}
-\lambda & 0 & \mu_n \\
\lambda & -\mu_1 & \\
\mu_1 & -\mu_2 & \\
\vdots & \vdots & 0 \\
0 & \mu_{n-1} & -\mu_n
\end{pmatrix}. \]
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

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

- The inter-arrival time of a demand is also exponentially distributed with a mean time of $\lambda^{-1}$.

- 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;
\( \mu^{-1} \), the mean production time for one unit of product;
\( \lambda^{-1} \), the mean inter-arrival time of a demand.

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.

![Markov Chain Diagram](image)

Figure 4.11. The Markov Chain for the Manufacturing System.
• 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

\[ q(i) = (1 - p) p^{n-i}, \quad i = n, n - 1, n - 2, \ldots \]

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:

\[
E(n) = I \sum_{i=0}^{n} (n - i)(1 - p)p^i + B \sum_{i=n+1}^{\infty} (i - n)(1 - p)p^i. \quad (4.13)
\]

\[ \text{inventory cost} \quad \text{backlog cost} \]
Proposition 4 The expected running cost $E(n)$ is minimized if the hedging point $n$ is chosen such that

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

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

$$E(n-1) \geq E(n) \iff p^n \geq \frac{I}{I + B} \quad \text{and} \quad E(n+1) \geq E(n) \iff p^{n+1} \leq \frac{I}{I + B}.$$  

Thus the optimal value of $n$ is the one such that $p^{n+1} \leq \frac{I}{I+B} \leq p^n$. 

$\Box$
In the world of limited resources and disposal capacities, there is a environmental pressure in using re-manufacturing system, a recycling process to reduce the amount of waste generated.

A major manufacturer of copy machines Xeron reported on annual savings of several hundred million dollars due to the re-manufacturing and re-use of equipment.

A return is first repaired/tested and then re-sell to the market. The result of re-manufacturing is that the manufacturers have to take into account of returns in their production plans.

(i) $\lambda^{-1}$, the mean inter-arrival time of demands,
(ii) $\mu^{-1}$, the mean inter-arrival time of returns,
(iii) $a$, the probability that a returned product is repairable,
(iv) $Q$, maximum inventory capacity,
(v) $I$, unit inventory cost,
(vi) $R$, cost per replenishment order.

![Diagram](image)

Figure 4.16. The Single-item Inventory Model.

The $(Q + 1) \times (Q + 1)$ system generator matrix is given as follows:

$$A = \begin{pmatrix}
0 & \lambda + a\mu & -\lambda & 0 \\
1 & -a\mu & \lambda + a\mu & -\lambda \\
\vdots & \ddots & \ddots & \ddots \\
Q & -\lambda & -a\mu & \lambda + a\mu & -\lambda \\
\end{pmatrix}. \quad (4.14)$$

The steady state probability distribution $p$ is given by

$$p_i = K(1 - \rho^{i+1}), \; i = 0, 1, \ldots, Q \quad (4.15)$$

where

$$\rho = \frac{a\mu}{\lambda} \quad \text{and} \quad K = \frac{1 - \rho}{(1 + Q)(1 - \rho) - \rho(1 - \rho^{Q+1})}.$$
Proposition 5 The expected inventory level is

\[ \sum_{i=1}^{Q} i_{p_i} = \sum_{i=1}^{Q} K(i - i\rho^{i+1}) = K \left( \frac{Q(Q + 1)}{2} + \frac{Q\rho^{Q+2}}{1 - \rho} - \frac{\rho^2(1 - \rho^Q)}{(1 - \rho)^2} \right), \]

the average rejection rate of returns is

\[ \mu p_Q = \mu K(1 - \rho^{Q+1}) \]

and the mean replenishment rate is

\[ \lambda \times p_0 \times \frac{\lambda^{-1}}{\lambda^{-1} + (a\mu)^{-1}} = \frac{\lambda K(1 - \rho)\rho}{(1 + \rho)}. \]
Proposition 6 If $\rho < 1$ and $Q$ is large then $K \approx (1 + Q)^{-1}$ and the approximated average running cost (inventory and replenishment cost)

$$C(Q) \approx \frac{QI}{2} + \frac{\lambda(1 - \rho)\rho R}{(1 + \rho)(1 + Q)}.$$ 

The optimal replenishment size

$$Q^* + 1 \approx \sqrt{\frac{2\lambda(1 - \rho)\rho R}{(1 + \rho)I}} = \sqrt{\frac{2a\mu R}{I} \left( \frac{2\lambda}{\lambda + a\mu} - 1 \right)}.$$ (4.16)
4.9  Queueing Systems with Two Types of Customers

In this section, we discuss queueing systems with two types of customers. The queueing system has no waiting space. There are two possible cases: infinite-server case and finite-server case.

4.9.1  Infinite-Server Queue

Consider the infinite-server queue with two types of customers. The arrival process of customers of type $i$ ($i = 1, 2$) is Poisson with rate $\lambda_i$ and their service times are independent, identically distributed, exponential random variables with mean $\mu_i^{-1}$ ($i = 1, 2$).

• We define the 2-dimensional states $\{E_{j_1, j_2}\}$, where $j_i$ is the number of customers of type $i$ in the system, with corresponding equilibrium distribution $\{p(j_1, j_2)\}$, then clearly the Markov property still holds.

• Here $p(j_1, j_2)$ is the steady-state probability that there are $j_1$ type 1 customers and $j_2$ type 2 customers in the system.
By equating expected **rate out** to expected **rate in** for each state, the equilibrium state equations are

\[
(\lambda_1 + \lambda_2 + j_1\mu_1 + j_2\mu_2)p(j_1, j_2) = \lambda_1 p(j_1 - 1, j_2) + (j_1 + 1)\mu_1 p(j_1 + 1, j_2) \\
+ \lambda_2 p(j_1, j_2 - 1) + (j_2 + 1)\mu_2 p(j_1, j_2 + 1)
\]

and \([p(-1, j_2) = p(j_1, -1) = 0 ; j_1 = 0, 1, \ldots ; j_2 = 0, 1, \ldots.]

![Markov Chain Diagram]

Figure 4.12. The Markov Chain of the System at State \(E_{j_1,j_2}\).
• In addition, the probabilities must satisfy the normalization equation

\[ \sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} p(j_1, j_2) = 1. \]

• In this case, we already know the answer. Since the number of servers is infinite, the two types of customers do not affect one another.

• Thus the marginal distribution of the number of customers of each type is that which would be obtained by solving the corresponding one-dimensional problem, namely the Poisson distribution:

\[
\begin{align*}
\left\{ 
\begin{array}{l}
p_1(j) = \sum_{k=0}^{\infty} p(j, k) = \frac{(\lambda_1/\mu_1)^j}{j!} e^{-(\lambda_1/\mu_1)}, \\
p_2(j) = \sum_{k=0}^{\infty} p(k, j) = \frac{(\lambda_2/\mu_2)^j}{j!} e^{-(\lambda_2/\mu_2)}.
\end{array}
\right.
\end{align*}
\]  

(4.17)
• Since the number of customers present of each type is independent of the number present of the other type, therefore

\[
p(j_1, j_2) = p_1(j_1)p_2(j_2) = \frac{(\lambda_1/\mu_1)^{j_1}}{j_1!} \frac{(\lambda_2/\mu_2)^{j_2}}{j_2!} e^{-[(\lambda_1/\mu_1) + (\lambda_2/\mu_2)]}.
\]

(4.18)

• Solution of Product Form:

The fact that the solution \( p(j_1, j_2) \) can be decomposed into a product of two factors has enabled us to solve the problem with ease.
• In fact one may try

\[ p(j_1, j_2) = \frac{(\lambda_1/\mu_1)^{j_1}}{j_1!} \frac{(\lambda_2/\mu_2)^{j_2}}{j_2!} C \] (4.19)

where the constant \( C \) is determined from the normalization equation. In this case

\[
\sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} p(j_1, j_2) = \sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} \frac{(\lambda_1/\mu_1)^{j_1}}{j_1!} \frac{(\lambda_2/\mu_2)^{j_2}}{j_2!} C
\]

\[
= C e^{\lambda_2/\mu_2} \sum_{j_1=0}^{\infty} \frac{(\lambda_1/\mu_1)^{j_1}}{j_1!}
\]

\[
= C e^{[(\lambda_1/\mu_1)+(\lambda_2/\mu_2)]}.
\]

Hence

\[ C = e^{-[(\lambda_1/\mu_1)+(\lambda_2/\mu_2)]}. \]

• In practice, a good strategy for finding solutions to equations of the form (4.17) is to assume a product solution of the form (4.19); and see if such a solution satisfies (4.17).

• If it goes, then the solution has been obtained. If it doesn’t, then try a different approach. In this case it works!
• The situation is similar to that of the previous example except now the system has finitely many servers.

• The system is again described by equation (4.17), which is valid now only for $j_1 + j_2 < s$.

• When $j_1 + j_2 = s$, then the states $E_{j_1+1,j_2}$ and $E_{j_1,j_2+1}$ cannot occur and the equation becomes

\[
(j_1\mu_1 + j_2\mu_2)p(j_1, j_2) = \lambda_1 p(j_1 - 1, j_2) + \lambda_2 p(j_2, j_2 - 1).
\]  

(4.20)

• Observe that (4.20) can be obtained from (4.17) by deleting the first two terms on the left and the last two terms on the right.
Since the product-form solution (4.19) satisfies the equation (4.17) and also the equation with only the deleted terms

\[(\lambda_1 + \lambda_2)p(j_1, j_2) = (j_1 + 1)\mu_1p(j_1 + 1, j_2) + (j_2 + 1)\mu_2p(j_1, j_2 + 1),\]

therefore it also satisfies (4.20). Thus the product solution (4.19) is a solution of this problem.

In particular, if we don’t distinguish the two types of customers, then the probability \(p(j)\) that there are \(j\) customers (including type 1 and type 2) in service is given by

\[p(j) = \sum_{j_1 + j_2 = j} p(j_1, j_2) = \sum_{j_1 = 0}^{j} p(j_1, j - j_1).\]

With the help of binomial theorem, we have

\[p(j) = C \sum_{j_1 = 0}^{j} \frac{(\lambda_1)}{j_1!} \frac{(\lambda_2)}{j - j_1!} = C \frac{j!}{j!} \sum_{j_1 = 0}^{j} \frac{(\lambda_1)}{j_1!(j - j_1)!} \frac{(\lambda_2)}{j_1!(j - j_1)!} = C \frac{1}{j!} \left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}\right)^j.\]
• The normalization equation

\[ \sum_{k=0}^{s} p(k) = 1 \]

implies that

\[ C = \left\{ \sum_{k=0}^{s} \frac{1}{k!} \left( \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \right)^k \right\}^{-1}. \]

We conclude that

\[ p(j) = \frac{a^j / j!}{\sum_{k=0}^{s} a^k / k!} \quad (j = 0, 1, \ldots, s) \]

where

\[ a = \left( \frac{\lambda_1}{\mu_1} \right) + \left( \frac{\lambda_2}{\mu_2} \right). \]

• We note that for the case of infinite-server queue (we let \( s \to \infty \)) we have

\[ p(j) = \frac{a^j e^{-a}}{j!} \quad (j = 0, 1, \ldots ). \]
4.10 Queues in Tandem

- Consider two sets of servers arranged in tandem, so that the output from the first set of servers is the input of the second set of servers.

- Assume that the arrival process at the first stage of this tandem queueing system is Poisson with rate $\lambda$, the service times in the first stage are exponentially distributed with mean $\mu_1^{-1}$, and the queueing discipline is blocked customers delayed.

- The customers completing service in the first stage will enter the second stage (and wait if all servers in second stage are busy), where the service times are assumed to be exponentially distributed with mean $\mu_2^{-1}$. The number of servers in stage $i$ is $s_i (i = 1, 2)$.
Let \( p(j_1, j_2) \) be the probability that there are \( j_1 \) customers in stage 1 and \( j_2 \) customers in stage 2. Let

\[
\mu_i(j) = \begin{cases} 
  j\mu_i & (j = 0, 1, \ldots, s_i) \\
  s_i\mu_i & (j = s_i + 1, s_i + 2, \ldots)
\end{cases}
\]

be the departure rates. Then equating rate in to rate out for each state, we obtain

\[
(\lambda + \mu_1(j_1) + \mu_2(j_2)) p(j_1, j_2) = \lambda p(j_1 - 1, j_2) + \mu_1(j_1 + 1)p(j_1 + 1, j_2 - 1) \\
+ \mu_2(j_2 + 1)p(j_1, j_2 + 1)
\]

\[
[p(-1, j_2) = p(j_1, -1) = 0; \ j_1 = 0, 1, \ldots; \ j_2 = 0, 1, \ldots]
\]
• Since the first stage of this system is precisely an \textbf{Erlang delay system}, so that the marginal distribution of the number of customers in stage one is given by the \textbf{Erlang delay probabilities}.

• Let us try (hopefully it would work) a product solution of the form

\[ p(j_1, j_2) = p_1(j_1)p_2(j_2) \tag{4.22} \]

with the factor \( p_1(j_1) \) given by the Erlang delay probabilities:

\[
p_1(j_1) = \begin{cases} 
    C_1 \left( \frac{\lambda}{\mu_1} \right)^{j_1} \left( j_1 = 0, 1, \ldots, s_1 - 1 \right), \\
    C_1 \left( \frac{\lambda}{\mu_1} \right)^{j_1} \frac{j_1!}{s_1!} \left( j_1 = s_1, s_1 + 1, \ldots \right). 
\end{cases} \tag{4.23}
\]
• We shall substitute the assumed product solution (4.22) and (4.23) into the equilibrium state equations (4.21), with the hope that the system of equations will be reduced to a one-dimensional set of equations that can be solved for the remaining factor $p_2(j_2)$. Indeed, (4.21) is reduced to

$$[\lambda + \mu_2(j_2)]p_2(j_2) = \lambda p_2(j_2 - 1) + \mu_2(j_2 + 1)p_2(j_2 + 1)$$

$$[p_2(-1) = 0 ; \ j_2 = 0, 1, \ldots]$$

(4.24) is the equilibrium state equations that define the Erlang delay probabilities.

• We conclude that $p_2(j_2)$ is given by

$$p_2(j_2) = \begin{cases} 
C_2 \frac{(\lambda/\mu_2)^{j_2}}{j_2!} & (j_2 = 0, 1, 2, \ldots, s_2 - 1), \\
C_2 \frac{\lambda^{j_2}}{s_2! s_2^{j_2 - s_2}} & (j_2 = s_2, s_2 + 1, \ldots). 
\end{cases}$$

(4.25)
• Using
\[ \sum_{j=0}^{\infty} p_i(j) = 1 \quad (i = 1, 2) \]
we get
\[ C'_i = \left( \sum_{k=0}^{s_i-1} \frac{\rho_i^k}{k!} + \frac{\rho_i^{s_i}}{s_i!(1 - \rho_i/s_i)} \right)^{-1} \]
where \( \rho_i = \frac{\lambda}{\mu_i} \) (i = 1, 2).

This equation implies that a proper joint distribution exists only when \( \lambda/\mu_1 < s_1 \) and \( \lambda/\mu_2 < s_2 \).

• It is a remarkable result that the number of customers in the second stage also follows the Erlang delay distribution, that is, the distribution of customers in the second stage is the same as if the first stage were not present and the customers arrived according to a Poisson process directly at the second stage.
• It also suggests that the output from the first stage is a Poisson process. This is in fact true in general and we state without proof the Burke’s theorem as follows.

**Proposition 7** (Burke’s Theorem) The statistical equilibrium output of an $M/M/s$ queue with arrival rate $\lambda$ and mean service times $\mu^{-1}$ is a Poisson process with the same rate $\lambda$.

**Remark 24** In 1956 Burke showed that the departure process of an M/M/s queue was Poisson.

4.11 Queues in Parallel

Case (i) We assume there is no inter-action between the two queues. There are $s_i$ servers in queue $i (i = 1, 2)$ and there is infinite many waiting spaces in each queue and we assume blocked customers are delayed. The arrival rate of queue $i$ is $\lambda_i$ and the service completion rate of a serve in queue $i$ is $\mu_i$.

Figure 4.14. Two Queues in Parallel.
Hence we have the steady-state probability that queue \( i \) has \( j \) customers given by

\[
p_i(j) = \begin{cases} \frac{\rho_i^j}{j!} C_i & (j = 0, 1, \ldots, s_i) \\ \frac{\rho_i^j}{s_i!s_i^{j-s_i}} C_i & (j = s_i + 1, \ldots) \end{cases}
\]

where \( \rho_i = \lambda_i / \mu_i \). Moreover if \( \rho_i < s_i \) then

\[
C_i = \left( \sum_{k=0}^{s_i-1} \frac{\rho_i^k}{k!} + \sum_{k=s_i}^{\infty} \frac{\rho_i^k}{s_i!s_i^{k-s_i}} \right)^{-1} = \left( \sum_{k=0}^{s_i-1} \frac{\rho_i^k}{k!} + \frac{\rho_i^{s_i}}{(s_i - 1)!(s_i - \rho_i)} \right)^{-1}.
\]

- If \( \rho_i \geq s_i \), the infinite geometric sum diverges then \( C_i = 0 \) and hence \( p_i(j) = 0 \) for all finite \( j \).

- If \( \rho_i < s_i \) then the steady-state probability that there are \( i \) customers in Queue 1 and \( j \) customers in Queue 2 is given by

\[
p(i, j) = p_1(i)p_2(j) \quad i, j = 0, 1, 2, \ldots.
\]
Case (ii) There are $s_i(i = 1, 2)$ servers in queue $i(i = 1, 2)$ and there are finite many waiting spaces in each queue and we assume blocked customers are cleared. The arrival rate of queue $i$ is $\lambda_i$ and the service completion rate of a serve in queue $i$ is $\mu_i$.

We assume there is inter-action between the two queues as follows. Whenever Queue 2(1) is full, an arrived customer of type 2(1) will overflow to Queue 1(2) provided that Queue 1(2) is not yet full. Let us consider a simple example as follows. We assume that Queue 1 and 2 are M/M/1/1 queue. The following figure gives the Markov chain of the queueing system.

![Markov Chain](figure.png)

Figure 4.15. The Markov Chain of the Two Queue Overflow System.
• The generator matrix for this queueing problem is given by

\[
A_8 = \begin{pmatrix}
(0, 0) & \begin{pmatrix}
* & \mu_1 & 0 & \mu_2 & 0 & 0 & 0 & 0 & 0 \\
\lambda_1 & * & \mu_1 & 0 & \mu_2 & 0 & 0 & 0 & 0 \\
0 & \lambda_1 & * & 0 & 0 & \mu_2 & 0 & 0 & 0 \\
\lambda_2 & 0 & 0 & * & \mu_1 & 0 & \mu_2 & 0 & 0 \\
0 & \lambda_2 & 0 & \lambda_1 & * & \mu_1 & 0 & \mu_2 & 0 \\
0 & 0 & \tilde{\lambda} & 0 & \lambda_1 & * & 0 & 0 & \mu_2 \\
0 & 0 & 0 & \lambda_2 & 0 & 0 & * & \mu_1 & 0 \\
0 & 0 & 0 & 0 & \lambda_2 & 0 & \tilde{\lambda} & * & \mu_1 \\
0 & 0 & 0 & 0 & 0 & \tilde{\lambda} & 0 & \tilde{\lambda} & *
\end{pmatrix}
\end{pmatrix}
\]

(4.26)

• Here “*” is such that the column sum is zero.

• Unfortunately there is no analytic solution for the steady-state probabilities of this system. Direct method or numerical method are common methods for solving the steady-state probabilities.
4.12 A Summary of Learning Outcomes

• Able to use the Kendall’s notation to describe a queueing system.

• Able to compute and interpret an M/M/s/n queue including:
  - 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.

• Able to state and show the Little’s queueing formula and Burke’s theorem.

• Able to give a system performance analysis of a Markovian queueing system:
  - Expected number of customers.
  - Expected number of busy servers.
  - Mean waiting time.

• Able to apply queues in tandem and in parallel to real problems such as:
  - Employment of operators.
  - Unreliable machine system.
  - Manufacturing system and inventory system.
4.13 Exercises

1. Customers arrive according to a Poisson process with rate $\lambda$ at a single server with $n$ waiting positions. Customers who find all waiting positions occupied are cleared. All other customers wait as long as necessary for service. The mean service time is $\mu^{-1}$, and $\rho = \frac{\lambda}{\mu}$. Making no assumption about the form of the service time distribution function, show that

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

Here $p_i$ is the steady-state probability that there are $i$ customers in the system.

2. A company has 3 telephone lines. Incoming calls are generated by the customers according to a Poisson process with a mean rate of 20 calls per hour. Calls that find all telephone lines busy are rejected and those calls that are able to get through will be served for a length of time that is exponentially distributed with mean 4 minutes.

(a) Let the number of busy lines be the state of the system. Write down the Markov chain and the generator matrix for this system. Hence solve the steady-state probability distribution for the system.

(b) Find the proportion of calls that are rejected.
3. An company offers services that can be modeled as an $s$-server Erlang loss system (M/M/s/0 queue).

Suppose the arrival rate is 2 customers per hour and the average service time is 1 hour. The entrepreneur earns $2.50 for each customer served and the company’s operating cost is $1.00 per server per hour (whether the server is busy or idle).

(a) Write down the expected hourly net profit $C(s)$ in terms of $s$.

(b) Show that

$$\lim_{s \to \infty} \frac{C(s)}{s} = -1$$

and interprets this result.

(c) If the maximum number of servers available is 5, what is the optimal number of servers which maximizes the expected hourly net profit? What is the expected hourly net profit earned when the optimal number of servers is provided?
4. Consider an Erlang loss system with two servers. The arrival rate is 1 and mean service time is \( \mu_i^{-1} \) for server \( i (i = 1, 2) \). When the system is idle, an arrived customer will visit the first server. Denote the states of the system as

\[(0, 0), (1, 0), (0, 1), (1, 1)\]

where \((0, 0)\) denotes the state that both servers are idle, \((1, 0)\) denotes the state that the first server is busy and the second server is idle, \((0, 1)\) denotes the state that the first server is idle and the second server is busy and \((1, 1)\) denotes the state that both servers are busy.

(a) We adopt the order of states: \((0, 0), (1, 0), (0, 1), (1, 1)\), write down the generator matrix for this Markov process.

(b) Find the proportion of customers loss \( B(\mu_1, \mu_2) \) in terms of \( \mu_1 \) and \( \mu_2 \)

(c) Show that it is better to put the faster server as the first server, in order to achieve lower proportion of customers loss.
5. We consider a closed queueing system of two one-server queues in tandem. There are \( n + 1 \) customers in this system. When a customer finishes his/her service at Server 1 he/she will go for the service at server 2 (join Queue 2). After he/she finishes the service at Server 2, he/she will go for the service at server 1 (join Queue 1) again. The service time at server \( i \) is exponentially distributed with mean \( \mu_i^{-1} \).

(a) Let \( E_{i,j} \) be the state that there are \( i \) customers in the Queue 1 (including the one at the server) and \( j \) customers in Queue 2 (including the one at the server). Write down the Markov chain for this queueing system.

(b) Compute the steady-state probability \( p_{i,j} \) (\( i \) customers in Queue 1 and \( j \) customers in Queue 2).
4.14 Suggested Solutions

1. The mean number of customers joining the system per mean service time is

\[ \rho(1 - p_{n+1}) \]

and the mean number of customers completing service per mean service time is \(1 - p_0\). Hence

\[ \rho(1 - p_{n+1}) = 1 - p_0 \]

and

\[ \rho = \frac{1 - p_0}{1 - p_{n+1}}. \]

2. (a) The Markov chain is given by

\[
\begin{array}{cccc}
0 & 1/4 & 2/4 & 3/4 \\
1/3 & 1/3 & 1/3 & 1/3 \\
1/3 & 1/3 & 1/3 & 1/3 \\
1/3 & 1/3 & 1/3 & 1/3 \\
\end{array}
\]

The generator matrix is a \(4 \times 4\) matrix.

\[
A = \begin{pmatrix}
-1/3 & 1/4 & 0 & 0 \\
1/3 & -1/3 - 1/4 & 2/4 & 0 \\
0 & 1/3 & -1/3 - 2/4 & 3/4 \\
0 & 0 & 1/3 & -3/4 \\
\end{pmatrix}.
\] (4.27)
Let the steady-state probability distribution be

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

We have

\[ p_1 = (4/3)p_0, \quad p_2 = 2((7/12)(4/3)p_0 - (1/3)p_0) = (8/9)p_0, \]
\[ p_3 = (4/3)(1/3)(8/9)p_0 = (32/81)p_0 \]

and

\[ p_0 = (1 + 4/3 + 8/9 + 32/81)^{-1} = \frac{81}{293}. \]

Hence

\[ \mathbf{p} = \frac{81}{293}(1, 4/3, 8/9, 32/81)^t. \]

(b) Proportion of rejected calls \( = p_3 = \frac{32}{81} \times \frac{81}{293} = \frac{32}{293}. \)
3. (a) \( \lambda = 2 \) and \( \mu = 1 \) and we let \( a = \lambda / \mu = 2 \) then

\[
p_0 = \left( \sum_{k=0}^{s} a^k / k! \right)^{-1} = \left( \sum_{k=0}^{s} 2^k / k! \right)^{-1}.
\]

The mean number of customers served per hour is given by

\[
a(1 - B(s, a)) = a\left(1 - \frac{a^s}{s!} p_0\right) = 2 \left(1 - \frac{2^s}{s!} \left(\sum_{k=0}^{s} \frac{2^k}{k!}\right)^{-1}\right).
\]

Hence the net profit is given by

\[
C(s) = 2.5 \times 2 \left(1 - \frac{2^s}{s!} \left(\sum_{k=0}^{s} \frac{2^k}{k!}\right)^{-1}\right) - 1 \times s.
\]
(b) We note that
\[ \lim_{s \to \infty} \left( \sum_{k=0}^{s} \frac{2^k}{k!} \right)^{-1} = e^{-2} \text{ and } \lim_{s \to \infty} \frac{2^s}{s!} = 0. \]

Hence
\[ \lim_{s \to \infty} \frac{C(s)}{s} = \lim_{s \to \infty} \frac{-s}{s} = -1. \]

This means that when the number of servers \( s \) is large, the profit will be dominated by the operating cost \(-s\).

(c) We note that
\[ C(1) = 0.667, \quad C(2) = 1, \quad C(3) = 0.947, \quad C(4) = 0.524, \quad C(5) = -0.183. \]

The optimal number of servers is 2 and the optimal profit is $1 per hour.
4. (a) The generator matrix is a $4 \times 4$ matrix.

$$A = \begin{pmatrix}
-1 & \mu_1 & \mu_2 & 0 \\
1 & -1 - \mu_1 & 0 & \mu_2 \\
0 & 0 & -1 - \mu_2 & \mu_1 \\
0 & 1 & 1 & -\mu_1 - \mu_2
\end{pmatrix}. \tag{4.28}$$

(b) From the third equation, we have

$$p_{(0,1)} = \frac{\mu_1}{1 + \mu_2} p_{(1,1)}.$$  

From the fourth equation

$$p_{(1,0)} = (\mu_1 + \mu_2 - \frac{\mu_1}{1 + \mu_2}) p_{(1,1)}.$$  

From the first equation

$$p_{(0,0)} = \frac{\mu_1 \mu_2 (\mu_1 + \mu_2 + 2)}{1 + \mu_2} p_{(1,1)}.$$  

Finally we have

$$B(\mu_1, \mu_2) = p_{(1,1)} = \left(1 + \frac{\mu_1}{1 + \mu_2} + (\mu_1 + \mu_2 - \frac{\mu_1}{1 + \mu_2}) + \frac{\mu_1 \mu_2 (\mu_1 + \mu_2 + 2)}{1 + \mu_2}\right)^{-1}.$$  

(c) It is clear that if $\mu_1 > \mu_2$ then $B(\mu_1, \mu_2) < B(\mu_2, \mu_1)$.  

135
5. (a) The Markov chain is as follows.

\[ E_{0,n+1} \xrightarrow{\mu_1} E_{1,n} \xrightarrow{\mu_1} \cdots \xrightarrow{\mu_1} E_{n,1} \xrightarrow{\mu_1} E_{n+1,0} \]

(note that states \( E_{i,j} \) do not exist if \( i + j \neq n + 1 \))

(b) We have

\[ p_{1,n} = \rho p_{0,n+1}, \quad p_{2,n-1} = \rho^2 p_{0,n+1}, \quad \ldots \quad p_{n+1,0} = \rho^{n+1} p_{0,n+1} \]

where \( \rho = \mu_2 / \mu_1 \).

Hence we have

\[ p_{i,n+1-i} = \frac{(1-\rho)\rho^i}{1-\rho^{n+2}} \quad i = 0, 1, \ldots, n + 1. \]