10

Queueing Theory and Simulation

Queueing is an unavoidable part of modern life, and it seems that queueing is everywhere in banks, restaurants, airports, call centers, and leisure parks. Queueing also applies to internet queries, email message deliveries, telephone conversations, and many other applications. Thus, modeling queueing behavior can be of both theoretical interest and practical importance. The Danish engineer, A.K. Erlang, was the first to study the queueing process in the telephone system, and the relevant key ideas can be applied to many other queueing systems. This chapter introduces the fundamentals of queueing models.

10.1 Introduction

10.1.1 Components of Queueing

Apart from gaining insight into the queueing characteristics, one of the main aims of modeling queues is to improve the efficiency of queue management and to minimize the overall waiting costs and service costs. In order to model queueing systems properly, we have to identify their common components such as the rates of arrival, service, and departure.

From our experience and empirical observations, we can analyze the basic components or factors in a queue, and they are:

- Arrival rate (λ): This can be represented as the rate or the number of arrivals per unit time in a queue.
- Service rate (μ): This is how quickly a service (e.g. serving a cup of coffee) is provided, which can be represented as the number of customers served per unit time or the time taken to serve a customer.
- Number of servers (s): This can be represented as the number of counters in a supermarket or the number of cashiers in a bank. If each queue requires one server only, the number of servers can also be considered as the number of queues.

Optimization Techniques and Applications with Examples, First Edition. Xin-She Yang.

The above components are most important. However, in practice, factors such as capacity in a restaurant or a bank is also important. In this case, we have to consider three additional factors or components:

- Maximum queueing length ($L_{\max}$): The capacity of a popular restaurant or a coffee shop may be limited by the number of tables, thus there is a maximum queuing length in practice.
- Population size (N): This can be considered as the pool for potential customers that can arrive at the queueing system. In practice, the population size is always finite, but in the simplest case of queue modeling, we can assume that the population size is infinite.
- Queueing rule (R): The simplest queueing rule can be first-come and first-serviced, that is, a first-in first-out (FIFO) rule. Other rules can also apply when appropriate. For example, LIFO means last in and first out, and SIRO means service in random order.

In terms of practical queue settings and managements, there are usually four possibilities:

1) Single phase and single server: This is the simplest case where there is one queue with one server. Examples include queues at a photocopier machine, ATM, or a small corner shop.
2) Single phase and multiservers: This corresponds to a queue with multiple servers. For example, queueing in banks and some information centers seems to use this popular system. In this case, each customer in the queue is given a number and once a server is available, the next customer in the system is called to be served. For example, many information desk systems belong to this category.
3) Multiphases and single server: Sometimes, the service can be more complicated and thus can be divided into different stages or phases. For example, a drive-in McDonald, a customer in the driving queue orders first, then drives onto the next window to wait and collect the meal. This also applies to many services such as restaurants, student applications where tasks may take more than one stage or require two or more servicing facilities.
4) Multiphases and multi-servers: This represents a very generalized case where multiple service stages are need to serve each customer and there are multiple servers available for the service. A good example is at the airport where customers check in first, then go through the security, and then go to the boarding gates. At each phase, there are queues and each phase has multiple servers.

10.1.2 Notations

In order to model a queueing system properly, the standard notations introduced by D. G. Kendall are usually used.

In essence, Kendall's notation consists of six parts, in general, which can be written as

$$A/B/s/L/P/R. \tag{10.1}$$

Here, the first part A denotes the arrival model. For example, if the arrival probability obeys a Poisson's distribution, which is also Markovian, we can write "A" as "M." Here, the main characteristics of a Markovian process is that it is a memoryless process and thus the arrivals should be independent of each other.

The second part B denotes the service time distribution. For example, the most commonly used distribution for service time is exponential, which is also Markovian. The third part "s" denotes the number of servers, which is just a positive integer. The fourth part "L" denotes the queueing capacity or the maximum queueing length. If not given explicitly, it is usually assumed that the capacity is infinite. In addition, the fifth part "P" denotes the population size, and it can be assumed to be infinite if not stated explicitly.

The final part is the queueing rule "R" or queueing discipline. The default rule is FIFO. In case of $L = \infty$, $P = \infty$, and D = FIFO, Kendall's notation can be often simplified to only the first three letters. Thus, the simplest queue model is

$$M/M/1, \tag{10.2}$$

which corresponds to the case of a single server with both arrival and service being Markovian. The implicit assumptions are that both the capacity and population size are infinite, while the queueing rule is FIFO.

To model a queueing process with Poisson arrivals and exponential service time for 5 servers with a population size of 100 and each server has a capacity of 15 with a queue discipline of FIFO, the system model can be written compactly as M/M/5/15/100/FIFO or simply as M/M/5/15/100.

Kendall's notation can be used to describe almost all queueing processes. However, in our present discussions here, we will focus on a single-phased queue with a single server. Thus, the following additional assumptions are made:

- The arrival of each individual is independent, thus bulk arrivals (e.g. a coach of tourists, a group of students) are not allowed.
- The service time is also independent. That is, the service time of the previous customer is independent of the service time of the next customer.
- The whole queuing system is stable in the sense that the arrival and service probability distributions remain unchanged.
- The queueing discipline is simply "first come, first served" (i.e. FIFO).

With these assumptions, the model can be simplified and we will not explicitly discuss these assumptions any further in the rest of this chapter.

10.2 Arrival Model

The most widely used model for arrivals is the Poisson process. Let us first review the Poisson distribution before modeling queues.

10.2.1 Poisson Distribution

The Poisson distribution can be thought of as the limit of the binomial distribution when the number of trial is very large and the probability is sufficiently small. Typically, it is concerned with the number of events that occur in a certain time interval (e.g. number of telephone calls in an hour) or spatial area. The Poisson distribution can be written as

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad (\lambda > 0), \tag{10.3}$$

where $x = 0, 1, 2, \ldots, n$ and λ is the mean of the distribution.

Obviously, the sum of all the probabilities must be equal to one. That is,

$$\begin{aligned}\sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} &= \frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \frac{\lambda^3 e^{-\lambda}}{3!} + \cdots \\ &= e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots\right. \\ &= e^{-\lambda} e^{\lambda} = e^{-\lambda+\lambda} = e^0 = 1. \end{aligned} \tag{10.4}$$

Many stochastic processes such as the number of phone calls in a call center, number of earthquakes in a given period, and the number of cars passing through a junction obey the Poisson distribution. Let us look at an example.

Example 10.1 Suppose you receive one email per hour on average. If you attend a 1-hour lesson, what is the probability of receiving exactly two emails after the lesson? What is the probability of no email at all?

Since the distribution is Poisson with $\lambda = 1$, the probability of receiving two emails is

$$P(X = 2) = \frac{\lambda^2 e^{-\lambda}}{2!} = \frac{1^2 e^{-1}}{2!} \approx 0.184.$$

The probability of no email is

$$P(X = 0) = \frac{\lambda^0 e^{-\lambda}}{0!} = \frac{1^0 e^{-1}}{0!} \approx 0.368.$$

The probability of receiving exactly one email is

$$P(X = 1) = \frac{\lambda^1 e^{-\lambda}}{1!} = \frac{1^1 e^{-1}}{1} \approx 0.368.$$

On the other hand, what is the probability of receiving two or more emails? This means $X = 2, 3, 4, \ldots$, which have an infinite number of terms. That is,

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + \cdots,$$

but how do we calculate this probability? Since the total probability is one or

$$\begin{aligned} 1 &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + \cdots \\ &= P(X = 0) + P(X = 1) + P(X \geq 2), \end{aligned}$$

we have

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0.184 - 0.368 \approx 0.264.$$

That is, the probability of receiving two or more emails is about 0.264.

Furthermore, what is the probability of receiving exactly one email in a 15-minute interval?

We already know that $\lambda = 1$ for one hour, so $\lambda_* = 1/4 = 0.25$ for a 15-minute period. The probability of receiving exactly one email in a 15-minute period is

$$P(X = 1) = \frac{\lambda_*^1 \mathrm{e}^{-\lambda_*}}{1!} = \frac{0.25^1 \mathrm{e}^{-0.25}}{1} \approx 0.195.$$

In this example, we have used $\lambda_* = \lambda t$ where t is the time interval in the same time unit when defining λ. In general, we should use λt to replace t when dealing with the arrivals in a fixed period t. Thus, the Poisson distribution becomes

$$P(X = n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}. \tag{10.5}$$

Therefore, in the above example, if the lesson is a two-hour session, the probability of getting exactly one email after the two-hour session is

$$P(X = 1) = \frac{(1 \times 2)^1 \mathrm{e}^{-1 \times 2}}{1!} \approx 0.271.$$

Using the definitions of mean and variance, it is straightforward to prove that $E(X) = \lambda$ and $\sigma^2 = \lambda$ for the Poisson distribution. For example, the mean or expectation $E(X)$ can be calculated by

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} xP(X = x) = \sum_{x=0}^{\infty} x \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} \\ &= 0 \times \mathrm{e}^{-\lambda} + 1 \times (\lambda \mathrm{e}^{-\lambda}) + 2 \times \left(\frac{\lambda^2 \mathrm{e}^{-\lambda}}{2!} \right) + 3 \times \left(\frac{\lambda^3 \mathrm{e}^{-\lambda}}{3!} + \cdots \right) \\ &= \lambda \mathrm{e}^{-\lambda} \left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots \right] = \lambda \mathrm{e}^{-\lambda} \mathrm{e}^{+\lambda} = \lambda. \end{aligned}$$

The parameter λ controls the location of the peak and the spread (or standard deviation). In essence, λ describes a Poisson distribution uniquely, so some textbooks use Poisson(λ) to denote a Poisson distribution with parameter λ.

Poisson distributions have an interesting property. For two independent random variables U and V that obey Poisson distributions: Poisson(λ_1) and Poisson(λ_2), respectively, $S = U + V$ obeys Poisson($\lambda_1 + \lambda_2$). From the basic Poisson distribution, we know that

$$\begin{aligned} P(S=n) &= \sum_{k=0}^{n} P(U=k, V=n-k) = \sum_{k=1}^{n} P(U=k)P(V=n-k) \\ &= \sum_{k=0}^{n} \frac{\lambda_1^k \mathrm{e}^{-\lambda_1}}{k!} \cdot \frac{\lambda_2^{n-k} \mathrm{e}^{-\lambda_2}}{(n-k)!} = \mathrm{e}^{-(\lambda_1+\lambda_2)} \frac{1}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k} \\ &= \mathrm{e}^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^n}{n!} = \mathrm{e}^{-\lambda} \frac{\lambda^n}{n!}, \end{aligned} \tag{10.6}$$

where $\lambda = \lambda_1 + \lambda_2$. In the above calculations, we have used the fact that U and V are independent (and thus the joint probability is the product of their probabilities). We have also used the binomial expansions

$$(\lambda_1 + \lambda_2)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k}. \tag{10.7}$$

Let us look at an example.

Example 10.2 For two students A and B, Student A receives on average 1 email per hour and Student B receives on average 1.5 emails per hour. What is the probability of receiving a total of exactly 4 emails in 1 hour?

Since $\lambda_1 = 1$ and $\lambda_2 = 1.5$, we have that A obeys Poisson(1) and B obeys Poisson(1.5), so A + B obeys Poisson($\lambda_1 + \lambda_2$) = Poisson(2.5). The probability of receiving exactly 4 emails is

$$P(A+B=4) = \frac{(\lambda_1+\lambda_2)^4 \mathrm{e}^{-(\lambda_1+\lambda_2)}}{4!} = \frac{2.5^4 \mathrm{e}^{-2.5}}{4!} \approx 0.134.$$

In addition, the probability of none of them receiving any email is

$$P(A+B=0) = \frac{2.5^0 \mathrm{e}^{-2.5}}{0!} \approx 0.082.$$

The arrival model of a Poisson distribution means that the inter-arrival times obey an exponential distribution

$$p(x) = \lambda \mathrm{e}^{-\lambda x} \quad (x > 0), \tag{10.8}$$

with a mean of $1/\lambda$.

10.2.2 Inter-arrival Time

For the ease of our discussion here, let us define the basic concepts of Poisson processes first. Loosely speaking, a Poisson process is an arrival process in which the inter-arrival times are independent and identically distributed (iid) random variables, and such inter-arrivals are exponentially distributed $f(t) = \lambda e^{-\lambda t}$.

An interesting property of the Poisson process is the memoryless properties of the inter-arrival time derived from the exponential distribution.

For a Poisson arrival process with an arrival sequence $(A_1, A_2, \ldots, A_k, \ldots)$ with the arrival time T_k for the kth arrival, the inter-arrival time is $\Delta T_k = T_{k+1} - T_k$ between arrival A_k and A_{k+1}.

From the Poisson's distribution, the first arrival at t means that there is no arrival in the interval $[0, t]$, thus its probability is

$$\text{Poisson}(T_1 > t) = P(n = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}. \tag{10.9}$$

This is also true that

$$P(\Delta T_k > t) = P[n(T_k + t) - n(T_k) = 0] = P(n = 0) = e^{-\lambda t}, \tag{10.10}$$

which means that it has a memoryless property. Therefore, the inter-arrival time obeys an exponential distribution with parameter λ.

As the total probability of all inter-arrival times must be one, the probability density function $f(t)$ should be divided by a scaling factor or normalization factor $\int_0^\infty e^{-\lambda t} dt = 1/\lambda$. Thus, we have

$$f(t) = \lambda e^{-\lambda t}. \tag{10.11}$$

Here, the simultaneous multiple arrivals are not allowed. For a higher arrival rate, as long as the time can be subdivided into sufficiently many small intervals, arrivals can always occur in sequence and thus the above results still hold.

10.3 Service Model

The most widely used service time model is the exponential distribution.

10.3.1 Exponential Distribution

The exponential distribution has the following probability density function

$$f(x) = \mu e^{-\mu x} \quad (\mu > 0, \quad x > 0), \tag{10.12}$$

and $f(x) = 0$ for $x \leq 0$. Its mean and variance are

$$\frac{1}{\mu}, \qquad \sigma^2 = \frac{1}{\mu^2}. \tag{10.13}$$

The expectation $E(X)$ of an exponential distribution is

$$E(X) = \int_{-\infty}^{\infty} x\mu e^{-\mu x} dx = \int_{0}^{\infty} x\mu e^{-\mu x} dx$$
$$= \left[-xe^{-\mu x} - \frac{1}{\mu}e^{-\mu x}\right]_0^{\infty} = \frac{1}{\mu}. \tag{10.14}$$

For $E(X^2)$, we have

$$E(X^2) = \int_0^{\infty} x^2 \mu e^{-\mu x} dx = \left[-x^2 e^{-\mu x}\right]_0^{\infty} + 2\int_0^{\infty} xe^{-\mu x} dx$$
$$= \left[-x^2 e^{-\mu x}\right]_0^{\infty} + \left[-\frac{2x}{\mu}e^{-\mu x} - \frac{2}{\mu^2}e^{-\mu x}\right]_0^{\infty} = \frac{2}{\mu^2}.$$

Here, we have used the fact that x and x^2 grow slower than $\exp(-\mu x)$ decreases. That is, $x\exp(-\mu x) \to 0$ and $x^2\exp(-\mu x) \to 0$ when $x \to \infty$.

From

$$E(X^2) = [E(X)]^2 + \sigma^2 = \frac{1}{\mu^2} + \text{Var}(X), \tag{10.15}$$

we have

$$\text{Var}(X) = \frac{2}{\mu^2} - \left(\frac{1}{\mu}\right)^2 = \frac{1}{\mu^2}. \tag{10.16}$$

Exponential distributions are widely used in queuing theory and simulating discrete events. As we have seen earlier, the arrival process of customers in a bank is a Poisson process and the time interval between arrivals (or inter-arrival time) obeys an exponential distribution.

The service time of a queue typically obeys an exponential distribution

$$P(t) = \begin{cases} \mu e^{-\mu t}, & t \geq 0, \\ 0, & t < 0, \end{cases} \tag{10.17}$$

where μ is the average number of customer served per unit time, and thus $\tau = 1/\mu$ is the mean service time. Thus, the service time as a random variable X less than some time (t) is the cumulative distribution

$$P(X \leq t) = \int_{-\infty}^{t} \mu e^{-\mu\tau} d\tau = \int_0^t \mu e^{-\mu\tau} d\tau = -e^{-\mu\tau}\Big|_0^t = 1 - e^{-\mu t}. \tag{10.18}$$

Obviously, as the total probability sum must be one, the probability of service time longer than a certain time is

$$P(X \geq t) = 1 - P(X \leq t) = e^{-\mu t}. \tag{10.19}$$

Example 10.3 If you are in a queue in a bank, you observe that it takes 2 minutes on average to service a customer. The service time obeys a cumulative distribution function (CDF)

$$P(X \le t) = 1 - \mathrm{e}^{-\mu t},$$

what is the probability of taking less than 1 minute to the next customer?

We know that $\mu = 1/2 = 0.5$ (2 minutes per customer or 0.5 customer per minute), so the probability is thus

$$P(t \le 1) = 1 - \mathrm{e}^{-0.5\times 1} \approx 0.393.$$

The probability of taking longer than 5 minutes is

$$P(t \ge 5) = \mathrm{e}^{-\mu t} = \mathrm{e}^{-0.5\times 5} = \mathrm{e}^{-2.5} \approx 0.008.$$

The arrival processes of many services can be approximated by this model.

10.3.2 Service Time Model

Since the service time T is exponentially distributed, it has a memoryless property. From the conditional probability

$$P(T > \tau + t) = P(T > \tau)P(T > t), \tag{10.20}$$

we have

$$P(T > \tau)P(T > t) = \mathrm{e}^{-\mu\tau}\mathrm{e}^{-\mu t} = \mathrm{e}^{-\mu(\tau+t)} = P(T > \tau + t), \tag{10.21}$$

where we have not considered the scaling/normaliation factors for simplicity. The above equation means that

$$P(T > \tau + t | T > t) = P(T > \tau), \tag{10.22}$$

which means that the time difference τ is independent of the time t, so it does not matter when the time starts (thus it is memoryless). This also means that the past has no effect on the future.

It is worth pointing out that the exponential distribution is the only memoryless distribution for continuous random variables.

10.3.3 Erlang Distribution

As the service time is a random variable S, which is exponentially distributed

$$h(t) = \begin{cases} \mu\mathrm{e}^{-\mu t}, & t \ge 0, \\ 0, & t < 0. \end{cases} \tag{10.23}$$

the joint service time distribution of two customers or two services with two iid random variables S_1 and S_2 can be calculated by the convolution integral

$$(h * h)(t) = \int_{-\infty}^{\infty} h(t-\tau)h(\tau)\mathrm{d}\tau. \tag{10.24}$$

Thus, the joint probability density of $S_1 + S_2$ is

$$h_2 = \begin{cases} \mu^2 t e^{-\mu t}, & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{10.25}$$

Similarly, the sum of random variables $S_1 + S_2 + S_3$ of the same iid is the convolution of h with h_2. Following the same line of thinking, we can conclude that the joint probability density of

$$T_s = \sum_{k=1}^{n} S_k = S_1 + S_2 + \cdots + S_n \tag{10.26}$$

can be calculated by

$$h_n = \begin{cases} \frac{\mu^n t^{n-1}}{(n-1)!} e^{-\mu t}, & t \geq 0, \\ 0, & t < 0, \end{cases} \tag{10.27}$$

which is the Erlang distribution for $t \geq 0$ and $\mu \geq 0$. This distribution is also a special case of the gamma distribution.

10.4 Basic Queueing Model

10.4.1 M/M/1 Queue

Now let us focus on the M/M/1 queue model with an arrival rate λ (obeying the Poisson process) and the service time μ obeying the exponential distribution (see Figure 10.1).

Obviously, if $\lambda > \mu$, the system will not be stable because the service is slower than the arrival and the queueing length will thus grow unboundedly (leading to infinity). Thus, in order to analyze the stable characteristics of a proper queueing system, it requires that the ratio $\rho = \lambda/\mu$ is less than 1. That is,

$$\rho = \frac{\lambda}{\mu} < 1, \tag{10.28}$$

which is a performance measure for the queueing system. In fact, ρ can be considered as the average server utilization because $\rho = 1$ means that the server is almost 100% busy (on average). Alternatively, ρ can be considered as the number of customers or the probability in service for the server. In other words, the probability of no customer at all is $p_0 = 1 - \rho$.

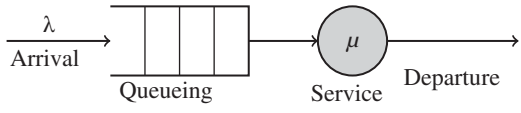

Figure 10.1 A simple queue representation of M/M/1.

The key idea here is to look at the long-time limiting behavior of the system (not the short-time transient behavior), which allows us to identify the main characteristics of the stable, steady-state queueing system. Thinking along this line, the probability of n customers is

$$p_n = \rho^n p_0 = \rho^n(1-\rho), \tag{10.29}$$

which is essentially a geometric distribution. It is worth pointing out that this probability is the limiting probability.

From the definition of the mean for a probability mass function, we can estimate the mean number of customers in the system as

$$\begin{aligned} E(n) &= \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{\infty} n\rho^n p_0 = p_0 \sum_{n=0}^{\infty} n\rho^n \\ &= p_0[1+\rho+2\rho^2+\cdots+n\rho^n+\cdots] = (1-\rho)\cdot\frac{\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho}, \end{aligned} \tag{10.30}$$

where we have used

$$\sum_{n=0}^{\infty} n\rho^n = \frac{\rho}{(1-\rho)^2}, \quad p_0 = 1-\rho. \tag{10.31}$$

Since ρ is a constant, the above equality can be proved by the following steps:

$$\begin{aligned} \sum_{n=0}^{\infty} n\rho^n &= \sum_{n=0}^{\infty} \rho\frac{d\rho^n}{d\rho} = \rho\frac{d}{d\rho}\left[\sum_{n=0}^{\infty}\rho^n\right] \\ &= \rho\frac{d}{d\rho}\left[\frac{1}{1-\rho}\right] = \rho\cdot\frac{1}{(1-\rho)^2} = \frac{\rho}{(1-\rho)^2}. \end{aligned} \tag{10.32}$$

Thus, the average number of customers in the system can be estimated by

$$L = E(n) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}. \tag{10.33}$$

It is worth pointing out that L will become infinite as $\rho \to 1$ or $\mu \to \lambda$. In fact, many quantities many diverge when $\rho \to 1$.

Since the number of customers in service is $\rho = \lambda/\mu$, the number of customers (L_q) in the queue (not served yet) can be estimated as

$$\begin{aligned} L_q &= L - \rho = \frac{\rho}{1-\rho} - \rho = \frac{\rho}{1-\rho} - \frac{\rho(1-\rho)}{1-\rho} \\ &= \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}. \end{aligned} \tag{10.34}$$

However, we have to be careful in order to calculate the average time (W) of a customer spent in the system correctly. One naive (but incorrect) approach is to use $W = L\tau = \lambda/\mu(\mu-\lambda)$, where $\tau = 1/\mu$ is the average service time.

The correct approach is as follows: since the service time is exponentially distributed with parameter μ, for a queueing system with n customers, its total service time obeys the Erlang distribution as discussed earlier:

$$q_n(t) = \frac{\mu^n}{(n-1)!} t^{n-1} \mathrm{e}^{-\mu t}. \tag{10.35}$$

Let T_w be the waiting time and distribution of the waiting time $Q(t) = p(T_w \leq t)$, so the no waiting at all means that

$$Q(0) = p_0 = 1 - \rho. \tag{10.36}$$

Since the service time is exponentially distributed and memoryless, the remaining service time of the customer being served should obey the same exponential distribution. For a detailed discussion on this issue, readers can refer to the book by Bhat and Miller (2002). For a small time increment δt for any time $t > 0$, the probability increment of T_w in the interval $[t, t + \delta t]$ (or $t < T_w \leq t + \delta t$) is given by

$$\begin{aligned}
\delta Q(t) &= \sum_{n=1}^{\infty} p_n q_n(t) \delta t = \sum_{n=1}^{\infty} (1-\rho)\rho^n \frac{\mu^n}{(n-1)!} t^{n-1} \mathrm{e}^{-\mu t} \delta t \\
&= (1-\rho)\mathrm{e}^{-\mu t} \left[\sum_{n=1}^{\infty} \frac{(\rho\mu)^n t^{n-1}}{(n-1)!} \right] \delta t = (1-\rho)\mathrm{e}^{-\mu t} \sum_{n=1}^{\infty} \frac{\lambda^n t^{n-1}}{(n-1)!} \delta t \\
&= (1-\rho)\mathrm{e}^{-\mu t} \left[\lambda \sum_{n=1}^{\infty} \frac{(\lambda t)^{n-1}}{(n-1)!} \right] \delta t \\
&= (1-\rho)\mathrm{e}^{-\mu t} \left[\lambda \mathrm{e}^{\lambda t} \right] \delta t = (1-\rho)\lambda \mathrm{e}^{-(1-\rho)\mu t} \delta t,
\end{aligned} \tag{10.37}$$

where we have used $\lambda = \rho\mu$ and the Taylor series of $\exp(\lambda t)$,

$$\exp(\lambda t) = \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} = \sum_{n=1}^{\infty} \frac{(\lambda t)^{n-1}}{(n-1)!}, \quad k = n - 1. \tag{10.38}$$

Thus, the probability density function of the waiting time in queue (T_{w}) is

$$P_{\mathrm{w}}(t) = (1-\rho)\lambda \mathrm{e}^{-\mu(1-\rho)t} = \lambda(1-\rho)\mathrm{e}^{-(\mu-\lambda)t} \quad (t > 0). \tag{10.39}$$

Using $\delta t = \mathrm{d}t$, the distribution $Q(t)$ becomes

$$\begin{aligned}
Q(T_{\mathrm{w}} \leq t) &= p_0 + \int_0^t [\delta Q] \mathrm{d}t = 1 - \rho + \lambda(1-\rho) \int_0^t \mathrm{e}^{-\mu(1-\rho)t} \mathrm{d}t \\
&= 1 - \rho \mathrm{e}^{-\mu(1-\rho)t} = 1 - \frac{\lambda}{\mu} \mathrm{e}^{-(\mu-\lambda)t},
\end{aligned} \tag{10.40}$$

which is the cumulative distribution function for waiting time $T_{\mathrm{w}} < t$. Thus, the probability of waiting time $T_{\mathrm{w}} > t$ is the complementary part, which means

$$H(T_{\mathrm{w}} > t) = 1 - Q(T_{\mathrm{w}} \leq t) = \rho \mathrm{e}^{-\mu(1-\rho)t} = \frac{\lambda}{\mu} \mathrm{e}^{-(\mu-\lambda)t}. \tag{10.41}$$

Finally, the mean waiting time can be calculated by

$$\begin{aligned} W_{\mathrm{q}} = E(T_{\mathrm{w}}) &= \int_0^\infty t P_{\mathrm{w}}(t)\mathrm{d}t = \int_0^\infty t\lambda(1-\rho)\mathrm{e}^{-\mu(1-\rho)t}\mathrm{d}t \\ &= \lambda(1-\rho)\left[\int_0^\infty t\mathrm{e}^{-\mu(1-\rho)t}\mathrm{d}t\right] \\ &= \lambda(1-\rho)\left[\frac{1}{\mu^2(1-\rho)^2}\right] = \frac{\lambda}{\mu^2(1-\rho)} = \frac{\rho}{\mu(1-\rho)}. \end{aligned} \tag{10.42}$$

Therefore, the average time of a customer spent in the queueing system (waiting time W_{q} plus the service time τ) is thus

$$\begin{aligned} W &= W_{\mathrm{q}} + \tau = \frac{\rho}{\mu(1-\rho)} + \frac{1}{\mu} \\ &= \frac{1}{\mu}\left[\frac{\rho}{1-\rho} + 1\right] = \frac{1}{\mu(1-\rho)} = \frac{1}{(\mu-\lambda)}. \end{aligned} \tag{10.43}$$

Example 10.4 A small shop usually has an arrival rate of about 20 customers per hour, while the counter can typically serve a customer every 2 minutes. Assuming the arrivals are Poisson and the service time is exponential, calculate the relevant quantities in the shop system.

The arrival rate is $\lambda = 20\mathrm{h}^{-1}$ and the service rate is $\mu = 60/2 = 30\mathrm{h}^{-1}$, so we have

$$\rho = \frac{20}{30} = \frac{2}{3} \approx 0.66667.$$

The expected queue length is

$$L = \frac{\rho^2}{1-\rho} = \frac{(2/3)^2}{1-\frac{2}{3}} = 1.333.$$

The mean waiting time is

$$W_{\mathrm{q}} = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{20}{30(30-20)} = \frac{1}{15}\ (\text{hour}) = 4\ (\text{minute}).$$

The probability of no waiting at all is

$$P_0 = 1 - \rho = \frac{1}{3}.$$

The probability of waiting longer than t is given by

$$P(T_{\mathrm{w}} > t) = \rho\exp[-(\mu-\lambda)t] \quad (t > 0).$$

For example, the probability of waiting longer than 5 minute is

$$P\left(T_{\mathrm{w}} > \frac{5}{60}\right) = \frac{2}{3}\exp\left[-(30-20)\times\frac{5}{60}\right] = 0.29.$$

Similarly, the probability of waiting longer than 15 minute is

$$P\left(T_{\mathrm{w}} > \frac{15}{60}\right) = \frac{2}{3}\exp\left[-(30-20)\times\frac{15}{60}\right] \approx 0.05.$$

It is worth pointing out that, from Eqs. (10.34) and (10.43), we have

$$W = \frac{1}{(\mu-\lambda)} = \frac{1}{\lambda}\frac{\lambda}{(\mu-\lambda)} = \frac{L}{\lambda} \tag{10.44}$$

or

$$L = \lambda W, \tag{10.45}$$

which is the well-known Little's law.

In addition, from $L_{\mathrm{q}} = \rho^2/(1-\rho)$ and $W_{\mathrm{q}} = \rho/[\mu(1-\rho)]$, we have

$$L_{\mathrm{q}} = \frac{\rho^2}{(1-\rho)} = \frac{\lambda\rho}{\mu(1-\rho)} = \lambda W_{\mathrm{q}}, \tag{10.46}$$

which means that Little's law also applies to the waiting time in queue and the number of customers in queue. Little's law is also valid for a queueing system with s servers (i.e. M/M/s model).

10.4.2 M/M/s Queue

For the extension of the results about the M/M/1 queue model with a single server to $s \geq 1$ servers, formal derivations usually require the full Markovian model, limiting probability and balance equations. Such derivations are beyond the scope of this book. Therefore, we will state the main results without formal proofs.

For a queueing system with s servers and n customers, it is obvious that there is no need to queue if $n \leq s$. The average utility measure is

$$U_\rho = \frac{\rho}{s} = \frac{\lambda}{s\mu} < 1. \tag{10.47}$$

It is worth pointing out that it requires $U_\rho = \rho/s < 1$ (not $\rho < 1$) here.

The probability of zero customer is given by

$$p_0 = \left[\sum_{k=0}^{s-1}\frac{\rho^k}{k!} + \frac{\rho^s}{s!}\left(\frac{s\mu}{s\mu-\lambda}\right)\right]^{-1} = \frac{1}{\rho^s/s!(1-U) + \sum_{k=0}^{s-1}\rho^k/k!}. \tag{10.48}$$

The probability of n customers is

$$p_n = \begin{cases} \frac{\rho^n}{n!}p_0 & (n \leq s), \\ \frac{\rho^n}{s!\,s^{n-s}}p_0 & (n > s). \end{cases} \tag{10.49}$$

The average number of customers in the system is

$$L = \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{s-1} np_n + \sum_{n=s}^{\infty} np_n = p_0 \left[\sum_{n=0}^{s-1} n\frac{\rho^n}{n!} + \sum_{n=s}^{\infty} s\frac{\rho^n}{s!s^{n-s}} \right]$$
$$= \frac{p_0}{s!} \left[\sum_{n=0}^{s-1} n\rho^n + s^s \sum_{n=s}^{\infty} n\frac{\rho^n}{s^n} \right] = \rho + p_0\rho \left[\frac{\rho^s}{(s-1)!(s-\rho)^2} \right]$$
$$= \rho + p_0 \left[\frac{\rho^{s+1}}{(s-1)!(s-\rho)^2} \right] = \frac{\lambda}{\mu} + p_0 \left[\frac{(\lambda/\mu)^s(\lambda\mu)}{(s-1)!(s\mu-\lambda)^2} \right]. \quad (10.50)$$

This means that the average number of customers in queue is

$$L_q = L - \rho = \frac{(\lambda/\mu)^s(\lambda\mu)}{(s-1)!\,(s\mu-\lambda)}. \quad (10.51)$$

The probability of waiting in the queue when $n > s$ is

$$P_{n>s} = p_0 \frac{\rho^{s+1}}{s!s(1-\rho/s)}. \quad (10.52)$$

The waiting time can be calculated by

$$W_{\mathrm{q}} = \frac{\rho^s p_0}{s!(s\mu)(1-\rho/s)^2} = \frac{\rho^s p_0}{s!(s\mu)} \frac{s^2}{(s-\rho)^2}. \quad (10.53)$$

In addition, the waiting time probability distribution $T_{\mathrm{w}} < t$ is given by

$$Q(T_{\mathrm{w}} < t) = 1 - \frac{\rho^s p_0}{s!(1-\rho/s)} \mathrm{e}^{-(s\mu-\lambda)t} = 1 - \frac{\rho^s p_0}{s!(1-\rho/s)} \mathrm{e}^{-\mu(s-\rho)t}, \quad (10.54)$$

thus the probability of waiting longer than t is given by

$$Q_{\mathrm{q}}(T_{\mathrm{w}} > t) = 1 - Q(T_{\mathrm{w}} < t) = \frac{\rho^s p_0}{s!(1-\rho/s)} \mathrm{e}^{-\mu(s-\rho)t}. \quad (10.55)$$

It is straightforward to verify that the above formulae will become the results for M/M/1 when $s = 1$ if we use $0! = 1$.

The average time of a customer spent in the system can be obtained using Little's law, and we have

$$W = \frac{L}{\lambda}. \quad (10.56)$$

Let us revisit the small shop example discussed earlier.

Example 10.5 Suppose the shop is getting busier at weekends, the arrival rate becomes 50 per hour, the shop has to open 2 counters at the same time and each counter can still serve (on average) one customer every 2 minutes. Assuming the other conditions remain the same (Poisson arrival, exponential service time), what are the new performance measures for the shop at weekends?

Now we have $\lambda = 50\text{h}^{-1}$, and $\mu = 60/2 = 30/\text{h}^{-1}$, and $s = 2$. The utility measure is

$$U = \frac{\lambda}{s\mu} = \frac{50}{2 \times 30} = \frac{5}{6} < 1,$$

with

$$\rho = \frac{50}{30} = \frac{5}{3}.$$

The probability of no customer is

$$p_0 = \left[\frac{\rho^s}{s!} \frac{s}{(s-\rho)} + \sum_{k=0}^{1} \frac{\rho^k}{k!} \right]^{-1} = \frac{1}{11} \approx 0.090\,909.$$

The expected average number of customers in the shop is

$$L = \rho + p_0 \left[\frac{\rho^{s+1}}{(s-1)!(s-\rho)^2} \right] = \frac{5}{3} + \frac{1}{11} \left[\frac{(5/3)^3}{(2-1)!(2-5/3)^2} \right] \approx 5.45$$

The expected waiting time in the shop is

$$W_q = \frac{\rho^s p_0}{s!(s\mu)(1-\rho/s)^2} = \frac{(5/3)^2 \times 1/11}{2!(2 \times 30)(1-5)/(3 \times 2)^2} = 0.075\,76 \text{ (hour)},$$

which is about 4.5 minute. The probability of no waiting at all $t = 0$ is

$$\begin{aligned} Q(0) &= 1 - \frac{\rho^s p_0}{s!(1-\rho/s)} = 1 - \frac{\rho^s p_0}{s!(1-U)} \\ &= 1 - \frac{(5/3)^2 \times 1/11}{2!(1-5/6)} \approx 0.2424. \end{aligned}$$

The probability of waiting longer than 10 minutes (or 10/60 hours) is

$$\begin{aligned} Q(T_w > 10 \text{ minute}) &= \frac{\rho^s p_0}{s!(1-\rho/s)} e^{-\mu(s-\rho)t} \\ &= \frac{(5/3)^2 \times 1/11}{2!(1-5/3 \times 2)} \exp\left[-30 \left(2 - \frac{5}{3} \right) \frac{10}{60} \right] \approx 0.14. \end{aligned}$$

Now let us discuss an important property about the queuing systems.

10.5 Little's Law

As we have seen earlier, Little's law of a queuing system states that the average number L of items in the system is equal to the average arrival rate λ multiplied by the average time W that an item spends in the queuing system. That is,

$$L = \lambda W.$$

This simple law provides some good insight into queueing systems, and relevant quantities can be estimated without any detailed knowledge of a particular queueing system.

Example 10.6 For example, a system has an arrival rate of 2 items per minute, and the average queue length is 8. What is the average waiting time for an item?

We know that $\lambda = 2$ and $L = 8$, so

$$W = \frac{L}{\lambda} = \frac{8}{2} = 4,$$

which means that an item usually waits 4 minutes before being processed.

The above results obtained for an M/M/1 or M/M/s system provide some useful insight into the queueing systems. However, real-world queues are more complicated because the assumptions we made may not be true. For example, real-world queues at a restaurant can be time-dependent because there are more customers at lunch time and in the evening. Therefore, the stable assumptions may not be true at all. More generalized queueing models should be considered. Interested readers can refer to more advanced literature on queueing theory and applications, listed at the end of this chapter.

10.6 Queue Management and Optimization

Queue management is crucially important to the success of many organizations and applications, from retail business and event management to call centers and the Internet routing. The conditions in real-world queues are dynamic, time-varying with uncertainty. Even though the mathematical models may no longer be valid, queues still have to be managed, and optimization still have to be carried out whenever appropriate.

The management of queues may include many aspects, from physical settings and structures to the estimation of key parameters and predictability of various quantities. For example, a business should observe and estimate the number of customers, queue length, waiting time, service time, and other quantities so as to be prepared for queueing management. Then, the number of servers should be able to vary so as to reduce the waiting time and queue length. Efficient queue management should aim to serve a majority (say 95%) within a fixed time limit (for example, 3 or 5 minutes). Customers' rating and satisfaction can be largely influenced by the waiting time and ease of exiting the queues. Amazon's online one-click checkout is a primary example for efficient queueing management.

In order to provide sufficiently accurate estimates of key parameters, it requires a multidisciplinary approach to use a variety of data and methodologies, including historical information, current arrival information, mobile

sensors and cameras, statistical analysis, forecasting, appointment systems, classification and clustering (sort queries from customers), machine learning, and artificial intelligence. Monitoring of queues and communications about the queue status, comfortable waiting environment, engagement and interactions with customers, and effective service mechanism are all part of queue management systems. A successful queue management system should be able to optimize customer experience so as to maintain a successful business in the long run.

Optimization can be carried out when estimating key parameters, dynamic allocation of servers, mining historical data, and predicting future trends.

Example 10.7 From the M/M/s queue theory discussed earlier, if the aim is to minimize the customer waiting W_q for given λ, μ and other parameters, we can adjust s so that

$$\text{minimize} \quad \frac{\rho^s p_0}{s!(s\mu)} \frac{s^2}{(s-\rho)^2}, \tag{10.57}$$

where s should be a positive integer $s \geq 1$. This is a nonlinear optimization problem, but it is not difficult to solve in principle because it is just a function of a single variable.

The main issue is that even with a good solution of s, it may be less useful in practice because the actual queue setting can be different such that the model is just a very crude approximation. However, in many applications, a simple estimate can be sufficient to provide enough information for proper queue management.

There is the queueing rule of thumb to estimate the servers needed for a particular setting. For a queue system with multiple servers s, if N is the total number of customers to be served during a total period of T, then we can use N/T to approximate the arrival rate λ. That is $\lambda = N/T$. In addition, the average service time τ can be estimated as $\tau = 1/\mu$. From the above discussion, we know that the queue system is valid and stable if

$$U = \frac{\lambda}{s\mu} < 1, \tag{10.58}$$

which becomes

$$U = \frac{N/T}{s\mu} = \frac{N\tau}{sT} < 1. \tag{10.59}$$

Based on this, Teknomo derived a queueing rule of thumb for the number of servers

$$s \geq \left\lceil \frac{N\tau}{T} \right\rceil, \tag{10.60}$$

which can be a handy estimate.

Example 10.8 For example, a supermarket can become very busy during lunch time. Suppose that there may be 500 customers for a 2-hour lunch period. If each customer can be served within 2 minutes at checkout, how many checkout counters should be available?

We know $N = 500$, $T = 2 \times 3600 = 7200$ seconds, and $\tau = 2 \times 60 = 120$ seconds, so we have

$$s = \left\lceil \frac{500 \times 120}{7200} \right\rceil = \left\lceil \frac{25}{3} \right\rceil = 9.$$

However, this simplified model does not give enough information about the queue length and waiting time. For such information, we need to use the complicated formulae discussed earlier in the chapter. Obviously, the dynamic, noisy nature of real-world settings requires an effective queue management system using real-time data and service management.

Exercises

10.1 An IT help desk typically receives 20 queries per hour and the help desk team can handle at most 30 queries per hour. Assuming this process obeys a Poisson model, write down the queue model for this process using Kendall's notation.

10.2 For the previous question, what is the average waiting time in the queue? What is the probability of no wait at all? What is the probability of waiting longer than 10 minutes?

10.3 A shop has the maximum of five counters and each counter can serve a customer every 2 minutes. The shop has 100 customers per hour, what is the average queue length if 3 counters are open? If the total number of customers in all the queues should not be more than 5, how many counters should be used?

10.4 A busy road has a traffic volume of about $Q = 250$ vehicles per hour. If a Poisson distribution is used, show that

$$P(t \geq h) = \mathrm{e}^{-h/T}, \quad P(t < h) = 1 - \mathrm{e}^{-h/T},$$

with $\lambda = Q/3600$ (cars per second) and the mean headway h (between successive cars) is $T = 1/\lambda = 3600/Q$. If it takes about 15 seconds to cross the road, what is the probability of no waiting at all? If some one walks slower and may take 20 seconds to cross the road, what is the probability of find a gap between 15 and 20 seconds?

10.5 A small shop has 5 parking spaces, and each customer takes on average about 10 minutes to shop. If there are 18 customers per hour driving to the shop, what is the probability of not finding a parking space upon arrival?

Bibliography

Bertsekas, D. and Gallager, R. (1992). *Data Networks*, 2e. Englewood Cliffs, NJ: Prentice Hall.

Bhat, U.N. (2008). *An Introduction to Queueing Theory: Modelling and Analysis in Applications*. Boston: Birkhäuser.

Bhat, U.N. and Miller, G.K. (2002). *Elements of Applied Stochastic Processes*, 3e. New York: Wiley.

Buzen, J.P. (1973). Computational algorithms for closed queueing networks with exponential servers. *Communications of the ACM* 16 (9): 527–531.

Daigle, J.N. (2010). *Queueing Theory and Applications to Packet Telecommunication*. New York: Springer.

Daley, D.J. and Servi, L.D. (1998). Idle and busy periods in stable M/M/k queues. *Journal of Applied Probability* 35 (4): 950–962.

Erlang, A.K. (1909). The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B* 20 (1): 33–39.

Gross, D. and Harris, C.M. (1998). *Fundamentals of Queueing Theory*. New York: Wiley.

Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29 (3): 567–588.

Harchol-Balter, M. (2013). *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge, UK: Cambridge University Press.

Jackson, J.R. (1957). Networks of waiting lines. *Operations Research* 5 (4): 518–521.

Kelly, F.P. (1975). Networks of queues with customers of different types. *Journal of Applied Probability* 12 (3), 542–554.

Kendall, D.G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Annals of Mathematical Statistics* 24 (3): 338–354.

Kingman, J.F.C. (2009). The first erlang century – and the next. *Queueing Systems* 63 (1): 3–4.

Lester, L. (2010). *Queueing Theory: A Linear Algebraic Approach*, 2e. New York: Springer.

Little, J.D.C. (1961). A proof for the queueing formula: $L = \lambda W$. *Operations Research* 9 (3): 383–387.

Mannering, F.L., Washburn, S.S., and Kilareshi, W.P. (2008). *Principles of Highway Engineering and Traffic Analysis*. Hoboken, NJ: Wiley.

Murdoch, J. (1978). *Queueing Theory: Worked Examples and Problems*. London UK: Palgrave Macmillan.

Newell, G.F. (1971). *Applications of Queueing Theory*. London: Chapman and Hall.

Saaty, T.L. (1961). *Elements of Queueing Theory*. New York: McGraw-Hill.

Simchi-Levi, D. and Trick, M.A. (2011). Introduction to Little's law as viewed on its 50th anniversary. *Operations Research* 59 (3): 535.

Teknomo, K. (2012). Queuing rule of thumb based on M/M/s queuing theory with applications in construction management. *Civil Engineering Dimension* 14 (3): 139–146.