Queueing Theory

Ivo Adan and Jacques Resing

Department of Mathematics and Computing Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

March 8, 2000

Contents

Chapter 1

Introduction

In general we do not like to wait. But reduction of the waiting time usually requires extra investments. To decide whether or not to invest, it is important to know the effect of the investment on the waiting time. So we need models and techniques to analyse such situations.

In this course we treat a number of elementary queueing models. Attention is paid to methods for the analysis of these models, and also to applications of queueing models. Important application areas of queueing models are production systems, transportation and stocking systems, communication systems and information processing systems. Queueing models are particularly useful for the design of these system in terms of layout, capacities and control.

In these lectures our attention is restricted to models with one queue. Situations with multiple queues are treated in the course "Networks of queues." More advanced techniques for the exact, approximative and numerical analysis of queueing models are the subject of the course "Algorithmic methods in queueing theory."

The organization is as follows. Chapter 2 first discusses a number of basic concepts and results from probability theory that we will use. The most simple interesting queueing model is treated in chapter 4, and its multi server version is treated in the next chapter. Models with more general service or interarrival time distributions are analysed in the chapters 6, 7 and 8. Some simple variations on these models are discussed in chapter 10. Chapter 9 is devoted to queueing models with priority rules. The last chapter discusses some insentive systems.

The text contains a lot of exercises and the reader is urged to try these exercises. This is really necessary to acquire skills to model and analyse new situations.

1.1 Examples

Below we briefly describe some situations in which queueing is important.

Example 1.1.1 *Supermarket.*
How long do customers have to wait at the checkouts? What happens with the waiting

time during peak-hours? Are there enough checkouts?

Example 1.1.2 *Production system.*
A machine produces different types of products.
What is the production lead time of an order? What is the reduction in the lead time when we have an extra machine? Should we assign priorities to the orders?

Example 1.1.3 *Post office.*
In a post office there are counters specialized in e.g. stamps, packages, financial transactions, etc.
Are there enough counters? Separate queues or one common queue in front of counters with the same specialization?

Example 1.1.4 *Data communication.*
In computer communication networks standard packages called cells are transmitted over links from one switch to the next. In each switch incoming cells can be buffered when the incoming demand exceeds the link capacity. Once the buffer is full incoming cells will be lost.
What is the cell delay at the switches? What is the fraction of cells that will be lost? What is a good size of the buffer?

Example 1.1.5 *Parking place.*
They are going to make a new parking place in front of a super market.
How large should it be?

Example 1.1.6 *Assembly of printed circuit boards.*
Mounting vertical components on printed circuit boards is done in an assembly center consisting of a number of parallel insertion machines. Each machine has a magazine to store components.
What is the production lead time of the printed circuit boards? How should the components necessary for the assembly of printed circuit boards be divided among the machines?

Example 1.1.7 *Call centers of an insurance company.*
Questions by phone, regarding insurance conditions, are handled by a call center. This call center has a team structure, where each team helps customers from a specific region only.
How long do customers have to wait before an operator becomes available? Is the number of incoming telephone lines enough? Are there enough operators? Pooling teams?

Example 1.1.8 *Main frame computer.*
Many cashomats are connected to a big main frame computer handling all financial transactions.
Is the capacity of the main frame computer sufficient? What happens when the use of cashomats increases?

Example 1.1.9 *Toll booths.*
Motorists have to pay toll in order to pass a bridge. Are there enough toll booths?

Example 1.1.10 *Traffic lights.*
How do we have to regulate traffic lights such that the waiting times are acceptable?

Chapter 2

Basic concepts from probability theory

This chapter is devoted to some basic concepts from probability theory.

2.1 Random variable

Random variables are denoted by capitals, X, Y, etc. The expected value or mean of X is denoted by $E(X)$ and its variance by $\sigma^2(X)$ where $\sigma(X)$ is the standard deviation of X.

An important quantity is the *coefficient of variation* of the positive random variable X defined as

$$c_X = \frac{\sigma(X)}{E(X)}.$$

The coefficient of variation is a (dimensionless) measure of the variability of the random variable X.

2.2 Generating function

Let X be a nonnegative discrete random variable with $P(X = n) = p(n)$, $n = 0, 1, 2, \ldots$. Then the generating function $P_X(z)$ of X is defined as

$$P_X(z) = E(z^X) = \sum_{n=0}^{\infty} p(n) z^n.$$

Note that $|P_X(z)| \leq 1$ for all $|z| \leq 1$. Further

$$P_X(0) = p(0), \qquad P_X(1) = 1, \qquad P'_X(1) = E(X),$$

and, more general,

$$P_X^{(k)}(1) = E(X(X-1)\cdots(X-k+1)),$$

where the superscript (k) denotes the kth derivative. For the generating function of the sum $Z = X + Y$ of two *independent* discrete random variables X and Y, it holds that

$$P_Z(z) = P_X(z) \cdot P_Y(z).$$

When Z is with probability q equal to X and with probability $1 - q$ equal to Y, then

$$P_Z(z) = qP_X(z) + (1 - q)P_Y(z).$$

2.3 Laplace-Stieltjes transform

The Laplace-Stieltjes transform $\widetilde{X}(s)$ of a nonnegative random variable X with distribution function $F(\cdot)$, is defined as

$$\widetilde{X}(s) = E(e^{-sX}) = \int_{x=0}^{\infty} e^{-sx} dF(x), \qquad s \geq 0.$$

When the random variable X has a density $f(\cdot)$, then the transform simplifies to

$$\widetilde{X}(s) = \int_{x=0}^{\infty} e^{-sx} f(x) dx, \qquad s \geq 0.$$

Note that $|\widetilde{X}(s)| \leq 1$ for all $s \geq 0$. Further

$$\widetilde{X}(0) = 1, \qquad \widetilde{X}'(0) = -E(X), \qquad \widetilde{X}^{(k)}(0) = (-1)^k E(X^k).$$

For the transform of the sum $Z = X + Y$ of two *independent* random variables X and Y, it holds that

$$\widetilde{Z}(s) = \widetilde{X}(s) \cdot \widetilde{Y}(s).$$

When Z is with probability q equal to X and with probability $1 - q$ equal to Y, then

$$\widetilde{Z}(s) = q\widetilde{X}(s) + (1 - q)\widetilde{Y}(s).$$

2.4 Useful probability distributions

This section discusses a number of important distributions which have been found useful for describing random variables in many applications.

2.4.1 Geometric distribution

A geometric random variable X with parameter p has probability distribution

$$P(X = n) = (1 - p)p^n, \qquad n = 0, 1, 2, \ldots$$

For this distribution we have

$$P_X(z) = \frac{1 - p}{1 - pz}, \qquad E(X) = \frac{p}{1 - p}, \qquad \sigma^2(X) = \frac{p}{(1 - p)^2}, \qquad c_X^2 = \frac{1}{p}.$$

2.4.2 Poisson distribution

A Poisson random variable X with parameter μ has probability distribution

$$P(X = n) = \frac{\mu^n}{n!}e^{-\mu}, \qquad n = 0, 1, 2, \ldots$$

For the Poisson distribution it holds that

$$P_X(z) = e^{-\mu(1-z)}, \qquad E(X) = \sigma^2(X) = \mu, \qquad c_X^2 = \frac{1}{\mu}.$$

2.4.3 Exponential distribution

The density of an exponential distribution with parameter μ is given by

$$f(t) = \mu e^{-\mu t}, \qquad t > 0.$$

The distribution function equals

$$F(t) = 1 - e^{-\mu t}, \qquad t \geq 0.$$

For this distribution we have

$$\widetilde{X}(s) = \frac{\mu}{\mu + s}, \qquad E(X) = \frac{1}{\mu}, \qquad \sigma^2(X) = \frac{1}{\mu^2}, \qquad c_X = 1.$$

An important property of an exponential random variable X with parameter μ is the *memoryless property*. This property states that for all $x \geq 0$ and $t \geq 0$,

$$P(X > x + t | X > t) = P(X > x) = e^{-\mu x}.$$

So the remaining lifetime of X, given that X is still alive at time t, is again exponentially distributed with the same mean $1/\mu$. We often use the memoryless property in the form

$$P(X < t + \Delta t | X > t) = 1 - e^{\mu \Delta t} = \mu \Delta t + o(\Delta t), \qquad (\Delta t \to 0), \tag{2.1}$$

where $o(\Delta t)$, $(\Delta t \to 0)$, is a shorthand notation for a function, $g(\Delta t)$ say, for which $g(\Delta t)/\Delta t$ tends to 0 when $\Delta t \to 0$ (see e.g. [4]).

If $X_1, \ldots, X_n$ are independent exponential random variables with parameters $\mu_1, \ldots, \mu_n$ respectively, then $\min(X_1, \ldots, X_n)$ is again an exponential random variable with parameter $\mu_1 + \cdots + \mu_n$ and the probability that X_i is the smallest one is given by $\mu_i/(\mu_1 + \cdots + \mu_n)$, $i = 1, \ldots, n$. (see exercise 2.7.1).

2.4.4 Erlang distribution

A random variable X has an *Erlang-k* ($k = 1, 2, \ldots$) distribution with mean k/μ if X is the sum of k independent random variables $X_1, \ldots, X_k$ having a common exponential distribution with mean $1/\mu$. The common notation is $E_k(\mu)$ or briefly E_k. The density of an $E_k(\mu)$ distribution is given by

$$f(t) = \mu \frac{(\mu t)^{k-1}}{(k-1)!} e^{-\mu t}, \qquad t > 0.$$

The distribution function equals

$$F(t) = 1 - \sum_{j=0}^{k-1} \frac{(\mu t)^j}{j!} e^{-\mu t}, \qquad t \geq 0.$$

The parameter μ is called the scale parameter, k is the shape parameter. A phase diagram of the E_k distribution is shown in figure 2.1.

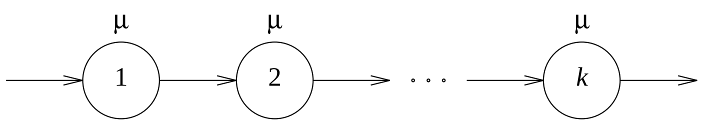

Figure 2.1: Phase diagram for the Erlang-k distribution with scale parameter μ

In figure 2.2 we display the density of the Erlang-k distribution with mean 1 (so $\mu = k$) for various values of k.

The mean, variance and squared coefficient of variation are equal to

$$E(X) = \frac{k}{\mu}, \qquad \sigma^2(X) = \frac{k}{\mu^2}, \qquad c_X^2 = \frac{1}{k}.$$

The Laplace-Stieltjes transform is given by

$$\widetilde{X}(s) = \left(\frac{\mu}{\mu + s} \right)^k.$$

A convenient distribution arises when we mix an E_{k-1} and E_k distribution with the same scale parameters. The notation used is $E_{k-1,k}$. A random variable X has an $E_{k-1,k}(\mu)$ distribution, if X is with probability p (resp. $1-p$) the sum of $k-1$ (resp. k) independent exponentials with common mean $1/\mu$. The density of this distribution has the form

$$f(t) = p\mu \frac{(\mu t)^{k-2}}{(k-2)!} e^{-\mu t} + (1-p)\mu \frac{(\mu t)^{k-1}}{(k-1)!} e^{-\mu t}, \qquad t > 0,$$

where $0 \leq p \leq 1$. As p runs from 0 to 1, the squared coefficient of variation of the mixed Erlang distribution varies from $1/(k-1)$ to $1/k$. It will appear (later on) that this distribution is useful for fitting a distribution if only the first two moments of a random variable are known.

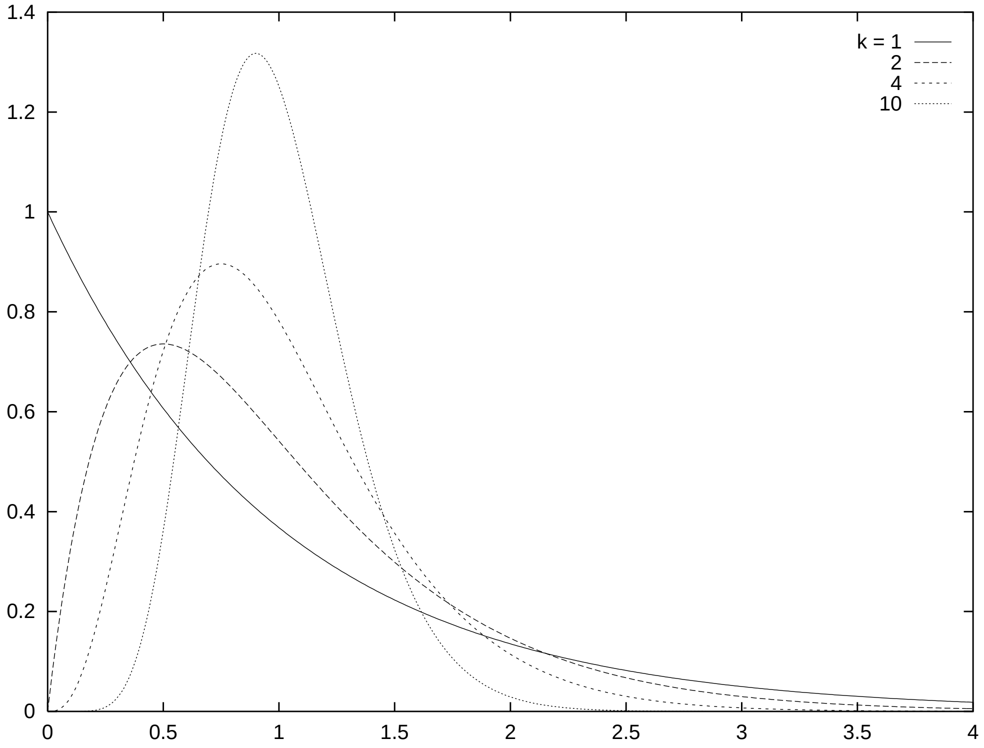

Figure 2.2: The density of the Erlang-k distribution with mean 1 for various values of k

2.4.5 Hyperexponential distribution

A random variable X is hyperexponentially distributed if X is with probability p_i, $i = 1, \ldots, k$ an exponential random variable X_i with mean $1/\mu_i$. For this random variable we use the notation $H_k(p_1, \ldots, p_k; \mu_1, \ldots, \mu_k)$, or simply H_k. The density is given by

$$f(t) = \sum_{i=1}^{k} p_i \mu_i e^{-\mu_i t}, \qquad t > 0,$$

and the mean is equal to

$$E(X) = \sum_{i=1}^{k} \frac{p_i}{\mu_i}.$$

The Laplace-Stieltjes transform satisfies

$$\widetilde{X}(s) = \sum_{i=1}^{k} \frac{p_i \mu_i}{\mu_i + s}.$$

The coefficient of variation c_X of this distribution is always greater than or equal to 1 (see exercise 2.7.3). A phase diagram of the H_k distribution is shown in figure 2.3.

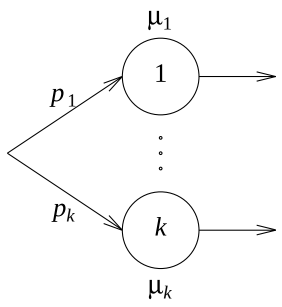

Figure 2.3: Phase diagram for the hyperexponential distribution

2.4.6 Phase-type distribution

The preceding distributions are all special cases of the *phase-type distribution*. The notation is PH. This distribution is characterized by a Markov chain with states $1, \ldots, k$ (the so-called phases) and a transition probability matrix P which is *transient*. This means that P^n tends to zero as n tends to infinity. In words, eventually you will always leave the Markov chain. The residence time in state i is exponentially distributed with mean $1/\mu_i$, and the Markov chain is entered with probability p_i in state i, $i = 1, \ldots, k$. Then the random variable X has a phase-type distribution if X is the total residence time in the preceding Markov chain, i.e. X is the total time elapsing from start in the Markov chain till departure from the Markov chain.

We mention two important classes of phase-type distributions which are *dense in the class of all non-negative distribution functions.* This is meant in the sense that for any non-negative distribution function $F(\cdot)$ a sequence of phase-type distributions can be found which pointwise converges at the points of continuity of $F(\cdot)$. The denseness of the two classes makes them very useful as a practical modelling tool. A proof of the denseness can be found in [23, 24]. The first class is the class of *Coxian distributions*, notation C_k, and the other class consists of *mixtures of Erlang distributions with the same scale parameters.* The phase representations of these two classes are shown in the figures 2.4 and 2.5.

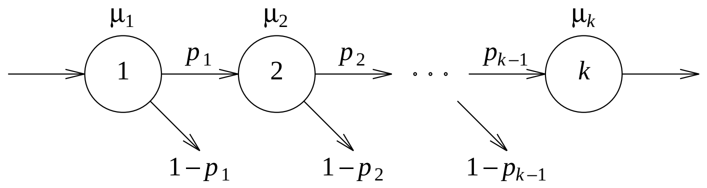

Figure 2.4: Phase diagram for the Coxian distribution

A random variable X has a Coxian distribution of order k if it has to go through up to at most k exponential phases. The mean length of phase n is μ_n, $n = 1, \ldots, k$. It starts in phase 1. After phase n it comes to an end with probability $1 - p_n$ and it enters the next phase with probability p_n. Obviously $p_k = 0$. For the Coxian-2 distribution it holds that

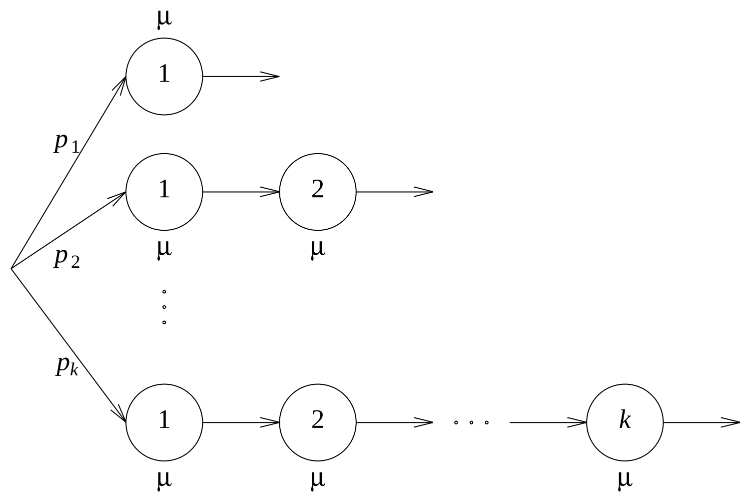

Figure 2.5: Phase diagram for the mixed Erlang distribution

the squared coefficient of variation is greater than or equal to 0.5 (see exercise 2.7.8).

A random variable X has a mixed Erlang distribution of order k if it is with probability p_n the sum of n exponentials with the same mean $1/\mu$, $n = 1, \ldots, k$.

2.5 Fitting distributions

In practice it often occurs that the only information of random variables that is available is their mean and standard deviation, or if one is lucky, some real data. To obtain an approximating distribution it is common to fit a phase-type distribution on the mean, $E(X)$, and the coefficient of variation, c_X, of a given positive random variable X, by using the following simple approach.

In case $0 < c_X < 1$ one fits an $E_{k-1,k}$ distribution (see subsection 2.4.4). More specifically, if

$$\frac{1}{k} \leq c_X^2 \leq \frac{1}{k-1},$$

for certain $k = 2, 3, \ldots$, then the approximating distribution is with probability p (resp. $1-p$) the sum of $k-1$ (resp. k) independent exponentials with common mean $1/\mu$. By choosing (see e.g. [28])

$$p = \frac{1}{1+c_X^2}[kc_X^2 - \{k(1+c_X^2) - k^2c_X^2\}^{1/2}], \qquad \mu = \frac{k-p}{E(X)},$$

the $E_{k-1,k}$ distribution matches $E(X)$ and c_X.

In case $c_X \geq 1$ one fits a $H_2(p_1, p_2; \mu_1, \mu_2)$ distribution. The hyperexponential distribution however is not uniquely determined by its first two moments. In applications, the H_2

distribution with *balanced means* is often used. This means that the normalization

$$\frac{p_1}{\mu_1} = \frac{p_2}{\mu_2}$$

is used. The parameters of the H_2 distribution with balanced means and fitting $E(X)$ and c_X (≥ 1) are given by

$$p_1 = \frac{1}{2}\left(1 + \sqrt{\frac{c_X^2 - 1}{c_X^2 + 1}}\right), \qquad p_2 = 1 - p_1,$$

$$\mu_1 = \frac{2p_1}{E(X)}, \qquad \mu_1 = \frac{2p_2}{E(X)}.$$

In case $c_X^2 \geq 0.5$ one can also use a Coxian-2 distribution for a two-moment fit. The following set is suggested by [18],

$$\mu_1 = 2E(X), \quad p_1 = 0.5/c_X^2, \quad \mu_2 = \mu_1 p_1.$$

It also possible to make a more sophisticated use of phase-type distributions by, e.g., trying to match the first three (or even more) moments of X or to approximate the shape of X (see e.g. [29, 11, 13]).

Phase-type distributions may of course also naturally arise in practical applications. For example, if the processing of a job involves performing several tasks, where each task takes an exponential amount of time, then the processing time can be described by an Erlang distribution.

2.6 Poisson process

Let $N(t)$ be the number of arrivals in $[0, t]$ for a *Poisson process* with rate λ, i.e. the time between successive arrivals is exponentially distributed with parameter λ and independent of the past. Then $N(t)$ has a *Poisson distribution* with parameter λt, so

$$P(N(t) = k) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}, \qquad k = 0, 1, 2, \ldots$$

The mean, variance and coefficient of variation of $N(t)$ are equal to (see subsection 2.4.2)

$$E(N(t) = \lambda t, \qquad \sigma^2(X) = \lambda t, \qquad c_{N(t)}^2 = \frac{1}{\lambda t}.$$

From (2.1) it is easily verified that

$$P(\text{arrival in } (t, t + \Delta t]) = \lambda \Delta t + o(\Delta t), \qquad (\Delta t \to 0).$$

Hence, for small Δt,

$$P(\text{arrival in } (t, t + \Delta t]) \approx \lambda \Delta t. \tag{2.2}$$

So in each small time interval of length Δt the occurence of an arrival is equally likely. In other words, Poisson arrivals occur completely random in time. In figure 2.6 we show a realization of a Poisson process and an arrival process with Erlang-10 interarrival times. Both processes have rate 1. The figure illustrates that Erlang arrivals are much more equally spread out over time than Poisson arrivals.

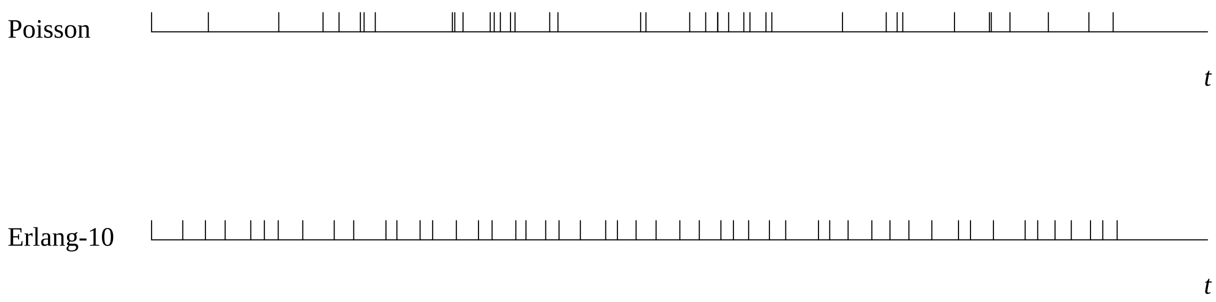

Figure 2.6: A realization of Poisson arrivals and Erlang-10 arrivals, both with rate 1

The Poisson process is an extremely useful process for modelling purposes in many practical applications, such as, e.g. to model arrival processes for queueing models or demand processes for inventory systems. It is empirically found that in many circumstances the arising stochastic processes can be well approximated by a Poisson process.

Next we mention two important properties of a Poisson process (see e.g. [20]).

(i) *Merging.*
Suppose that $N_1(t)$ and $N_2(t)$ are two independent Poisson processes with respective intensities λ_1 and λ_2. Then the sum $N_1(t) + N_2(t)$ of the two processes is again a Poisson process with rate $\lambda_1 + \lambda_2$.

(ii) *Splitting.*
Suppose that $N(t)$ is a Poisson process with rate λ and that each arrival is marked with probability p independent of all other arrivals. Let $N_1(t)$ and $N_2(t)$ denote respectively the number of marked and unmarked arrivals in $[0, t]$. Then $N_1(t)$ and $N_2(t)$ are both Poisson processes with respective intensities λp and $\lambda(1 - p)$. And these two processes are independent.

So Poisson processes remain Poisson processes under merging and splitting.

2.7 Exercises

Exercise 2.7.1
Let $X_1, \ldots, X_n$ be independent exponential random variables with mean $E(X_i) = 1/\mu_i$, $i = 1, \ldots, n$. Define

$$Y_n = \min(X_1, \ldots, X_n), \qquad Z_n = \max(X_1, \ldots, X_n).$$

(i) Determine the distributions of Y_n and Z_n.

(ii) Show that the probability that X_i is the smallest one among $X_1, \ldots, X_n$ is equal to $\mu_i/(\mu_1 + \cdots + \mu_n)$, $i = 1, \ldots, n$.

Exercise 2.7.2
Let $X_1, X_2, \ldots$ be independent exponential random variables with mean $1/\mu$ and let N be a discrete random variable with

$$P(N = k) = (1-p)p^{k-1}, \qquad k = 1, 2, \ldots,$$

where $0 \leq p < 1$ (i.e. N is a shifted geometric random variable). Show that S defined as

$$S = \sum_{n=1}^{N} X_n$$

is again exponentially distributed with parameter $(1-p)\mu$.

Exercise 2.7.3
Show that the coefficient of variation of a hyperexponential distribution is greater than or equal to 1.

Exercise 2.7.4 *(Poisson process)*
Suppose that arrivals occur at $T_1, T_2, \ldots$. The interarrival times $A_n = T_n - T_{n-1}$ are independent and have common exponential distribution with mean $1/\lambda$, where $T_0 = 0$ by convention. Let $N(t)$ denote the number of arrivals is $[0, t]$ and define for $n = 0, 1, 2, \ldots$

$$p_n(t) = P(N(t) = n), \qquad t > 0.$$

(i) Determine $p_0(t)$.

(ii) Show that for $n = 1, 2, \ldots$

$$p_n'(t) = -\lambda p_n(t) + \lambda p_{n-1}(t), \qquad t > 0,$$

with initial condition $p_n(0) = 0$.

(iii) Solve the preceding differential equations for $n = 1, 2, \ldots$

Exercise 2.7.5 *(Poisson process)*
Suppose that arrivals occur at $T_1, T_2, \ldots$. The interarrival times $A_n = T_n - T_{n-1}$ are independent and have common exponential distribution with mean $1/\lambda$, where $T_0 = 0$ by convention. Let $N(t)$ denote the number of arrivals is $[0,t]$ and define for $n = 0, 1, 2, \ldots$

$$p_n(t) = P(N(t) = n), \qquad t > 0.$$

(i) Determine $p_0(t)$.

(ii) Show that for $n = 1, 2, \ldots$

$$p_n(t) = \int_0^t p_{n-1}(t-x)\lambda e^{-\lambda x} dx, \qquad t > 0.$$

(iii) Solve the preceding integral equations for $n = 1, 2, \ldots$

Exercise 2.7.6 *(Poisson process)*
Prove the properties (i) and (ii) of Poisson processes, formulated in section 2.6.

Exercise 2.7.7 *(Fitting a distribution)*
Suppose that processing a job on a certain machine takes on the average 4 minutes with a standard deviation of 3 minutes. Show that if we model the processing time as a mixture of an Erlang-1 (exponential) distribution and an Erlang-2 distribution with density

$$f(t) = p\mu e^{-\mu t} + (1-p)\mu^2 t e^{-\mu t},$$

the parameters p and μ can be chosen in such a way that this distribution matches the mean and standard deviation of the processing times on the machine.

Exercise 2.7.8
Consider a random variable X with a Coxian-2 distribution with parameters μ_1 and μ_2 and branching probability p_1.

(i) Show that $c_X^2 \geq 0.5$.

(ii) Show that if $\mu_1 < \mu_2$, then this Coxian-2 distribution is identical to the Coxian-2 distribution with parameters $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{p}_1$ where $\hat{\mu}_1 = \mu_2$, $\hat{\mu}_2 = \mu_1$ and $\hat{p}_1 = 1 - (1-p_1)\mu_1/\mu_2$.

Part (ii) implies that for any Coxian-2 distribution we may assume without loss of generality that $\mu_1 \geq \mu_2$.

Exercise 2.7.9
Let X and Y be exponentials with parameters μ and λ, respectively. Suppose that $\lambda < \mu$. Let Z be equal to X with probability λ/μ and equal to $X+Y$ with probability $1-\lambda/\mu$. Show that Z is an exponential with parameter λ.

Exercise 2.7.10
Consider a H_2 distribution with parameters $\mu_1 > \mu_2$ and branching probabilities q_1 and q_2, respectively. Show that the C_2 distribution with parameters μ_1 and μ_2 and branching probability p_1 given by

$$p_1 = 1 - (q_1\mu_1 + q_2\mu_2)/\mu_1 \,,$$

is equivalent to the H_2 distribution.

Exercise 2.7.11 *(Poisson distribution)*
Let $X_1, \ldots, X_n$ be independent Poisson random variables with means $\mu_1, \ldots, \mu_n$, respectively. Show that the sum $X_1 + \cdots + X_n$ is Poisson distributed with mean $\mu_1 + \cdots + \mu_n$.

Chapter 3

Queueing models and some fundamental relations

In this chapter we describe the basic queueing model and we discuss some important fundamental relations for this model. These results can be found in every standard textbook on this topic, see e.g. [14, 20, 28].

3.1 Queueing models and Kendall's notation

The basic queueing model is shown in figure 3.1. It can be used to model, e.g., machines or operators processing orders or communication equipment processing information.

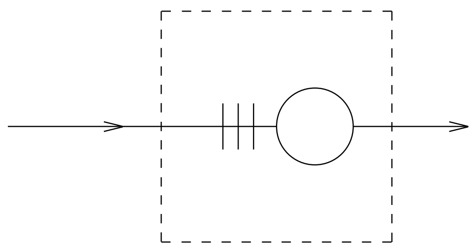

Figure 3.1: Basic queueing model

Among others, a queueing model is characterized by:

- The arrival process of customers.
  Usually we assume that the interarrival times are independent and have a common distribution. In many practical situations customers arrive according to a Poisson stream (i.e. exponential interarrival times). Customers may arrive one by one, or in batches. An example of batch arrivals is the customs office at the border where travel documents of bus passengers have to be checked.

- The behaviour of customers.
  Customers may be patient and willing to wait (for a long time). Or customers may be impatient and leave after a while. For example, in call centers, customers will hang up when they have to wait too long before an operator is available, and they possibly try again after a while.

- The service times.
  Usually we assume that the service times are independent and identically distributed, and that they are independent of the interarrival times. For example, the service times can be deterministic or exponentially distributed. It can also occur that service times are dependent of the queue length. For example, the processing rates of the machines in a production system can be increased once the number of jobs waiting to be processed becomes too large.

- The service discipline.
  Customers can be served one by one or in batches. We have many possibilities for the order in which they enter service. We mention:

  - first come first served, i.e. in order of arrival;
  - random order;
  - last come first served (e.g. in a computer stack or a shunt buffer in a production line);
  - priorities (e.g. rush orders first, shortest processing time first);
  - processor sharing (in computers that equally divide their processing power over all jobs in the system).

- The service capacity.
  There may be a single server or a group of servers helping the customers.

- The waiting room.
  There can be limitations with respect to the number of customers in the system. For example, in a data communication network, only finitely many cells can be buffered in a switch. The determination of good buffer sizes is an important issue in the design of these networks.

Kendall introduced a shorthand notation to characterize a range of these queueing models. It is a three-part code $a/b/c$. The first letter specifies the interarrival time distribution and the second one the service time distribution. For example, for a general distribution the letter G is used, M for the exponential distribution (M stands for Memoryless) and D for deterministic times. The third and last letter specifies the number of servers. Some examples are $M/M/1$, $M/M/c$, $M/G/1$, $G/M/1$ and $M/D/1$. The notation can be extended with an extra letter to cover other queueing models. For example, a system with exponential interarrival and service times, one server and having waiting room only for N customers (including the one in service) is abbreviated by the four letter code $M/M/1/N$.

In the basic model, customers arrive one by one and they are always allowed to enter the system, there is always room, there are no priority rules and customers are served in order of arrival. It will be explicitly indicated (e.g. by additional letters) when one of these assumptions does not hold.

3.2 Occupation rate

In a single-server system $G/G/1$ with arrival rate λ and mean service time $E(B)$ the amount of work arriving per unit time equals $\lambda E(B)$. The server can handle 1 unit work per unit time. To avoid that the queue eventually grows to infinity, we have to require that $\lambda E(B) < 1$. Without going into details, we note that the mean queue length also explodes when $\lambda E(B) = 1$, except in the $D/D/1$ system, i.e., the system with no randomness at all.

It is common to use the notation

$$\rho = \lambda E(B).$$

If $\rho < 1$, then ρ is called the *occupation rate* or *server utilization*, because it is the fraction of time the server is working.

In a multi-server system $G/G/c$ we have to require that $\lambda E(B) < c$. Here the occupation rate per server is $\rho = \lambda E(B)/c$.

3.3 Performance measures

Relevant performance measures in the analysis of queueing models are:

- The distribution of the waiting time and the sojourn time of a customer. The sojourn time is the waiting time plus the service time.
- The distribution of the number of customers in the system (including or excluding the one or those in service).
- The distribution of the amount of work in the system. That is the sum of service times of the waiting customers and the residual service time of the customer in service.
- The distribution of the *busy period* of the server. This is a period of time during which the server is working continuously.

In particular, we are interested in mean performance measures, such as the mean waiting time and the mean sojourn time.

Now consider the $G/G/c$ queue. Let the random variable $L(t)$ denote the number of customers in the system at time t, and let S_n denote the sojourn time of the nth customer in the system. Under the assumption that the occupation rate per server is less than one, it can be shown that these random variables have a limiting distribution as $t \to \infty$ and $n \to \infty$. These distributions are independent of the initial condition of the system.

Let the random variables L and S have the limiting distributions of $L(t)$ and S_n, respectively. So

$$p_k = P(L = k) = \lim_{t\to\infty} P(L(t) = k), \qquad F_S(x) = P(S \leq x) = \lim_{n\to\infty} P(S_n \leq x).$$

The probability p_k can be interpreted as the fraction of time that k customers are in the system, and $F_S(x)$ gives the probability that the sojourn time of an arbitrary customer entering the system is not greater than x units of time. It further holds with probability 1 that

$$\lim_{t\to\infty} \frac{1}{t} \int_{x=0}^{t} L(x)dx = E(L), \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} S_k = E(S).$$

So the long-run average number of customers in the system and the long-run average sojourn time are equal to $E(L)$ and $E(S)$, respectively. A very useful result for queueing systems relating $E(L)$ and $E(S)$ is presented in the following section.

3.4 Little's law

Little's law gives a very important relation between $E(L)$, the mean number of customers in the system, $E(S)$, the mean sojourn time and λ, the average number of customers entering the system per unit time. Little's law states that

$$E(L) = \lambda E(S). \tag{3.1}$$

Here it is assumed that the capacity of the system is sufficient to deal with the customers (i.e. the number of customers in the system does not grow to infinity).

Intuitively, this result can be understood as follows. Suppose that all customers pay 1 dollar per unit time while in the system. This money can be earned in two ways. The first possibility is to let pay all customers "continuously" in time. Then the average reward earned by the system equals $E(L)$ dollar per unit time. The second possibility is to let customers pay 1 dollar per unit time for their residence in the system when they leave. In equilibrium, the average number of customers leaving the system per unit time is equal to the average number of customers entering the system. So the system earns an average reward of $\lambda E(S)$ dollar per unit time. Obviously, the system earns the same in both cases. For a rigorous proof, see [17, 25].

To demonstrate the use of Little's law we consider the basic queueing model in figure 3.1 with one server. For this model we can derive relations between several performance measures by applying Little's law to suitably defined (sub)systems. Application of Little's law to the system consisting of queue plus server yields relation (3.1). Applying Little's law to the queue (excluding the server) yields a relation between the queue length L^q and the waiting time W, namely

$$E(L^q) = \lambda E(W).$$

Finally, when we apply Little's law to the server only, we obtain (cf. section 3.2)

$$\rho = \lambda E(B),$$

where ρ is the mean number of customers at the server (which is the same as the fraction of time the server is working) and $E(B)$ the mean service time.

3.5 PASTA property

For queueing systems with Poisson arrivals, so for $M/\cdot/\cdot$ systems, the very special property holds that arriving customers find on average the same situation in the queueing system as an outside observer looking at the system at an arbitrary point in time. More precisely, the fraction of customers finding on arrival the system in some state A is exactly the same as the fraction of time the system is in state A. This property is only true for Poisson arrivals.

In general this property is not true. For instance, in a $D/D/1$ system which is empty at time 0, and with arrivals at $1, 3, 5, \ldots$ and service times 1, every arriving customer finds an empty system, whereas the fraction of time the system is empty is $1/2$.

This property of Poisson arrivals is called PASTA property, which is the acrynom for Poisson Arrivals See Time Averages. Intuitively, this property can be explained by the fact that Poisson arrivals occur completely random in time (see (2.2)). A rigorous proof of the PASTA property can be found in [31, 32].

In the following chapters we will show that in many queueing models it is possible to determine mean performance measures, such as $E(S)$ and $E(L)$, directly (i.e. not from the distribution of these measures) by using the PASTA property and Little's law. This powerful approach is called the *mean value approach*.

3.6 Exercises

Exercise 3.6.1
In a gas station there is one gas pump. Cars arrive at the gas station according to a Poisson proces. The arrival rate is 20 cars per hour. An arriving car finding n cars at the station immediately leaves with probability $q_n = n/4$, and joins the queue with probability $1 - q_n$, $n = 0, 1, 2, 3, 4$. Cars are served in order of arrival. The service time (i.e. the time needed for pumping and paying) is exponential. The mean service time is 3 minutes.

(i) Determine the stationary distribution of the number of cars at the gas station.

(ii) Determine the mean number of cars at the gas station.

(iii) Determine the mean sojourn time (waiting time plus service time) of cars deciding to take gas at the station.

(iv) Determine the mean sojourn time and the mean waiting time of all cars arriving at the gas station.

Chapter 4

$M/M/1$ queue

In this chapter we will analyze the model with exponential interarrival times with mean $1/\lambda$, exponential service times with mean $1/\mu$ and a single server. Customers are served in order of arrival. We require that

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

since, otherwise, the queue length will explode (see section 3.2). The quantity ρ is the fraction of time the server is working. In the following section we will first study the time-dependent behaviour of this system. After that, we consider the limiting behaviour.

4.1 Time-dependent behaviour

The exponential distribution allows for a very simple description of the state of the system at time t, namely the number of customers in the system (i.e. the customers waiting in the queue and the one being served). Neither we do have to remember when the last customer arrived nor we have to register when the last customer entered service. Since the exponential distribution is memoryless (see 2.1), this information does not yield a better prediction of the future.

Let $p_n(t)$ denote the probability that at time t there are n customers in the system, $n = 0, 1, \ldots$ Based on property (2.1) we get, for $\Delta t \to 0$,

$$\begin{aligned} p_0(t+\Delta t) &= (1-\lambda\Delta t)p_0(t) + \mu\Delta t p_1(t) + o(\Delta t), \\ p_n(t+\Delta t) &= \lambda\Delta t)p_{n-1}(t) + (1-(\lambda+\mu)\Delta t)p_n(t) + \mu\Delta t p_{n+1}(t) + o(\Delta t), \\ & \qquad n = 1, 2, \ldots \end{aligned}$$

Hence, by letting $\Delta t \to 0$, we obtain the following infinite set of differential equations for the probabilities $p_n(t)$.

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu p_1(t), \\ p_n'(t) &= \lambda p_{n-1}(t) - (\lambda+\mu)p_n(t) + \mu p_{n+1}(t), \qquad n = 1, 2, \ldots \end{aligned} \tag{4.1}$$

It is difficult to solve these differential equations. An explicit solution for the probabilities $p_n(t)$ can be found in [14] (see p. 77). The expression presented there is an infinite sum of modified Bessel functions. So already one of the simplest interesting queueing models leads to a difficult expression for the time-dependent behavior of its state probabilities. For more general systems we can only expect more complexity. Therefore, in the remainder we will focus on the *limiting or equilibrium behavior* of this system, which appears to be much easier to analyse.

4.2 Limiting behavior

One may show that as $t \to \infty$, then $p_n'(t) \to 0$ and $p_n(t) \to p_n$ (see e.g. [8]). Hence, from (4.1) it follows that the limiting or equilibrium probabilities p_n satisfy the equations

$$\begin{aligned} 0 &= -\lambda p_0 + \mu p_1, && (4.2) \\ 0 &= \lambda p_{n-1} - (\lambda + \mu) p_n + \mu p_{n+1}, \qquad n = 1, 2, \ldots && (4.3) \end{aligned}$$

Clearly, the probabilities p_n also satisfy

$$\sum_{n=0}^{\infty} p_n = 1, \tag{4.4}$$

which is called the normalization equation. It is also possible to derive the equations (4.2) and (4.3) directly from a *flow diagram*, as shown in figure 4.1.

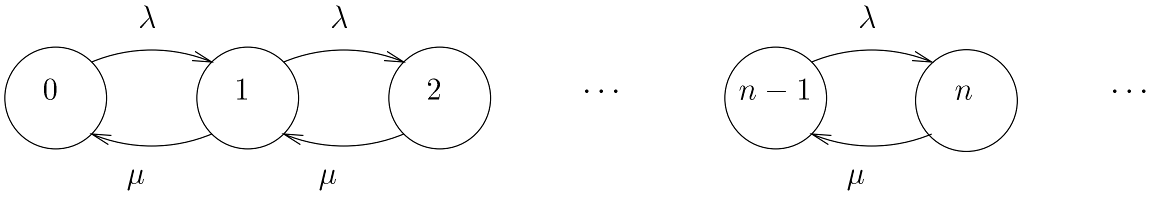

Figure 4.1: Flow diagram for the $M/M/1$ model

The arrows indicate possible transitions. The rate at which a transition occurs is λ for a transition from n to $n+1$ (an arrival) and μ for a transition from $n+1$ to n (a departure). The number of transitions per unit time from n to $n+1$, which is also called the *flow* from n to $n+1$, is equal to p_n, the fraction of time the system is in state n, times λ, the rate at arrivals occur while the system is in state n. The equilibrium equations (4.2) and (4.3) follow by equating the flow out of state n and the flow into state n.

For this simple model there are many ways to determine the solution of the equations (4.2)–(4.4). Below we discuss several approaches.

4.2.1 Direct approach

The equations (4.3) are a second order recurrence relation with constant coefficients. Its general solution is of the form

$$p_n = c_1 x_1^n + c_2 x_2^n, \qquad n = 0, 1, 2, \ldots \tag{4.5}$$

where x_1 and x_2 are roots of the quadratic equation

$$\lambda - (\lambda + \mu)x + \mu x^2 = 0.$$

This equation has two zeros, namely $x = 1$ and $x = \lambda/\mu = \rho$. So all solutions to (4.3) are of the form

$$p_n = c_1 + c_2 \rho^n, \qquad n = 0, 1, 2, \ldots$$

Equation (4.4), stating that the sum of all probabilities is equal to 1, of course directly implies that c_1 must be equal to 0. That c_1 must be equal to 0 also follows from (4.2) by substituting the solution (4.5) into (4.2).

The coefficient c_2 finally follows from the normalization equation (4.4), yielding that $c_2 = 1 - \rho$. So we can conclude that

$$p_n = (1 - \rho)\rho^n, \qquad n = 0, 1, 2, \ldots \tag{4.6}$$

Apparently, the equilibrium distribution depends upon λ and μ only through their ratio ρ.

4.2.2 Recursion

One can use (4.2) to express p_1 in p_0 yielding

$$p_1 = \rho p_0.$$

Substitution of this relation into (4.3) for $n = 1$ gives

$$p_2 = \rho^2 p_0.$$

By substituting the relations above into (4.3) for $n = 2$ we obtain p_3, and so on. Hence we can recursively express all probabilities in terms of p_0, yielding

$$p_n = \rho^n p_0, \qquad n = 0, 1, 2, \ldots.$$

The probability p_0 finally follows from the normalization equation (4.4).

4.2.3 Generating function approach

The probability generating function of the random variable L, the number of customers in the system, is given by

$$P_L(z) = \sum_{n=0}^{\infty} p_n z^n, \tag{4.7}$$

which is properly defined for z with $|z| \leq 1$. By multiplying the nth equilibrium equation with z^n and then summing the equations over all n, the equilibrium equations for p_n can be transformed into the following single equation for $P_L(z)$,

$$0 = \mu p_0(1 - z^{-1}) + (\lambda z + \mu z^{-1} - (\lambda + \mu))P_L(z).$$

The solution of this equation is

$$P_L(z) = \frac{p_0}{1 - \rho z} = \frac{1 - \rho}{1 - \rho z} = \sum_{n=0}^{\infty} (1 - \rho)\rho^n z^n, \tag{4.8}$$

where we used that $P(1) = 1$ to determine $p_0 = 1 - \rho$ (cf. section 3.2). Hence, by equating the coefficients of z^n in (4.7) and (4.8) we retrieve the solution (4.6).

4.2.4 Global balance principle

The global balance principle states that for *each set of states* A, the flow out of set A is equal to the flow into that set. In fact, the equilibrium equations (4.2)–(4.3) follow by applying this principle to a single state. But if we apply the balance principle to the set $A = \{0, 1, \ldots, n-1\}$ we get the very simple relation

$$\lambda p_{n-1} = \mu p_n, \qquad n = 1, 2, \ldots$$

Repeated application of this relation yields

$$p_n = \rho^n p_0, \qquad n = 0, 1, 2, \ldots$$

so that, after normalization, the solution (4.6) follows.

4.3 Mean performance measures

From the equilibrium probabilities we can derive expressions for the mean number of customers in the system and the mean time spent in the system. For the first one we get

$$E(L) = \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1 - \rho},$$

and by applying Little's law,

$$E(S) = \frac{1/\mu}{1 - \rho}. \tag{4.9}$$

If we look at the expressions for $E(L)$ and $E(S)$ we see that both quantities grow to infinity as ρ approaches unity. The dramatic behavior is caused by the variation in the arrival and service process. This type of behavior with respect to ρ is characteristic for almost every queueing system.

In fact, $E(L)$ and $E(S)$ can also be determined directly, i.e. without knowing the probabilities p_n, by combining Little's law and the PASTA property (see section 3.5). Based on PASTA we know that the average number of customers in the system seen by an arriving customer equals $E(L)$ and each of them (also the one in service) has a (residual) service time with mean $1/\mu$. The customer further has to wait for its own service time. Hence

$$E(S) = E(L)\frac{1}{\mu} + \frac{1}{\mu}.$$

This relation is known as the *arrival relation.* Together with

$$E(L) = \lambda E(S)$$

we find expression (4.9). This approach is called the *mean value approach.*

The mean number of customers in the queue, $E(L^q)$, can be obtained from $E(L)$ by subtracting the mean number of customers in service, so

$$E(L^q) = E(L) - \rho = \frac{\rho^2}{1-\rho}.$$

The mean waiting time, $E(W)$, follows from $E(S)$ by subtracting the mean service time (or from $E(L^q)$ by applying Little's law). This yields

$$E(W) = E(S) - 1/\mu = \frac{\rho/\mu}{1-\rho}.$$

4.4 Distribution of the sojourn time and the waiting time

It is also possible to derive the distribution of the sojourn time. Denote by L^a the number of customers in the system just before the arrival of a customer and let B_k be the service time of the kth customer. Of course, the customer in service has a residual service time instead of an ordinary service time. But these are the same, since the exponential service time distribution is memoryless. So the random variables B_k are independent and exponentially distributed with mean $1/\mu$. Then we have

$$S = \sum_{k=1}^{L^a+1} B_k. \tag{4.10}$$

By conditioning on L^a and using that L^a and B_k are independent it follows that

$$P(S > t) = P(\sum_{k=1}^{L^a+1} B_k > t) = \sum_{n=0}^{\infty} P(\sum_{k=1}^{n+1} B_k > t) P(L^a = n). \tag{4.11}$$

The problem is to find the probability that an arriving customer finds n customers in the system. PASTA states that the fraction of customers finding on arrival n customers in the system is equal to the fraction of time there are n customers in the system, so

$$P(L^a = n) = p_n = (1-\rho)\rho^n. \tag{4.12}$$

Substituting (4.12) in (4.11) and using that $\sum_{k=1}^{n+1} B_k$ is Erlang-$(n+1)$ distributed, yields (cf. exercise 2.7.2)

$$\begin{aligned} P(S > t) &= \sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{(\mu t)^k}{k!} e^{-\mu t}(1-\rho)\rho^n \\ &= \sum_{k=0}^{\infty}\sum_{n=k}^{\infty} \frac{(\mu t)^k}{k!} e^{-\mu t}(1-\rho)\rho^n \\ &= \sum_{k=0}^{\infty} \frac{(\mu\rho t)^k}{k!} e^{-\mu t} \\ &= e^{-\mu(1-\rho)t}, \qquad t \geq 0. \end{aligned} \tag{4.13}$$

Hence, S is exponentially distributed with parameter $\mu(1-\rho)$. This result can also be obtained via the use of transforms. From (4.10) it follows, by conditioning on L^a, that

$$\begin{aligned} \widetilde{S}(s) &= E(e^{-sS}) \\ &= \sum_{n=0}^{\infty} P(L^a = n) E(e^{-s(B_1+\ldots+B_{n+1})}) \\ &= \sum_{n=0}^{\infty} (1-\rho)\rho^n E(e^{-sB_1}) \cdots E(e^{-sB_{n+1}}). \end{aligned}$$

Since B_k is exponentially distributed with parameter μ, we have (see subsection 2.4.3)

$$E(e^{-sB_k}) = \frac{\mu}{\mu + s},$$

so

$$\widetilde{S}(s) = \sum_{n=0}^{\infty} (1-\rho)\rho^n \left(\frac{\mu}{\mu+s}\right)^{n+1} = \frac{\mu(1-\rho)}{\mu(1-\rho)+s},$$

from which we can conclude that S is an exponential random variable with parameter $\mu(1-\rho)$.

To find the distribution of the waiting time W, note that $S = W + B$, where the random variable B is the service time. Since W and B are independent, it follows that

$$\widetilde{S}(s) = \widetilde{W}(s) \cdot \widetilde{B}(s) = \widetilde{W}(s) \cdot \frac{\mu}{\mu+s}.$$

and thus,

$$\widetilde{W}(s) = \frac{(1-\rho)(\mu+s)}{\mu(1-\rho)+s} = (1-\rho)\cdot 1 + \rho \cdot \frac{\mu(1-\rho)}{\mu(1-\rho)+s}.$$

From the transform of W we conclude (see subsection 2.3) that W is with probability $(1-\rho)$ equal to zero, and with probability ρ equal to an exponential random variable with parameter $\mu(1-\rho)$. Hence

$$P(W > t) = \rho e^{-\mu(1-\rho)t}, \qquad t \geq 0. \tag{4.14}$$

The distribution of W can, of course, also be obtained along the same lines as (4.13). Note that

$$P(W > t | W > 0) = \frac{P(W > t)}{P(W > 0)} = e^{-\mu(1-\rho)t},$$

so the *conditional waiting time* $W|W > 0$ is exponentially distributed with parameter $\mu(1-\rho)$.

In table 4.1 we list for increasing values of ρ the mean waiting time and some waiting time probabilities. From these results we see that randomness in the arrival and service process leads to (long) waiting times and the waiting times explode as the server utilization tends to one.

ρ	$E(W)$		$P(W > t)$		
		t	5	10	20
0.5	1		0.04	0.00	0.00
0.8	4		0.29	0.11	0.02
0.9	9		0.55	0.33	0.12
0.95	19		0.74	0.58	0.35

Table 4.1: Performance characteristics for the $M/M/1$ with mean service time 1

Remark 4.4.1 *(PASTA property)*
For the present model we can also derive relation (4.12) directly from the flow diagram 4.1. Namely, the average number of customers per unit time finding on arrival n customers in the system is equal to λp_n. Dividing this number by the average number of customers arriving per unit time gives the desired fraction, so

$$P(L^a = n) = \frac{\lambda p_n}{\lambda} = p_n.$$

4.5 Priorities

In this section we consider an $M/M/1$ system serving different types of customers. To keep it simple we suppose that there are two types only, type 1 and 2 say, but the analysis can easily be extended the situation with more types of customers (see also chapter 9). Type 1 and type 2 customers arrive according to independent Poisson processes with rate λ_1, and

λ_2 respectively. The service times of all customers are exponentially distributed with the same mean $1/\mu$. We assume that

$$\rho_1 + \rho_2 < 1,$$

where $\rho_i = \lambda_i/\mu$, i.e. the occupation rate due to type i customers. Type 1 customers are treated with priority over type 2 jobs. In the following subsections we will consider two priority rules, preemptive-resume priority and non-preemptive priority.

4.5.1 Preemptive-resume priority

In the preemptive resume priority rule, type 1 customers have absolute priority over type 2 jobs. Absolute priority means that when a type 2 customer is in service and a type 1 customer arrives, the type 2 service is interrupted and the server proceeds with the type 1 customer. Once there are no more type 1 customers in the system, the server resumes the service of the type 2 customer at the point where it was interrupted.

Let the random variable L_i denote the number of type i customers in the system and S_i the sojourn time of a type i customer. Below we will determine $E(L_i)$ and $E(S_i)$ for $i = 1, 2$.

For type 1 customers the type 2 customers do not exist. Hence we immediately have

$$E(S_1) = \frac{1/\mu}{1-\rho_1}, \qquad E(L_1) = \frac{\rho_1}{1-\rho_1}. \tag{4.15}$$

Since the (residual) service times of all customers are exponentially distributed with the same mean, the total number of customers in the system does not depend on the order in which the customers are served. So this number is the same as in the system where all customers are served in order of arrival. Hence,

$$E(L_1) + E(L_2) = \frac{\rho_1 + \rho_2}{1 - \rho_1 - \rho_2}, \tag{4.16}$$

and thus, inserting (4.15),

$$E(L_2) = \frac{\rho_1 + \rho_2}{1 - \rho_1 - \rho_2} - \frac{\rho_1}{1-\rho_1} = \frac{\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)},$$

and applying Little's law,

$$E(S_2) = \frac{E(L_2)}{\lambda_2} = \frac{1/\mu}{(1-\rho_1)(1-\rho_1-\rho_2)}.$$

Example 4.5.1 For $\lambda_1 = 0.2$, $\lambda_2 = 0.6$ and $\mu = 1$, we find in case all customers are treated in order of arrival,

$$E(S) = \frac{1}{1-0.8} = 5,$$

and in case type 1 customers have absolute priority over type 2 jobs,

$$E(S_1) = \frac{1}{1-0.2} = 1.25, \qquad E(S_2) = \frac{1}{(1-0.2)(1-0.8)} = 6.25.$$

4.5.2 Non-preemptive priority

We now consider the situation that type 1 customers have nearly absolute priority over type 2 jobs. The difference with the previous rule is that type 1 customers are not allowed to interrupt the service of a type 2 customers. This priority rule is therefore called *non-preemptive*.

For the mean sojourn time of type 1 customers we find

$$E(S_1) = E(L_1)\frac{1}{\mu} + \frac{1}{\mu} + \rho_2\frac{1}{\mu}.$$

The last term reflects that when an arriving type 1 customer finds a type 2 customer in service, he has to wait until the service of this type 2 customer has been completed. According to PASTA the probability that he finds a type 2 customer in service is equal to the fraction of time the server spends on type 2 customers, which is ρ_2. Together with Little's law,

$$E(L_1) = \lambda_1 E(S_1),$$

we obtain

$$E(S_1) = \frac{(1+\rho_2)/\mu}{1-\rho_1}, \qquad E(L_1) = \frac{(1+\rho_2)\rho_1}{1-\rho_1}.$$

For type 2 customers it follows from (4.16) that

$$E(L_2) = \frac{(1-\rho_1(1-\rho_1-\rho_2))\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)},$$

and applying Little's law,

$$E(S_2) = \frac{(1-\rho_1(1-\rho_1-\rho_2))/\mu}{(1-\rho_1)(1-\rho_1-\rho_2)}.$$

Example 4.5.2 For $\lambda_1 = 0.2$, $\lambda_2 = 0.6$ and $\mu = 1$, we get

$$E(S_1) = \frac{1+0.6}{1-0.2} = 2, \qquad E(S_2) = \frac{1-0.2(1-0.8)}{(1-0.2)(1-0.8)} = 6.$$

4.6 Busy period

In a servers life we can distinguish *cycles*. A cycle is the time that elapses between two consecutive arrivals finding an empty system. Clearly, a cycle starts with a *busy period* BP during which the server is helping customers, followed by an *idle period* IP during which the system is empty.

Due to the memoryless property of the exponential distribution (see subsection 2.4.3), an idle period IP is exponentially distributed with mean $1/\lambda$. In the following subsections we determine the mean and the distribution of a busy period BP.

4.6.1 Mean busy period

It is clear that the mean busy period divided by the mean cycle length is equal to the fraction of time the server is working, so

$$\frac{E(BP)}{E(BP)+E(IP)} = \frac{E(BP)}{E(BP)+1/\lambda} = \rho.$$

Hence,

$$E(BP) = \frac{1/\mu}{1-\rho}.$$

4.6.2 Distribution of the busy period

Let the random variable C_n be the time till the system is empty again if there are now n customers present in the system. Clearly, C_1 is the length of a busy period, since a busy period starts when the first customer after an idle period arrives and it ends when the system is empty again. The random variables C_n satisfy the following recursion relation. Suppose there are $n(>0)$ customers in the system. Then the next event occurs after an exponential time with parameter $\lambda+\mu$: with probability $\lambda/(\lambda+\mu)$ a new customer arrives, and with probability $\mu/(\lambda+\mu)$ service is completed and a customer leaves the system. Hence, for $n = 1, 2, \ldots$,

$$C_n = X + \begin{cases} C_{n+1} & \text{with probability } \lambda/(\lambda+\mu), \\ C_{n-1} & \text{with probability } \mu/(\lambda+\mu), \end{cases} \tag{4.17}$$

where X is an exponential random variable with parameter $\lambda+\mu$. From this relation we get for the Laplace-Stieltjes transform $\widetilde{C}_n(s)$ of C_n that

$$\widetilde{C}_n(s) = \frac{\lambda+\mu}{\lambda+\mu+s}\left(\widetilde{C}_{n+1}(s)\frac{\lambda}{\lambda+\mu} + \widetilde{C}_{n-1}(s)\frac{\mu}{\lambda+\mu}\right),$$

and thus, after rewriting,

$$(\lambda+\mu+s)\widetilde{C}_n(s) = \lambda\widetilde{C}_{n+1}(s) + \mu\widetilde{C}_{n-1}(s), \qquad n = 1, 2, \ldots$$

For *fixed* s this equation is a very similar to (4.3). Its general solution is

$$\widetilde{C}_n(s) = c_1 x_1^n(s) + c_2 x_2^n(s), \qquad n = 0, 1, 2, \ldots$$

where $x_1(s)$ and $x_2(s)$ are the roots of the quadratic equation

$$(\lambda+\mu+s)x = \lambda x^2 + \mu,$$

satisfying $0 < x_1(s) \leq 1 < x_2(s)$. Since $0 \leq \widetilde{C}_n(s) \leq 1$ it follows that $c_2 = 0$. The coefficient c_1 follows from the fact that $C_0 = 0$ and hence $\widetilde{C}_0(s) = 1$, yielding $c_1 = 1$. Hence we obtain

$$\widetilde{C}_n(s) = x_1^n(s),$$

and in particular, for the Laplace-Stieltjes transform $\widetilde{BP}(s)$ of the busy period BP, we find

$$\widetilde{BP}(s) = \widetilde{C}_1(s) = x_1(s) = \frac{1}{2\lambda}\left(\lambda + \mu + s - \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu}\right).$$

By inverting this transform (see e.g. [1]) we get for the density $f_{BP}(t)$ of BP,

$$f_{BP}(t) = \frac{1}{t\sqrt{\rho}} e^{-(\lambda+\mu)t} I_1(2t\sqrt{\lambda\mu}), \qquad t > 0,$$

where $I_1(\cdot)$ denotes the modified Bessel function of the first kind of order one, i.e.

$$I_1(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{2k+1}}{k!(k+1)!}.$$

In table 4.2 we list for some values of ρ the probability $P(BP > t)$ for a number of t values. If you think of the situation that $1/\mu$ is one hour, then 10% of the busy periods lasts longer than 2 days (16 hours) and 5% percent even longer than 1 week, when $\rho = 0.9$. Since the mean busy period is 10 hours in this case, it is not unlikely that in a month time a busy period longer than a week occurs.

ρ	$P(BP > t)$						
t	1	2	4	8	16	40	80
0.8	0.50	0.34	0.22	0.13	0.07	0.02	0.01
0.9	0.51	0.36	0.25	0.16	0.10	0.05	0.03
0.95	0.52	0.37	0.26	0.18	0.12	0.07	0.04

Table 4.2: Probabilities for the busy period duration for the $M/M/1$ with mean service time equal to 1

4.7 Exercises

Exercise 4.7.1 *(bulk arrivals)*
In a work station orders arrive according to a Poisson arrival process with arrival rate λ. An order consists of N independent jobs. The distribution of N is given by

$$P(N = k) = (1 - p)p^{k-1}$$

with $k = 1, 2, \ldots$ and $0 \leq p < 1$. Each job requires an exponentially distributed amount of processing time with mean $1/\mu$.

(i) Derive the distribution of the total processing time of an order.

(ii) Determine the distribution of the number of orders in the system.

Exercise 4.7.2 *(variable production rate)*
Consider a work station where jobs arrive according to a Poisson process with arrival rate λ. The jobs have an exponentially distributed service time with mean $1/\mu$. So the service completion rate (the rate at which jobs depart from the system) is equal to μ.
If the queue length drops below the threshold Q_L the service completion rate is lowered to μ_L. If the queue length reaches Q_H, where $Q_H \geq Q_L$, the service rate is increased to μ_H. (L stands for low, H for high.)
Determine the queue length distribution and the mean time spent in the system.

Exercise 4.7.3
A repair man fixes broken televisions. The repair time is exponentially distributed with a mean of 30 minutes. Broken televisions arrive at his repair shop according to a Poisson stream, on average 10 broken televisions per day (8 hours).

(i) What is the fraction of time that the repair man has no work to do?

(ii) How many televisions are, on average, at his repair shop?

(iii) What is the mean throughput time (waiting time plus repair time) of a television?

Exercise 4.7.4
In a gas station there is one gas pump. Cars arrive at the gas station according to a Poisson process. The arrival rate is 20 cars per hour. Cars are served in order of arrival. The service time (i.e. the time needed for pumping and paying) is exponentially distributed. The mean service time is 2 minutes.

(i) Determine the distribution, mean and variance of the number of cars at the gas station.

(ii) Determine the distribution of the sojourn time and the waiting time.

(iii) What is the fraction of cars that has to wait longer than 2 minutes?

An arriving car finding 2 cars at the station immediately leaves.

(iv) Determine the distribution, mean and variance of the number of cars at the gas station.

(v) Determine the mean sojourn time and the mean waiting time of all cars (including the ones that immediately leave the gas station).

Exercise 4.7.5
A gas station has two pumps, one for gas and the other for LPG. For each pump customers arrive according to a Poisson proces. On average 20 customers per hour for gas and 5 customers for LPG. The service times are exponential. For both pumps the mean service time is 2 minutes.

(i) Determine the distribution of the number of customers at the gas pump, and at the LPG pump.

(ii) Determine the distribution of the *total* number of customers at the gas station.

Exercise 4.7.6
Consider an $M/M/1$ queue with two types of customers. The mean service time of all customers is 5 minutes. The arrival rate of type 1 customers is 4 customers per hour and for type 2 customers it is 5 customers per hour. Type 1 customers are treated with priority over type 2 customers.

(i) Determine the mean sojourn time of type 1 and 2 customers under the preemptive-resume priority rule.

(ii) Determine the mean sojourn time of type 1 and 2 customers under the non-preemptive priority rule.

Exercise 4.7.7
Consider an $M/M/1$ queue with an arrival rate of 60 customers per hour and a mean service time of 45 seconds. A period during which there are 5 or more customers in the system is called crowded, when there are less than 5 customers it is quiet. What is the mean number of crowded periods per day (8 hours) and how long do they last on average?

Exercise 4.7.8
Consider a machine where jobs arrive according to a Poisson stream with a rate of 20 jobs per hour. The processing times are exponentially distributd with a mean of $1/\mu$ hours. The processing cost is 16μ dollar per hour, and the waiting cost is 20 dollar per order per hour.
Determine the processing speed μ minimizing the average cost per hour.

Chapter 5

$M/M/c$ queue

In this chapter we will analyze the model with exponential interarrival times with mean $1/\lambda$, exponential service times with mean $1/\mu$ and c parallel identical servers. Customers are served in order of arrival. We suppose that the occupation rate per server,

$$\rho = \frac{\lambda}{c\mu},$$

is smaller than one.

5.1 Equilibrium probabilities

The state of the system is completely characterized by the number of customers in the system. Let p_n denote the equilibrium probability that there are n customers in the system. Similar as for the $M/M/1$ we can derive the equilibrium equations for the probabilities p_n from the flow diagram shown in figure 5.1.

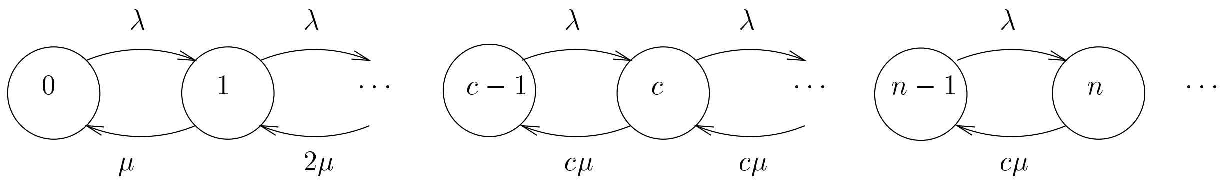

Figure 5.1: Flow diagram for the $M/M/c$ model

Instead of equating the flow into and out of a single state n, we get simpler equations by equating the flow out of and into the set of states $\{0, 1, \ldots, n-1\}$. This amounts to equating the flow between the two neighboring states $n-1$ and n yielding

$$\lambda p_{n-1} = \min(n, c)\mu p_n, \qquad n = 1, 2, \ldots$$

Iterating gives

$$p_n = \frac{(c\rho)^n}{n!} p_0, \qquad n = 0, \ldots, c$$

and

$$p_{c+n} = \rho^n p_c = \rho^n \frac{(c\rho)^c}{c!} p_0, \qquad n = 0, 1, 2, \ldots$$

The probability p_0 follows from normalization, yielding

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!} \cdot \frac{1}{1-\rho} \right)^{-1}.$$

An important quantity is the probability that a job has to wait. Denote this probability by Π_W. It is usually referred to as the *delay probability*. By PASTA it follows that

$$\begin{aligned}
\Pi_W &= p_c + p_{c+1} + p_{c+2} + \cdots \\
&= \frac{p_c}{1-\rho} \\
&= \frac{(c\rho)^c}{c!} \left((1-\rho) \sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!} \right)^{-1}.
\end{aligned} \tag{5.1}$$

Remark 5.1.1 *(Computation of Π_W)*
It will be clear that the computation of Π_W by using (5.1) leads to numerical problems when c is large (due to the terms $(c\rho)^c/c!$). In remark 11.3.2 we will formulate a nuermically stable procedure to compute Π_W.

5.2 Mean queue length and mean waiting time

From the equilibrium probabilities we directly obtain for the mean queue length,

$$\begin{aligned}
E(L^q) &= \sum_{n=0}^{\infty} n p_{c+n} \\
&= \frac{p_c}{1-\rho} \sum_{n=0}^{\infty} n(1-\rho)\rho^n \\
&= \Pi_W \cdot \frac{\rho}{1-\rho},
\end{aligned} \tag{5.2}$$

and then from Little's law,

$$E(W) = \Pi_W \cdot \frac{1}{1-\rho} \cdot \frac{1}{c\mu}. \tag{5.3}$$

These formulas for $E(L^q)$ and $E(W)$ can also be found by using the mean value technique. If not all servers are busy on arrival the waiting time is zero. If all servers are busy and there are zero or more customers waiting, then a new arriving customers first has to wait until the first departure and then continues to wait for as many departures as there were customers waiting upon arrival. An interdeparture time is the minimum of c exponential (residual) service times with mean $1/\mu$, and thus it is exponential with mean $1/c\mu$ (see exercise 2.7.1). So we obtain

$$E(W) = \Pi_W \frac{1}{c\mu} + E(L^q)\frac{1}{c\mu}.$$

Together with Little's law we retrieve the formulas (5.2)–(5.3). Table 5.1 lists the delay probability and the mean waiting time in an $M/M/c$ with mean service time 1 for $\rho = 0.9$.

c	Π_W	$E(W)$
1	0.90	9.00
2	0.85	4.26
5	0.76	1.53
10	0.67	0.67
20	0.55	0.28

Table 5.1: Performance characteristics for the $M/M/c$ with $\mu = 1$ and $\rho = 0.9$

We see that the delay probability slowly decreases as c increases. The mean waiting time however decreases fast (a little faster than $1/c$). One can also look somewhat differently at the performance of the system. We do not look at the occupation rate of a machine, but at the average number of idle machines. Let us call this the surplus capacity. Table 5.2 shows for fixed surplus capacity (instead of for fixed occupation rate as in the previous table) and c varying from 1 to 20 the mean waiting time and the mean number of customers in the system.

c	ρ	$E(W)$	$E(L)$
1	0.90	9.00	9
2	0.95	9.26	19
5	0.98	9.50	51
10	0.99	9.64	105
20	0.995	9.74	214

Table 5.2: Performance characteristics for fixed surplus capacity of 0.1 server

Although the mean number of customers in the system sharply increases, the mean waiting time remains nearly constant.

5.3 Distribution of the waiting time and the sojourn time

The derivation of the distribution of the waiting time is very similar to the one in section 4.4 for the $M/M/1$. By conditioning on the state seen on arrival we obtain

$$P(W > t) = \sum_{n=0}^{\infty} P(\sum_{k=1}^{n+1} D_k > t) p_{c+n},$$

where D_k is the kth interdeparture time. Clearly, the random variables D_k are independent and exponentially distributed with mean $1/c\mu$. Analogously to (4.13) we find

$$\begin{aligned}
P(W > t) &= \sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{(c\mu t)^k}{k!} e^{-c\mu t} p_c \rho^n \\
&= \sum_{k=0}^{\infty}\sum_{n=k}^{\infty} \frac{(c\mu t)^k}{k!} e^{-c\mu t} p_c \rho^n \\
&= \frac{p_c}{1-\rho} \sum_{k=0}^{\infty} \frac{(c\mu\rho t)^k}{k!} e^{-c\mu t} \\
&= \Pi_W e^{-c\mu(1-\rho)t}, \qquad t \geq 0.
\end{aligned}$$

This yields for the conditional waiting time,

$$P(W > t | W > 0) = \frac{P(W > t)}{P(W > 0)} = e^{-c\mu(1-\rho)t}, \qquad t \geq 0.$$

Hence, the conditional waiting time $W|W > 0$ is exponentially distributed with parameter $c\mu(1-\rho)$. To determine the distribution of the sojourn time we condition on the length of the service time, so

$$\begin{aligned}
P(S > t) &= P(W + B > t) \\
&= \int_{x=0}^{\infty} P(W + x > t) \mu e^{-\mu x} dx \\
&= \int_{x=0}^{t} P(W > t - x) \mu e^{-\mu x} dx + \int_{x=t}^{\infty} \mu e^{-\mu x} dx \\
&= \int_{x=0}^{t} \Pi_W e^{-c\mu(1-\rho)(t-x)} \mu e^{-\mu x} dx + e^{-\mu t} \\
&= \frac{\Pi_W}{\rho} \left(e^{-\mu(1-\rho)t} - e^{-\mu t} \right) + e^{-\mu t} \\
&= \frac{\Pi_W}{\rho} e^{-\mu(1-\rho)t} + \left(1 - \frac{\Pi_W}{\rho}\right) e^{-\mu t}.
\end{aligned}$$

It is easily seen that $\Pi_W \leq \rho$ (how?), so we can conclude that S has an hyperexponential distribution.

5.4 Exercises

Exercise 5.4.1 *(a fast and a slow machine)*
Consider two parallel machines with a common buffer where jobs arrive according to a Poisson stream with rate λ. The processing times are exponentially distributed with mean $1/\mu_1$ on machine 1 and $1/\mu_2$ on machine 2 ($\mu_1 > \mu_2$). Jobs are processed in order of arrival. A job arriving when both machines are idle is assigned to the fast machine. We assume that

$$\rho = \frac{\lambda}{\mu_1 + \mu_2}$$

is less than one.

(i) Determine the distribution of the number of jobs in the system.

(ii) Use this distribution to derive the mean number of jobs in the system.

(iii) When is it better to not use the slower machine at all?

(iv) Calculate for the following two cases the mean number of jobs in the system with and without the slow machine.

(a) $\lambda = 2$, $\mu_1 = 5$, $\mu_2 = 1$;

(b) $\lambda = 3$, $\mu_1 = 5$, $\mu_2 = 1$.

Note: In [21] it is shown that one should not remove the slow machine if $r > 0.5$ where $r = \mu_2/\mu_1$. When $0 \leq r < 0.5$ the slow machine should be removed (and the resulting system is stable) whenever $\rho \leq \rho_c$, where

$$\rho_c = \frac{2 + r^2 - \sqrt{(2 + r^2)^2 + 4(1 + r^2)(2r - 1)(1 + r)}}{2(1 + r^2)}.$$

Exercise 5.4.2
One is planning to build new telephone boxes near the railway station. The question is how many boxes are needed. Measurements showed that approximately 80 persons per hour want to make a phone call. The duration of a call is approximately exponentially distributed with mean 1 minute. How many boxes are needed such that the mean waiting time is less than 2 minutes?

Exercise 5.4.3
An insurance company has a call center handling questions of customers. Nearly 40 calls per hour have to be handled. It takes on average 3 minutes to help a customer. How many operators are needed such that only 5% of the customers has to wait longer than 2 minutes.

Exercise 5.4.4
In a dairy barn there are two water troughs (i.e. drinking places). From each trough only one cow can drink at the same time. When both troughs are occupied new arriving cows wait patiently for their turn. It takes an exponential time to drink with mean 3 minutes. Cows arrive at the water troughs according to a Poisson process with rate 20 cows per hour.

(i) Determine the probability that there are i cows at the water troughs (waiting or drinking), $i = 0, 1, 2, \ldots$

(ii) Determine the mean number of cows waiting at the troughs and the mean waiting time.

(iii) What is the fraction of cows finding both troughs occupied on arrival?

(iv) How many troughs are needed such that at most 10% of the cows find all troughs occupied on arrival?

Exercise 5.4.5
A computer consists of three processors. Their main task is to execute jobs from users. These jobs arrive according to a Poisson process with rate 15 jobs per minute. The execution time is exponentially distributed with mean 10 seconds. When a processor completes a job and there are no other jobs waiting to be executed, the processor starts to execute maintenance jobs. These jobs are always available and they take an exponential time with mean 5 seconds. But as soon as a job from a user arrives, the processor interrupts the execution of the maintenance job and starts to execute the new job. The execution of the maintenance job will be resumed later (at the point where it was interrupted).

(i) What is the mean number of processors busy with executing jobs from users?

(ii) How many maintenance jobs are completed per minute?

(iii) What is the probability that a job from a user has to wait?

(iv) Determine the mean waiting time of a job from a user.

Chapter 6

$M/E_r/1$ queue

Before analyzing the $M/G/1$ queue, we first study the $M/E_r/1$ queue. The Erlang distribution can be used to model service times with a low coefficient of variation (less than one), but it can also arise naturally. For instance, if a job has to pass, stage by stage, through a series of r independent production stages, where each stage takes a exponentially distributed time. The analysis of the $M/E_r/1$ queue is similar to that of the $M/M/1$ queue.

We consider a single-server queue. Customers arrive according to a Poisson process with rate λ and they are treated in order of arrival. The service times are Erlang-r distributed with mean r/μ (see subsection 2.4.4). For stability we require that the occupation rate

$$\rho = \lambda \cdot \frac{r}{\mu} \tag{6.1}$$

is less than one. In the following section we will explain that there are two ways in which one can describe the state of the system.

6.1 Two alternative state descriptions

The natural way to describe the state of a nonempty system is by the pair (k, l) where k denotes the number of customers in the system and l the remaining number of service phases of the customer in service. Clearly this is a two-dimensional description. An alternative way to describe the state is by counting the total number of uncompleted phases of work in the system. Clearly, there is a one to one correspondence between this number and the pair (k, l). The number of uncompleted phases of work in the system is equal to the number $(k-1)r + l$ (for the customer in service we have l phases of work instead of r). From here on we will work with the one-dimensional phase description.

6.2 Equilibrium distribution

For the one-dimensional phase description we get the flow diagram of figure 6.1.

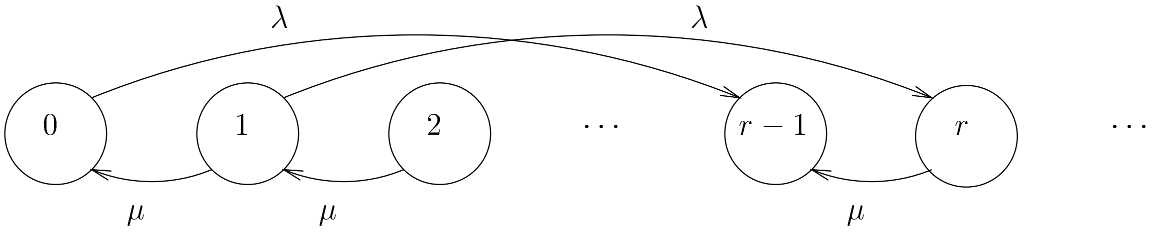

Figure 6.1: One-dimensional flow diagram for the $M/E_r/1$ model

Let p_n be the equilibrium probability of n phases work in the system. By equating the flow out of state n and the flow into state n we obtain the following set of equilibrium equations for p_n.

$$p_0\lambda = p_1\mu, \tag{6.2}$$

$$p_n(\lambda+\mu) = p_{n+1}\mu, \qquad n = 1,\dots,r-1, \tag{6.3}$$

$$p_n(\lambda+\mu) = p_{n-r}\lambda + p_{n+1}\mu, \qquad n = r, r+1, r+2, \dots \tag{6.4}$$

These equations may be solved as follows. We first look for solutions of (6.4) of the form

$$p_n = x^n, \qquad n = 0,1,2,\dots \tag{6.5}$$

and then we construct a linear combination of these solutions also satisfying the boundary equations (6.2)–(6.3) and the normalization equation

$$\sum_{n=0}^{\infty} p_n = 1.$$

An alternative solution approach is based on generating functions (see exercise 6.5.2).

Substitution of (6.5) into (6.4) and then dividing by the common power x^{n-r} yields the polynomial equation

$$(\lambda+\mu)x^r = \lambda + \mu x^{r+1}. \tag{6.6}$$

One root is $x = 1$, but this one is not useful, since we must be able to normalize the solution of the equilibrium equations. Provided condition (6.1) holds, it can be shown that equation (6.6) has exactly r distinct roots x with $|x| < 1$, say $x_1, \dots, x_r$ (see exercise 6.5.1). We now consider the linear combination

$$p_n = \sum_{k=1}^{r} c_k x_k^n, \qquad n = 0,1,2,\dots \tag{6.7}$$

For each choice of the coefficients c_k this linear combination satisfies (6.4). This freedom is used to also satisfy the equations (6.2)–(6.3) and the normalization equation. Note that, since the equilibrium equations are dependent, equation (6.2) (or one of the equations in (6.3)) may be omitted. Substitution of (6.7) into these equations yields a set of r linear equations for r unknown coefficients. It can be shown that this set of equations has a unique solution, given by

$$c_k = \frac{1-\rho}{\prod_{j\neq k}(1 - x_j/x_k)}, \qquad k = 1, \ldots, r,$$

where the subscript j runs from 1 to r. This completes the determination of the equilibrium probabilities p_n. The important conclusion is that for the $M/E_r/1$ queue the equilibrium probabilities can be expressed as a mixture of r geometric distributions.

From the distribution of the number of phases in the system we can easily find the distribution of the number of customers in the system. Let q_i be the probability of i customers in the system. Obviously, $q_0 = p_0$ and for $i \geq 1$ it follows that

$$\begin{aligned}
q_i &= \sum_{n=(i-1)r+1}^{ir} p_n \\
&= \sum_{n=(i-1)r+1}^{ir} \sum_{k=1}^{r} c_k x_k^n \\
&= \sum_{k=1}^{r} \sum_{n=(i-1)r+1}^{ir} c_k x_k^n \\
&= \sum_{k=1}^{r} c_k (x_k^{-r+1} + x_k^{-r+2} + \cdots + 1)(x_k^r)^i.
\end{aligned}$$

Hence, the queue length probabilities can also be expressed as a mixture of r geometric distributions.

We finally remark that the roots of equation (6.6) can be numerically determined very efficiently (cf. [2, 3]), and that the results in this section can be easily extended to the case that the service times are mixtures of Erlang distributions with the same scale parameters (see subsection 2.4.6).

Example 6.2.1

Consider the $M/E_2/1$ queue with $\lambda = 1$ and $\mu = 6$. The equilibrium equations are given by

$$\begin{aligned}
p_0 &= 6p_1, \\
7p_1 &= 6p_2, \\
7p_n &= p_{n-2} + 6p_{n+1}, \qquad n = 2, 3, 4, \ldots
\end{aligned}$$

Substitution of $p_n = x^n$ into the equations for $n \geq 2$ and dividing by x^{n-2} yields

$$7x^2 = 1 + 6x^3,$$

the roots of which are $x = 1$, $x = 1/2$ and $x = -1/3$. Hence, we set

$$p_n = c_1 \left(\frac{1}{2}\right)^n + c_2 \left(-\frac{1}{3}\right)^n, \qquad n = 0, 1, 2, \ldots$$

and determine c_1 and c_2 from the equilibrium equation in $n = 0$ and the normalization equation,

$$\begin{aligned}
&c_1 + c_2 = 3c_1 - 2c_2, \\
&\frac{c_1}{1-1/2} + \frac{c_2}{1+1/3} = 1.
\end{aligned}$$

The solution is $c_1 = 2/5$ and $c_2 = 4/15$ (note that the equilibrium equation in $n = 1$ is also satisfied). So we obtain

$$p_n = \frac{2}{5}\left(\frac{1}{2}\right)^n + \frac{4}{15}\left(-\frac{1}{3}\right)^n, \qquad n = 0, 1, 2, \ldots$$

For the queue length probabilities it follows that $q_0 = p_0 = 2/3$ and for $i \geq 1$,

$$q_i = p_{2i-1} + p_{2i} = \frac{6}{5}\left(\frac{1}{4}\right)^i - \frac{8}{15}\left(\frac{1}{9}\right)^i.$$

Note that the formula above is also valid for $i = 0$.

6.3 Mean waiting time

Let the random variable L^f denote the number of phases work in the system. Then by PASTA it follows that

$$E(W) = E(L^f)\frac{1}{\mu},$$

and from (6.7),

$$\begin{aligned}
E(L^f) &= \sum_{n=0}^{\infty} n p_n \\
&= \sum_{n=0}^{\infty}\sum_{k=1}^{r} c_k n x_k^n \\
&= \sum_{k=1}^{r}\sum_{n=0}^{\infty} c_k n x_k^n \\
&= \sum_{k=1}^{r} \frac{c_k x_k}{(1-x_k)^2}
\end{aligned}$$

From Little's law we can also find $E(L^q)$, the mean number of customers waiting in the queue. A much more direct way to determine $E(W)$ and $E(L^q)$ is the mean value approach.

An arriving customer has to wait for the customers in the queue and, if the server is busy, for the one in service. According to the PASTA property the mean number of customers waiting in the queue is equal to $E(L^q)$ and the probability that the server is busy on arrival is equal to ρ, i.e. the fraction of time the server is busy. Hence,

$$E(W) = E(L^q)\frac{r}{\mu} + \rho E(R), \tag{6.8}$$

where the random variable R denotes the *residual service time* of the customer in service. If the server is busy on arrival, then with probability $1/r$ he is busy with the first phase of the service time, also with probability $1/r$ he is busy with the second phase, and so on. So the mean residual service time $E(R)$ is equal to

$$\begin{aligned} E(R) &= \frac{1}{r} \cdot \frac{r}{\mu} + \frac{1}{r} \cdot \frac{r-1}{\mu} + \cdots + \frac{1}{r} \cdot \frac{1}{\mu} \\ &= \frac{r+1}{2} \cdot \frac{1}{\mu}. \end{aligned} \tag{6.9}$$

Substitution of this expression into (6.8) yields

$$E(W) = E(L^q)\frac{r}{\mu} + \rho \cdot \frac{r+1}{2} \cdot \frac{1}{\mu}.$$

Together with Little's law, stating that

$$E(L^q) = \lambda E(W)$$

we find

$$E(W) = \frac{\rho}{1-\rho} \cdot \frac{r+1}{2} \cdot \frac{1}{\mu}.$$

6.4 Distribution of the waiting time

The waiting time can be expressed as

$$W = \sum_{i=1}^{L^f} B_i,$$

where B_i is the amount of work for the kth phase. So the random variables B_i are independent and exponentially distributed with mean $1/\mu$. By conditioning on L^f and using that L^f and B_i are independent it follows, very similar to (4.13), that

$$\begin{aligned} P(W > t) &= \sum_{n=1}^{\infty} P(\sum_{i=1}^{n} B_i > t) p_n \\ &= \sum_{n=1}^{\infty} \sum_{i=0}^{n-1} \frac{(\mu t)^i}{i!} e^{-\mu t} \sum_{k=1}^{r} c_k x_k^n \\ &= \sum_{k=1}^{r} c_k \sum_{i=0}^{\infty} \sum_{n=i+1}^{\infty} \frac{(\mu t)^i}{i!} e^{-\mu t} x_k^n \\ &= \sum_{k=1}^{r} \frac{c_k x_k}{1-x_k} \sum_{i=0}^{\infty} \frac{(\mu x_k t)^i}{i!} e^{-\mu t} \\ &= \sum_{k=1}^{r} \frac{c_k}{1-x_k} x_k e^{-\mu(1-x_k)t}, \qquad t \geq 0. \end{aligned}$$

Note that this distribution is a generalization of the one for the $M/M/1$ model (namely, not one, but a mixture of exponentials).

In table 6.1 we list for varying values of ρ and r the mean waiting time and some waiting time probabilities. The squared coefficient of variation of the service time is denoted by c_B^2. We see that the variation in the service times is important to the behavior of the system. Less variation in the service times leads to smaller waiting times.

ρ	r	c_B^2	$E(W)$		$P(W > t)$		
				t	5	10	20
0.8	1	1	4		0.29	0.11	0.02
	2	0.5	3		0.21	0.05	0.00
	4	0.25	2.5		0.16	0.03	0.00
	10	0.1	2.2		0.12	0.02	0.00
0.9	1	1	9		0.55	0.33	0.12
	2	0.5	6.75		0.46	0.24	0.06
	4	0.25	5.625		0.41	0.18	0.04
	10	0.1	4.95		0.36	0.14	0.02

Table 6.1: Performance characteristics for the $M/E_r/1$ with mean service time equal to 1

6.5 Exercises

Exercise 6.5.1 *(Rouché's theorem)*
Rouché's theorem reads as follows.
Let the bounded region D have as its boundary a contour C. Let the functions $f(z)$ and $g(z)$ be analytic both in D and on C, and assume that $|f(z)| < |g(z)|$ on C. Then $f(z)+g(z)$ has in D the same number of zeros as $g(z)$, all zeros counted according to their multiplicity.

(i) Use Rouché's theorem to prove that the polynomial equation (6.6) has exactly r (possibly complex) roots x with $|x| < 1$.
(*Hint*: take $f(z) = \mu z^{r+1}$ and $g(z) = -(\lambda+\mu)z^r + \lambda$, and take as contour C the circle with center 0 and radius $1-\epsilon$ with ϵ small and positive.)

(ii) Show that all roots are simple.
(Hint: Show that there are no z for which both $f(z)+g(z)$ and $f'(z)+g'(z)$ vanish at the same time.)

Exercise 6.5.2 *(Generating function approach)*
Consider the $M/E_r/1$ queue with arrival rate λ and mean service time r/μ. Let $P(z)$ be the generating function of the probabilities p_n, so

$$P(z) = \sum_{n=0}^{\infty} p_n z^n, \qquad |z| \leq 1.$$

(i) Show, by multiplying the equilibrium equations (6.2)–(6.4) with z^n and adding over all states n, that

$$P(z)(\lambda+\mu) - p_0\mu = P(z)\lambda z^r + (P(z) - p_0)\mu z^{-1}.$$

(ii) Show that $P(z)$ is given by

$$P(z) = \frac{(1-\rho)\mu}{\mu - \lambda(z^r + z^{r-1} + \cdots + z)}. \tag{6.10}$$

(iii) Show, by partial fraction decomposition of $P(z)$, that the probabilities p_n can be written as

$$p_n = \sum_{k=1}^{r} c_k \left(\frac{1}{z_k}\right)^n, \qquad n = 0, 1, 2, \ldots$$

where $z_1, \ldots, z_r$ are the zeros of the numerator in (6.10).

Exercise 6.5.3
Orders arrive according to a Poisson process, so the interarrival times are independent and exponentially distributed. Each order has to pass, stage by stage, through a series of r independent production stages. Each stage takes a exponentially distributed time. The total production time, the workload, of the order is the sum of r independent, identically distributed random variables. So the distribution of the workload is the r-stage Erlang distribution.

(i) The state of this system (i.e., the work station together with its queue) can be characterized by the number of orders in the system and by the number of not yet completed stages for the order in the work station.
Describe the flow diagram and give the set of equations for the equilibrium state probabilities.

(ii) The state of the system at a certain time can also be described by the total number of production stages that have to be completed before all the orders in the system are ready.
Give the flow diagram and formulate the equations for the equilibrium state probabilities.

(iii) Define

$$\begin{aligned} p_n &= P(\text{total number of orders in the system} = n), \\ q_k &= P(\text{total number of production stages in the system } = k). \end{aligned}$$

Give the relation between these two probabilities.

Exercise 6.5.4
Orders arrive in so called bulks. Each bulk consists of r independent orders. The bulks themselves arrive according to a Poisson process. The sequence in which orders of one bulk are processed in the work station is unimportant. All orders of a bulk have to wait until all orders of bulks that have arrived earlier are completed. The workload of each order is exponentially distributed. The state of the system is simply characterized by the number of orders in the system.
Describe the flow diagram for this situation. Compare the result with that of the previous exercise, part (ii).

Exercise 6.5.5
Jobs arrive at a machine according to a Poisson process with a rate of 16 jobs per hour. Each job consists of 2 tasks. Each task has an exponentially distributed processing time with a mean of 1 minute. Jobs are processed in order of arrival.

(i) Determine the distribution of not yet completed tasks in the system.

(ii) Determine the distribution of the number of jobs in the system.

(iii) What is the mean number of jobs in the system?

(iv) What is the mean waiting time of a job?

Exercise 6.5.6
Consider an $M/E_2/1/2$ queue with arrival rate 1 and mean service time 2. At time $t = 0$ the system is empty. Determine the mean time till the first customer is rejected on arrival.

Exercise 6.5.7
Consider the $M/E_2/1$ queue with an arrival rate of 4 customers per hour and a mean service time of 8 minutes.

(i) Determine the distribution of the waiting time.

(i) What is the fraction of customers that has to wait longer than 5 minutes?

Exercise 6.5.8
Customers arrive in groups at a single-server queue. These groups arrive according to a Poisson process with a rate of 3 groups per hour. With probability 1/3 a group consists of 2 customers, and with probability 2/3 it consists of 1 customer only. All customers have an expontial service time with a mean of 6 minutes. Groups are served in order of arrival. Within a group, customers are served in random order.

(i) Determine the distribution of the number of customers in the system.

(ii) Determine the mean sojourn time of an arbitrary customer.

Exercise 6.5.9
In a repair shop of an airline company defective airplane engines are repaired. Defects occur according to a Poisson with a rate of 1 defective engine per 2 weeks. The mean repair time of an engine is 2/3 week. The repair time distribution can be well approximated by an Erlang-2 distribution. In the repair shop only one engine can be in repair at the same time.

(i) Show that p_n, the probability that there are n engines in the repair shop, is given by

$$p_n = \frac{6}{5}\left(\frac{1}{4}\right)^n - \frac{8}{15}\left(\frac{1}{9}\right)^n .$$

(ii) Determine the mean sojourn time (waiting time plus repair time) of an engine in the repair shop.

The airline company has several spare engines in a depot. When a defect occurs, then the defective engine is immediately replaced by a spare engine (if there is one available) and the defective engine is send to the repair shop. After repair the engine is as good as new, and it is transported to the depot. When a defect occurs and no spare engine is available, the airplane has to stay on the ground and it has to wait till a spare engine becomes available.

(iii) Determine the minimal number of spare engines needed such that for 99% of the defects there is a spare engine available.

Exercise 6.5.10
Jobs arrive according to a Poisson process at a machine. The arrival rate is 25 jobs per hour. With probability 2/5 a job consists of 1 task, and with probability 3/5 it consists of 2 tasks. Tasks have an exponentially distributed processing time with a mean of 1 minute. Jobs are processed in order of arrival.

(i) Determine the distribution of the number of uncompleted tasks at the machine.

(ii) Determine the mean waiting time of a job.

(iii) Determine the mean jobs in the system.

Exercise 6.5.11
Customers arrive in groups at a server. The group size is 2 and groups arrive according to a Poisson stream with a rate of 2 groups per hour. Customers are served one by one, and they require an exponential service time with a mean of 5 minutes.

(i) Determine the distribution of the number of customers in the system.

(ii) What is the mean waiting time of the first customer in a group?

(iii) And what is the mean waiting time of the second one?

Chapter 7

$M/G/1$ queue

In the $M/G/1$ queue customers arrive according to a Poisson process with rate λ and they are treated in order of arrival. The service times are independent and identically distributed with distribution function $F_B(\cdot)$ and density $f_B(\cdot)$. For stability we have to require that the occupation rate

$$\rho = \lambda E(B) \tag{7.1}$$

is less than one. In this chapter we will derive the limiting or equilibrium distribution of the number of customers in the system and the distributions of the sojourn time, the waiting time and the busy period duration. It is further shown how the means of these quatities can be obtained by using the mean value approach.

7.1 Which limiting distribution?

The state of the $M/G/1$ queue can be described by the pair (n, x) where n denotes the number of customers in the system and x the service time already received by the customer in service. We thus need a two-dimensional state description. The first dimension is still discrete, but the other one is continuous and this essentially complicates the analysis. However, if we look at the system just after departures, then the state description can be simplified to n only, because $x = 0$ for the new customer (if any) in service. Denote by L_k^d the number of customers left behind by the kth departing customer. In the next section we will determine the limiting distribution

$$d_n = \lim_{k\to\infty} P(L_k^d = n).$$

The probability d_n can be interpreted as the fraction of customers that leaves behind n customers. But in fact we are more interested in the limiting distribution p_n defined as

$$p_n = \lim_{t\to\infty} P(L(t) = n),$$

where $L(t)$ is the number of customers in the system at time t. The probability p_n can be interpreted as the fraction of time there are n customers in the system. From this distribution we can compute, e.g., the mean number of customers in the system. Another perhaps even more important distribution is the limiting distribution of the number of customers in the system seen by an arriving customer, i.e.,

$$a_n = \lim_{k\to\infty} P(L_k^a = n),$$

where L_k^a is the number of customers in the system just before the kth arriving customer. From this distribution we can compute, e.g., the distribution of the sojourn time. What is the relation between these three distributions? It appears that they all are the same.

Of course, from the PASTA property we already know that $a_n = p_n$ for all n. We will now explain why also $a_n = d_n$ for all n. Taking the state of the system as the number of customers therein, the changes in state are of a nearest-neighbour type: if the system is in state n, then an arrival leads to a transition from n to $n+1$ and a departure from n to $n-1$. Hence, in equilibrium, the number of transitions per unit time from state n to $n+1$ will be the same as the number of transitions per unit time from $n+1$ to n. The former transitions correspond to arrivals finding n customers in the system, the frequency of which is equal to the total number of arrivals per unit time, λ, multiplied with the fraction of customers finding n customers in the system, a_n. The latter transitions correspond to departures leaving behind n customers. The frequency of these transitions is equal to the total number of departures per unit time, λ, multiplied with the fraction of customers leaving behind n customers, d_n. Equating both frequencies yields $a_n = d_n$. Note that this equality is valid for any system where customers arrive and leave one by one (see also exercise 7.10.12). Thus it also holds for, e.g., the G/G/c queue.

Summarizing, for the $M/G/1$ queue, arrivals, departures and outside observers all see the same distribution of number of customers in the system, i.e., for all n,

$$a_n = d_n = p_n.$$

7.2 Departure distribution

In this section we will determine the distribution of the number of customers left behind by a departing customer when the system is in equilibrium.

Denote by L_k^d the number of customers left behind by the kth departing customer. We first derive an equation relating the random variable L_{k+1}^d to L_k^d. The number of customers left behind by the $k+1$th customer is clearly equal to the number of customers present when the kth customer departed minus one (since the $k+1$th customer departs himself) plus the number of customers that arrives during his service time. This last number is denoted by the random variable A_{k+1}. Thus we have

$$L_{k+1}^d = L_k^d - 1 + A_{k+1},$$

which is valid if $L_k^d > 0$. In the special case $L_k^d = 0$, it is readily seen that

$$L_{k+1}^d = A_{k+1}.$$

From the two equations above it is immediately clear that the sequence $\{L_k^d\}_{k=0}^\infty$ forms a Markov chain. This Markov chain is usually called the *imbedded Markov chain*, since we look at imbedded points on the time axis, i.e., at departure instants.

We now specify the transition probabilities

$$p_{i,j} = P(L_{k+1}^d = j | L_k^d = i).$$

Clearly $p_{i,j} = 0$ for all $j < i - 1$ and $p_{i,j}$ for $j \geq i - 1$ gives the probability that exactly $j - i + 1$ customers arrived during the service time of the $k + 1$th customer. This holds for $i > 0$. In state 0 the kth customer leaves behind an empty system and then $p_{0,j}$ gives the probability that during the service time of the $k + 1$th customer exactly j customers arrived. Hence the matrix P of transition probabilities takes the form

$$P = \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \cdots \\ 0 & 0 & \alpha_0 & \alpha_1 & \cdots \\ 0 & 0 & 0 & \alpha_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where α_n denotes the probability that during a service time exactly n customers arrive. To calculate α_n we note that given the duration of the service time, t say, the number of customers that arrive during this service time is Poisson distributed with parameter λt. Hence, we have

$$\alpha_n = \int_{t=0}^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_B(t) dt. \tag{7.2}$$

The transition probability diagram is shown in figure 7.1.

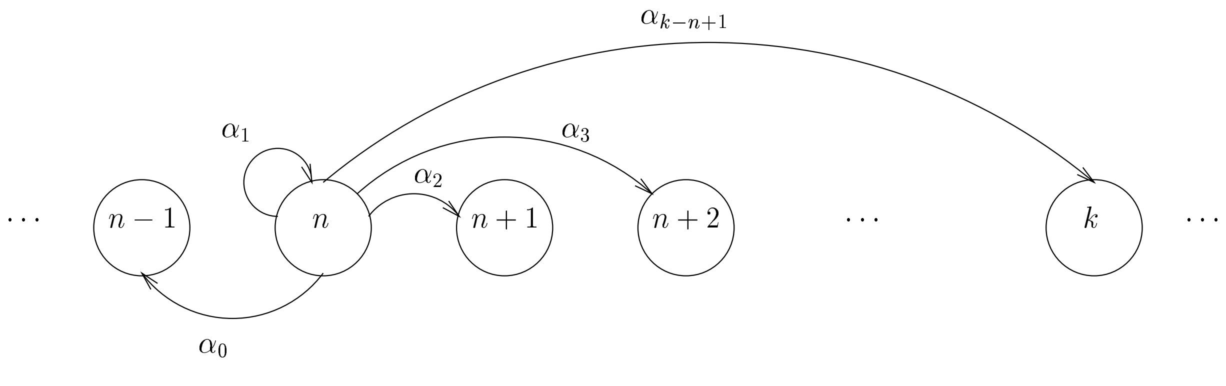

Figure 7.1: Transition probability diagram for the $M/G/1$ imbedded Markov chain

This completes the specification of the imbedded Markov chain. We now wish to determine its limiting distribution. Denote the limiting distribution of L_k^d by $\{d_n\}_{n=0}^\infty$ and the limiting random variable by L^d. So

$$d_n = P(L^d = n) = \lim_{k \to \infty} P(L_k^d = n).$$

The limiting probabilities d_n, which we know are equal to p_n, satisfy the equilibrium equations

$$\begin{aligned}
d_n &= d_{n+1}\alpha_0 + d_n\alpha_1 + \cdots + d_1\alpha_n + d_0\alpha_n \\
&= \sum_{k=0}^{n} d_{n+1-k}\alpha_k + d_0\alpha_n, \qquad n = 0, 1, \ldots
\end{aligned} \tag{7.3}$$

To solve the equilibrium equations we will use the generating function approach. Let us introduce the probability generating functions

$$P_{L^d}(z) = \sum_{n=0}^{\infty} d_n z^n, \qquad P_A(z) = \sum_{n=0}^{\infty} \alpha_n z^n,$$

which are defined for all $z \le 1$. Multiplying (7.3) by z^n and summing over all n leads to

$$\begin{aligned}
P_{L^d}(z) &= \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} d_{n+1-k}\alpha_k + d_0\alpha_n\right) z^n \\
&= z^{-1}\sum_{n=0}^{\infty}\sum_{k=0}^{n} d_{n+1-k}z^{n+1-k}\alpha_k z^k + \sum_{n=0}^{\infty} d_0\alpha_n z^n \\
&= z^{-1}\sum_{k=0}^{\infty}\sum_{n=k}^{\infty} d_{n+1-k}z^{n+1-k}\alpha_k z^k + d_0 P_A(z) \\
&= z^{-1}\sum_{k=0}^{\infty}\alpha_k z^k \sum_{n=k}^{\infty} d_{n+1-k}z^{n+1-k} + d_0 P_A(z) \\
&= z^{-1}P_A(z)(P_{L^d}(z) - d_0) + d_0 P_A(z).
\end{aligned}$$

Hence we find

$$P_{L^d}(z) = \frac{d_0 P_A(z)(1 - z^{-1})}{1 - z^{-1}P_A(z)}.$$

To determine the probability d_0 we note that d_0 is equal to p_0, which is the fraction of time the system is empty. Hence $d_0 = p_0 = 1 - \rho$ (alternatively, d_0 follows from the requirement $P_{L^d}(1) = 1$). So, by multiplying numerator and denominator by $-z$ we obtain

$$P_{L^d}(z) = \frac{(1-\rho)P_A(z)(1-z)}{P_A(z) - z}. \tag{7.4}$$

By using (7.2), the generating function $P_A(z)$ can be rewritten as

$$\begin{aligned}
P_A(z) &= \sum_{n=0}^{\infty}\int_{t=0}^{\infty}\frac{(\lambda t)^n}{n!}e^{-\lambda t}f_B(t)dtz^n \\
&= \int_{t=0}^{\infty}\sum_{n=0}^{\infty}\frac{(\lambda t z)^n}{n!}e^{-\lambda t}f_B(t)dt \\
&= \int_{t=0}^{\infty}\sum_{n=0}^{\infty} e^{-(\lambda-\lambda z)t}f_B(t)dt \\
&= \widetilde{B}(\lambda - \lambda z)
\end{aligned} \tag{7.5}$$

Substitution of (7.5) into (7.4) finally yields

$$P_{L^d}(z) = \frac{(1-\rho)\widetilde{B}(\lambda - \lambda z)(1-z)}{\widetilde{B}(\lambda - \lambda z) - z}. \tag{7.6}$$

This formula is one form of the *Pollaczek-Khinchin formula*. In the following sections we will derive similar formulas for the sojourn time and waiting time. By differentiating formula (7.6) we can determine the moments of the queue length (see section 2.2). To find its distribution, however, we have to invert formula (7.6), which usually is very difficult. In the special case that $\widetilde{B}(s)$ is a quotient of polynomials in s, i.e., *a rational function*, then in principle the right-hand side of (7.6) can be decomposed into partial fractions, the inverse transform of which can be easily determined. The service time has a rational transform for, e.g., mixtures of Erlang distributions or Hyperexponential distributions (see section 2.4). The inversion of (7.6) is demonstrated below for exponential and Erlang-2 service times.

Example 7.2.1 *($M/M/1$)*
Suppose the service time is exponentially distributed with mean $1/\mu$. Then

$$\widetilde{B}(s) = \frac{\mu}{\mu + s}.$$

Thus

$$P_{L^d}(z) = \frac{(1-\rho)\frac{\mu}{\mu+\lambda-\lambda z}(1-z)}{\frac{\mu}{\mu+\lambda-\lambda z} - z} = \frac{(1-\rho)\mu(1-z)}{\mu - z(\mu+\lambda-\lambda z)} = \frac{(1-\rho)\mu(1-z)}{(\mu-\lambda z)(1-z)} = \frac{1-\rho}{1-\rho z}.$$

Hence

$$d_n = p_n = (1-\rho)\rho^n, \qquad n = 0, 1, 2, \ldots$$

Example 7.2.2 *($M/E_2/1$)*
Suppose the service time is Erlang-2 distributed with mean $2/\mu$. Then (see subsection 2.4.4)

$$\widetilde{B}(s) = \left(\frac{\mu}{\mu + s}\right)^2,$$

so

$$\begin{aligned}
P_{L^d}(z) &= \frac{(1-\rho)\left(\frac{\mu}{\mu+\lambda-\lambda z}\right)^2(1-z)}{\left(\frac{\mu}{\mu+\lambda-\lambda z}\right)^2 - z} \\
&= \frac{(1-\rho)\mu^2(1-z)}{\mu^2 - z(\mu+\lambda-\lambda z)^2} \\
&= \frac{(1-\rho)(1-z)}{1 - z(1+\rho(1-z)/2)^2} \\
&= \frac{1-\rho}{1-\rho z - \rho^2 z(1-z)/4}.
\end{aligned}$$

For $\rho = 1/3$ we then find

$$
\begin{aligned}
P_{L^d}(z) &= \frac{2/3}{1 - z/3 - z(1-z)/36} = \frac{24}{36 - 13z + z^2} \\
&= \frac{24}{(4-z)(9-z)} = \frac{24/5}{4-z} - \frac{24/5}{9-z} = \frac{6/5}{1-z/4} - \frac{8/15}{1-z/9}.
\end{aligned}
$$

Hence,

$$
d_n = p_n = \frac{6}{5}\left(\frac{1}{4}\right)^n - \frac{8}{15}\left(\frac{1}{9}\right)^n, \qquad n = 0, 1, 2, \ldots
$$

Example 7.2.3 *($M/H_2/1$)*
Suppose that $\lambda = 1$ and that the service time is hyperexponentially distributed with parameters $p_1 = 1 - p_2 = 1/4$ and $\mu_1 = 1$, $\mu_2 = 2$. So the mean service time is equal to $1/4 \cdot 1 + 3/4 \cdot 1/2 = 5/8$. The Laplace-Stieltjes transform of the service time is given by (see subsection 2.4.5)

$$
\widetilde{B}(s) = \frac{1}{4} \cdot \frac{1}{1+s} + \frac{3}{4} \cdot \frac{2}{2+s} = \frac{1}{4} \cdot \frac{8+7s}{(1+s)(2+s)}.
$$

Thus we have

$$
\begin{aligned}
P_{L^d}(z) &= \frac{\frac{3}{8}\frac{1}{4}\frac{15-7z}{(2-z)(3-z)}(1-z)}{\frac{1}{4}\frac{15-7z}{(2-z)(3-z)} - z} \\
&= \frac{3}{8} \cdot \frac{(15-7z)(1-z)}{(15-7z) - 4z(2-z)(3-z)} \\
&= \frac{3}{8} \cdot \frac{15-7z}{(3-2z)(5-2z)} \\
&= \frac{3}{8} \cdot \frac{9/4}{3-2z} + \frac{3}{8} \cdot \frac{5/4}{5-2z/5} \\
&= \frac{9/32}{1-2z/3} + \frac{3/32}{1-2z/5}.
\end{aligned}
$$

So

$$
d_n = p_n = \frac{9}{32}\left(\frac{2}{3}\right)^n + \frac{3}{32}\left(\frac{2}{5}\right)^n, \qquad n = 0, 1, 2, \ldots
$$

7.3 Distribution of the sojourn time

We now turn to the calculation of how long a customer spends in the system. We will show that there is a nice relationship between the transforms of the time spent in the system and the departure distribution.

Let us consider a customer arriving at the system in equilibrium. Denote the total time spent in the system for this customer by the random variable S with distribution function $F_S(\cdot)$ and density $f_S(\cdot)$. The distribution of the number of customers left behind upon departure of our customer is equal to $\{d_n\}_{n=0}^{\infty}$ (since the system is in equilibrium). In considering a first-come first-served system it is clear that all customers left behind are precisely those who arrived during his stay in the system. Thus we have (cf. (7.2))

$$d_n = \int_{t=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_S(t) dt.$$

Hence, we find similarly to (7.5) that

$$P_{L^d}(z) = \widetilde{S}(\lambda - \lambda z).$$

Substitution of this relation into (7.6) yields

$$\widetilde{S}(\lambda - \lambda z) = \frac{(1-\rho)\widetilde{B}(\lambda - \lambda z)(1-z)}{\widetilde{B}(\lambda - \lambda z) - z}.$$

Making the change of variable $s = \lambda - \lambda z$ we finally arrive at

$$\widetilde{S}(s) = \frac{(1-\rho)\widetilde{B}(s)s}{\lambda \widetilde{B}(s) + s - \lambda}. \tag{7.7}$$

This formula is also known as the *Pollaczek-Khinchin formula.*

Example 7.3.1 *(M/M/1)*
For exponential service times with mean $1/\mu$ we have

$$\widetilde{B}(s) = \frac{\mu}{\mu + s}.$$

Thus

$$\widetilde{S}(s) = \frac{(1-\rho)\frac{\mu}{\mu+s}s}{\lambda\frac{\mu}{\mu+s} + s - \lambda} = \frac{(1-\rho)\mu s}{\lambda\mu + (s-\lambda)(\mu+s)} = \frac{(1-\rho)\mu s}{(\mu-\lambda)s + s^2} = \frac{\mu(1-\rho)}{\mu(1-\rho)+s}.$$

Hence, S is exponentially distributed with parameter $\mu(1-\rho)$, i.e.,

$$F_S(t) = P(S \le t) = 1 - e^{-\mu(1-\rho)t}, \qquad t \ge 0.$$

Example 7.3.2 *(M/E_2/1)*
Suppose that $\lambda = 1$ and that the service time is Erlang-2 distributed with mean $2/3$, so

$$\widetilde{B}(s) = \left(\frac{3}{3+s}\right)^2.$$

Then it follows that (verify)

$$F_S(t) = \frac{8}{5}(1 - e^{-3t}) - \frac{3}{5}(1 - e^{-8t}), \qquad t \ge 0.$$

Example 7.3.3 *(M/H_2/1)*
Consider example 7.2.3 again. From (7.7) we obtain (verify)

$$F_S(t) = \frac{27}{32}(1 - e^{-t/2}) + \frac{5}{32}(1 - e^{-3t/2}), \qquad t \ge 0.$$

7.4 Distribution of the waiting time

We have that S, the time spent in the system by a customer, is the sum of W (his waiting time) and B (his service time), where W and B are independent. Since the transform of the sum of two independent random variables is the product of the transforms of these two random variables (see section 2.3), it holds that

$$\widetilde{S}(s) = \widetilde{W}(s) \cdot \widetilde{B}(s). \tag{7.8}$$

Together with (7.7) it follows that

$$\widetilde{W}(s) = \frac{(1-\rho)s}{\lambda\widetilde{B}(s) + s - \lambda}, \tag{7.9}$$

which is the third form of the Pollaczek-Khinchin formula.

Example 7.4.1 *(M/M/1)*
For exponential service times with mean $1/\mu$ we have

$$\widetilde{B}(s) = \frac{\mu}{\mu + s}.$$

Then from (7.9) it follows that (verify)

$$\widetilde{W}(s) = (1-\rho) + \rho \cdot \frac{\mu(1-\rho)}{\mu(1-\rho) + s}.$$

The inverse transform yields

$$F_W(t) = P(W \le t) = (1-\rho) + \rho(1 - e^{-\mu(1-\rho)t}), \qquad t \ge 0.$$

Hence, with probability $(1-\rho)$ the waiting time is zero (i.e., the system is empty on arrival) and, given that the waiting time is positive (i.e., the system is not empty on arrival), the waiting time is exponentially distributed with parameter $\mu(1-\rho)$ (see also section 4.4).

7.5 Lindley's equation

In the literature there are several alternative approaches to derive the Pollaczek-Khinchin formulas. In this section we present one that is based on a fundamental relationship between the waiting time of the nth customer and the $n+1$th customer.

Denote by A_n the interarrival time between the nth and $n+1$th customer arriving at the system. Further let W_n and B_n denote the waiting time and the service time, respectively, for the nth customer. Clearly, if $A_n \le W_n + B_n$, then the waiting time of the $n+1$th customer is equal to $W_n + B_n - A_n$. If on the other hand $A_n > W_n + B_n$, then his waiting time is equal to zero. Summarizing, we have

$$W_{n+1} = \max(W_n + B_n - A_n, 0). \tag{7.10}$$

This equation is commonly referred to as *Lindley's equation*. We denote the limit of W_n as n goes to infinity by W (which exists if ρ is less than one). Then, by letting n go to infinity in (7.10) we obtain the limiting form

$$W = \max(W + B - A, 0) \tag{7.11}$$

(where we also dropped the subscripts of A_n and B_n, since we are considering the limiting situation). With $S = W + B$, this equation may also be written as

$$W = \max(S - A, 0) \tag{7.12}$$

Note that the interarrival time A is exponentially distributed with parameter λ and that the random variables S and A are independent. We now calculate the transform of W directly from equation (7.12). This gives

$$\begin{aligned}
\widetilde{W}(s) &= E(e^{-sW}) \\
&= E(e^{-s\max(S-A,0)}) \\
&= \int_{x=0}^{\infty}\int_{y=0}^{\infty} e^{-s\max(x-y,0)} f_S(x)\lambda e^{-\lambda y} dx dy \\
&= \int_{x=0}^{\infty}\int_{y=0}^{x} e^{-s(x-y)} f_S(x)\lambda e^{-\lambda y} dx dy + \int_{x=0}^{\infty}\int_{y=x}^{\infty} f_S(x)\lambda e^{-\lambda y} dx dy \\
&= \frac{\lambda}{s-\lambda}\int_{x=0}^{\infty}(e^{-\lambda x} - e^{-sx}) f_S(x) dx + \int_{x=0}^{\infty} e^{-\lambda x} f_S(x) dx \\
&= \frac{\lambda}{s-\lambda}(\widetilde{S}(\lambda) - \widetilde{S}(s)) + \widetilde{S}(\lambda) \\
&= \frac{s}{s-\lambda}\widetilde{S}(\lambda) - \frac{\lambda}{s-\lambda}\widetilde{S}(s).
\end{aligned}$$

Substitution of (7.8) then yields

$$\widetilde{W}(s) = \frac{\widetilde{S}(\lambda)s}{\lambda\widetilde{B}(s) + s - \lambda}. \tag{7.13}$$

It remains to find $\widetilde{S}(\lambda)$. Realizing that

$$\widetilde{S}(\lambda) = \int_{x=0}^{\infty} e^{-\lambda x} f_S(x) dx = \int_{x=0}^{\infty} P(A > x) f_S(x) dx = P(A > S),$$

it follows that $\widetilde{S}(\lambda)$ is precisely the probability that the system is empty on arrival, so

$$\widetilde{S}(\lambda) = 1 - \rho.$$

Substitution of this relation into (7.13) finally yields formula (7.9).

7.6 Mean value approach

The mean waiting time can of course be calculated from the Laplace-Stieltjes transform (7.9) by differentiating and substituting $s = 0$ (see section 2.3). In this section we show that the mean waiting time can also be determined directly (i.e., without transforms) with the *mean value approach*.

A new arriving customer first has to wait for the *residual service time* of the customer in service (if there is one) and then continues to wait for the servicing of all customers who were already waiting in the queue on arrival. By PASTA we know that with probability ρ the server is busy on arrival. Let the random variable R denote the residual service time and let L^q denote the number of customers waiting in the queue. Hence,

$$E(W) = E(L^q)E(B) + \rho E(R),$$

and by Little's law (applied to the queue),

$$E(L^q) = \lambda E(W).$$

So we find

$$E(W) = \frac{\rho E(R)}{1-\rho}. \tag{7.14}$$

Formula (7.14) is commonly referred to as the Pollaczek-Khinchin mean value formula. It remains to calculate the mean residual service time. In the following section we will show that

$$E(R) = \frac{E(B^2)}{2E(B)}, \tag{7.15}$$

which may also be written in the form

$$E(R) = \frac{E(B^2)}{2E(B)} = \frac{\sigma_B^2 + E(B)^2}{2E(B)} = \frac{1}{2}(c_B^2 + 1)E(B). \tag{7.16}$$

An important observation is that, clearly, the mean waiting time only depends upon the first two moments of service time (and not upon its distribution). So in practice it is sufficient to know the mean and standard deviation of the service time in order to estimate the mean waiting time.

7.7 Residual service time

Suppose that our customer arrives when the server is busy and denote the total service time of the customer in service by X. Further let $f_X(\cdot)$ denote the density of X. The basic observation to find $f_X(\cdot)$ is that it is more likely that our customer arrives in a long service time than in a short one. So the probability that X is of length x should be proportional to the length x as well as the frequency of such service times, which is $f_B(x)dx$. Thus we may write

$$P(x \leq X \leq x + dx) = f_X(x)dx = Cxf_B(x)dx,$$

where C is a constant to normalize this density. So

$$C^{-1} = \int_{x=0}^{\infty} xf_B(x)dx = E(B).$$

Hence

$$f_X(x) = \frac{xf_B(x)}{E(B)}\,.$$

Given that our customer arrives in a service time of length x, the arrival instant will be a random point within this service time, i.e., it will be uniformly distributed within the service time interval $(0, x)$. So

$$P(t \leq R \leq t + dt | X = x) = \frac{dt}{x}, \qquad t \leq x.$$

Of course, this conditional probability is zero when $t > x$. Thus we have

$$P(t \leq R \leq t + dt) = f_R(t)dt = \int_{x=t}^{\infty} \frac{dt}{x} f_X(x)dx = \int_{x=t}^{\infty} \frac{f_B(x)}{E(B)}\,dxdt = \frac{1 - F_B(t)}{E(B)}\,dt.$$

This gives the final result

$$f_R(t) = \frac{1 - F_B(t)}{E(B)}\,,$$

from which we immediately obtain, by partial integration,

$$E(R) = \int_{t=0}^{\infty} tf_R(t)dt = \frac{1}{E(B)} \int_{t=0}^{\infty} t(1 - F_B(t))dt = \frac{1}{E(B)} \int_{t=0}^{\infty} \frac{1}{2}t^2 f_B(t)dt = \frac{E(B^2)}{2E(B)}\,.$$

This computation can be repeated to obtain all moments of R, yielding

$$E(R^n) = \frac{E(B^{n+1})}{(n+1)E(B)}\,.$$

Example 7.7.1 *(Erlang service times)*
For an Erlang-r service time with mean r/μ we have

$$E(B) = \frac{r}{\mu}\,, \qquad \sigma^2(B) = \frac{r}{\mu^2}\,,$$

so

$$E(B^2) = \sigma^2(B) + (E(B))^2 = \frac{r(1+r)}{\mu^2}\,.$$

Hence (see also (6.9))

$$E(R) = \frac{1+r}{2\mu}$$

7.8 Variance of the waiting time

The mean value approach fails to give more information than the mean of the waiting time (or other performance characteristics). But to obtain information on the variance or higher moments of the waiting time we can use formula (7.9) for the Laplace-Stieltjes transform of the waiting time. This is demonstrated below.

We first rewrite formula (7.9) as

$$\widetilde{W}(s) = \frac{1-\rho}{1-\rho\left(\frac{1-\widetilde{B}(s)}{sE(B)}\right)},$$

The term between brackets can be recognised as the transform of R, since by partial integration we have

$$\begin{aligned}\widetilde{R}(s) &= E(e^{-sR}) = \int_{t=0}^{\infty} e^{-st} f_R(t)dt = \frac{1}{E(B)}\int_{t=0}^{\infty} e^{-st}(1-F_B(t))dt \\ &= \frac{1}{E(B)}\left(\frac{1}{s} - \int_{t=0}^{\infty}\frac{1}{s}e^{-st}f_B(t)dt\right) = \frac{1-\widetilde{B}(s)}{sE(B)}.\end{aligned}$$

Hence

$$\widetilde{W}(s) = \frac{1-\rho}{1-\rho\widetilde{R}(s)}. \tag{7.17}$$

By differentiating (7.17) and substituting $s = 0$ we retrieve formula (7.14), i.e.,

$$E(W) = \frac{\rho E(R)}{1-\rho}.$$

By differentiating (7.17) twice and substituting $s = 0$ we find

$$E(W^2) = 2(E(W))^2 + \frac{\rho E(R^2)}{1-\rho}.$$

Hence, for the variance of the waiting time we now obtain

$$\sigma^2(W) = E(W^2) - (E(W))^2 = \frac{\rho}{(1-\rho)^2}\left(\rho(E(R))^2 + (1-\rho)E(R^2)\right).$$

The first two moments of the *conditional waiting time* $W|W > 0$ can be calculated from

$$E(W|W>0) = \frac{E(W)}{\rho}, \qquad E(W^2|W>0) = \frac{E(W^2)}{\rho}.$$

It then follows, after some algebra, that the squared coefficient of variation of $W|W > 0$ is given by the simple equation

$$c^2_{W|W>0} = \rho + (1-\rho)c_R^2 = 1 + (1-\rho)(c_R^2 - 1).$$

This implies that for ρ near to 1 the squared coefficient of variation of the conditional waiting time is close to 1. Hence, as a rule of thumb, we may think of the conditional waiting time as an exponential random variable (which has a coefficient of variation equal to 1) and thus obtain a rough estimate for its distribution.

7.9 Distribution of the busy period

The mean duration of a busy period in the $M/G/1$ queue can be determined in exactly the same way as for the $M/M/1$ queue in subsection 4.6.1. Thus we have

$$E(BP) = \frac{E(B)}{1-\rho}. \tag{7.18}$$

The calculation of the distribution of a busy period in the $M/G/1$ queue is more complicated.

We first derive an important relation for the duration of a busy period. What happens in a busy period? We start with the service of the first customer. His service time is denoted by B_1. During his service new customers may arrive to the system. Denote this (random) number by N and label these customers by $C_1, \ldots, C_N$. At the departure of the first customer we take customer C_1 into service. However, instead of letting the other customers $C_2, \ldots, C_N$ wait for their turn, we choose to take them temporarily out of the system. Customer C_2 will be put into the system again and taken into service as soon as the system is empty again. So customers arriving during the service of C_1 will be served first (as well as the ones arriving during their service, and so on). Thus it is as if C_1 initiates a new busy period for which C_2 has to wait. This busy period will be called a sub-busy period. In the same way C_3 has to wait for the sub-busy period initiated by C_2, and so on. Finally the sub-busy period due to C_N completes the major busy period. Note that our customers $C_2, \ldots, C_N$ are not treated first-come first-served anymore. But this does not affect the duration of the (major) busy period, because its duration is independent of the order in which customers are served. Denote the busy period by BP and the sub-busy period due to C_i by BP_i. Then we have the important relation

$$BP = B_1 + BP_1 + \cdots + BP_N \tag{7.19}$$

where the random variables $BP, BP_1, BP_2, \ldots$ are independent and all have the same distribution, and further, they are independent of N. The number N of course depends on the service time B_1.

We will now use relation (7.19) to derive the Laplace-Stieltjes transform of the busy period BP. By conditioning on the length of B_1 we get

$$\widetilde{BP}(s) = E(e^{-sBP}) = \int_{t=0}^{\infty} E(e^{-sBP}|B_1 = t) f_B(t) dt$$

and by also conditioning on N,

$$\begin{aligned}
E(e^{-sBP}|B_1 = t) &= \sum_{n=0}^{\infty} E(e^{-sBP}|B_1 = t, N = n) P(N = n|B_1 = t) \\
&= \sum_{n=0}^{\infty} E(e^{-s(B_1+BP_1+\cdots+BP_N)}|B_1 = t, N = n) \frac{(\lambda t)^n}{n!} e^{-\lambda t} \\
&= \sum_{n=0}^{\infty} E(e^{-s(t+BP_1+\cdots+BP_n)}) \frac{(\lambda t)^n}{n!} e^{-\lambda t}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=0}^{\infty} E(e^{-st} \cdot e^{-sBP_1} \cdots e^{-sBP_n}) \frac{(\lambda t)^n}{n!} e^{-\lambda t} \\
&= \sum_{n=0}^{\infty} e^{-st} (E(e^{-sBP}))^n \frac{(\lambda t)^n}{n!} e^{-\lambda t} \\
&= e^{-(s+\lambda-\lambda \widetilde{BP}(s))t}
\end{aligned}
$$

Thus we have

$$
\widetilde{BP}(s) = \int_{t=0}^{\infty} e^{-(s+\lambda-\lambda \widetilde{BP}(s))t} f_B(t) dt,
$$

which can be recognised as

$$
\widetilde{BP}(s) = \widetilde{B}(s + \lambda - \lambda \widetilde{BP}(s)). \tag{7.20}
$$

In the example below it is shown that even for the simple $M/M/1$ queue it is not easy to determine the distribution of BP from this equation.

Example 7.9.1 *($M/M/1$)*
For the $M/M/1$ we have

$$
\widetilde{B}(s) = \frac{\mu}{\mu + s}.
$$

In this case equation (7.20) reduces to

$$
\widetilde{BP}(s) = \frac{\mu}{\mu + s + \lambda - \lambda \widetilde{BP}(s)}.
$$

So $\widetilde{BP}(s)$ is a root of the quadratic equation

$$
\lambda (\widetilde{BP}(s))^2 - (\lambda + \mu + s) \widetilde{BP}(s) + \mu = 0.
$$

Solving this equation and restricting our solution to the case for which $0 \leq \widetilde{BP}(s) \leq 1$ for all $s \geq 0$ yields

$$
\widetilde{BP}(s) = \frac{1}{2\lambda} \left(\lambda + \mu + s - \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu} \right).
$$

This equation may be inverted to obtain the density of the busy period. This gives an expression involving a Bessel function (see subsection 4.6.2).

It is, however, straightforward to determine the moments of BP from equation (7.20) by differentiating and substituting $s = 0$ (see section 2.3). This will be demonstrated below.

Differentiating (7.20) and substituting $s = 0$ yields

$$
-E(BP) = \widetilde{BP}^{(1)}(0) = \widetilde{B}^{(1)}(0) \cdot (1 - \lambda \widetilde{BP}^{(1)}(0)) = -E(B)(1 + \lambda E(BP)).
$$

Hence, with $\rho = \lambda E(B)$, we retrieve formula (7.18), i.e.,

$$E(BP) = \frac{E(B)}{1-\rho}. \tag{7.21}$$

By differentiating (7.20) twice and substituting $s = 0$ we get

$$\begin{aligned} E(BP^2) &= \widetilde{BP}^{(2)}(0) \\ &= \widetilde{B}^{(2)}(0) \cdot (1 - \lambda \widetilde{BP}^{(1)}(0))^2 + \widetilde{B}^{(1)}(0) \cdot (-\lambda \widetilde{BP}^{(2)}(0)) \\ &= E(B^2)(1 + \lambda E(BP))^2 + \lambda E(B) E(BP^2). \end{aligned}$$

From this we obtain

$$E(BP^2) = \frac{E(B^2)}{(1-\rho)^3}. \tag{7.22}$$

From (7.21) and (7.22) it follows that the squared coefficient of variation, c^2_{BP}, is given by

$$c^2_{BP} = \frac{c^2_B}{1-\rho} + \frac{\rho}{1-\rho}. \tag{7.23}$$

Clearly, as ρ tends to one, the variation in BP explodes.

Example 7.9.2 *($M/M/1$)*
Since $c^2_B = 1$ for exponential service times it follows from (7.23) that

$$c^2_{BP} = \frac{1+\rho}{1-\rho}.$$

In table 7.1 we list c^2_{BP} for increasing values of ρ.

ρ	0.5	0.6	0.7	0.8	0.9	0.95
c^2_{BP}	3	4	5.7	9	19	39

Table 7.1: The squared coefficient of the busy period for the $M/M/1$ queue

7.10 Exercises

Exercise 7.10.1 *(Post office)*
At a post office customers arrive according to a Poisson process with a rate of 30 customers per hour. A quarter of the customers wants to cash a cheque. Their service time is exponentially distributed with a mean of 2 minutes. The other customers want to buy stamps and their service times are exponentially distributed with a mean of 1 minute. In the post office there is only one server.

(i) Determine the generating function $P_L(z)$ of the number of customers in the system.

(ii) Determine the distribution of the number of customers in the system.

(iii) Determine the mean number of customers in the system.

(iv) Determine $\widetilde{S}(s)$, the Laplace-Stieltjes transform of the sojourn time.

(v) Determine the mean and the distribution of the sojourn time.

(vi) Determine the mean busy period duration.

(vii) Determine the mean number of customers in the system and the mean sojourn time in case all customers have an exponentially distributed service time with a mean of 75 seconds.

Exercise 7.10.2
Consider a single machine where jobs arrive according to a Poisson stream with a rate of 10 jobs per hour. The processing time of a job consists of two phases. Each phase takes an exponential time with a mean of 1 minute.

(i) Determine the Lapace-Stieltjes transform of the processing time.

(ii) Determine the distribution of the number of jobs in the system.

(iii) Determine the mean number of jobs in the system and the mean production lead time (waiting time plus processing time).

Exercise 7.10.3 *(Post office)*
At a post office customers arrive according to a Poisson process with a rate of 60 customers per hour. Half of the customers have a service time that is the sum of a fixed time of 15 seconds and an exponentially distributed time with a mean of 15 seconds. The other half have an exponentially distributed service time with a mean of 1 minute.
Determine the mean waiting time and the mean number of customers waiting in the queue.

Exercise 7.10.4 *(Two phase production)*
A machine produces products in two phases. The first phase is standard and the same for all products. The second phase is customer specific (the finishing touch). The first (resp. second) phase takes an exponential time with a mean of 10 (resp. 2) minutes. Orders for the production of one product arrive according to a Poisson stream with a rate of 3 orders per hour. Orders are processed in order of arrival.
Determine the mean production lead time of an order.

Exercise 7.10.5
Consider a machine where jobs are being processed. The mean production time is 4 minutes and the standard deviation is 3 minutes. The mean number of jobs arriving per hour is 10. Suppose that the interarrival times are exponentially distributed.
Determine the mean waiting time of the jobs.

Exercise 7.10.6
Consider an $M/G/1$ queue, where the server successfully completes a service time with probability p. If a service time is not completed successfully, it has to be repeated until it is successful. Determine the mean sojourn time of a customer in the following two cases:

(i) The repeated service times are identical.

(ii) The repeated service times are independent, and thus (possibly) different.

Exercise 7.10.7
At the end of a production process an operator manually performs a quality check. The time to check a product takes on average 2 minutes with a standard deviation of 1 minute. Products arrive according to a Poisson stream. One considers to buy a machine that is able to automatically check the products. The machine needs exactly 84 seconds to perform a quality check. Since this machine is expensive, one decides to buy the machine only if it is able to reduce lead time for a quality check (i.e. the time that elapses from the arrival of a product till the completion of its quality check) to one third of the lead time in the present situation.
So only if the arrival rate of products exceeds a certain threshold, one will decide to buy the machine. Calculate the value of this threshold.

Exercise 7.10.8 *(Robotic dairy barn)*
In a robotic dairy barn cows are automatically milked by a robot. The cows are lured into the robot by a feeder with nice food that can only be reached by first passing through the robot. When a cow is in the robot, the robot first detects whether the cow has to be milked. If so, then the cow will be milked, and otherwise, the cow can immediately leave the robot and walk to the feeder. A visit to the robot with (resp. without) milking takes an exponential time with a mean of 6 (resp. 3) minutes. Cows arrive at the robot according to a Poisson process with a rate of 10 cows per hour, a quarter of which will be milked.

(i) Show that the Laplace-Stieltjes transform of the service time in minutes of an arbitrary cow at the robot is given by

$$\widetilde{B}(s) = \frac{1}{4} \cdot \frac{4+21s}{(1+6s)(1+3s)}.$$

(ii) Show that the Laplace-Stieltjes transform of the waiting time in minutes of an arbitrary cow in front of the robot is given by

$$\widetilde{W}(s) = \frac{3}{8} + \frac{9}{16} \cdot \frac{1}{1+12s} + \frac{1}{16} \cdot \frac{1}{1+4s}.$$

(iii) Determine the fraction of cows for which the waiting time in front of the robot is less than 3 minutes.

(iv) Determine the mean waiting time in front of the robot.

Exercise 7.10.9
Consider a machine processing parts. These parts arrive according to a Poisson stream with a rate of 1 product per hour. The machine processes one part at a time. Each part receives two operations. The first operation takes an exponential time with a mean of 15 minutes. The second operation is done immediately after the first one and it takes an exponential time with a mean of 20 minutes.

(i) Show that the Laplace-Stieltjes transform of the production lead time (waiting time plus processing time) in hours is given by

$$\widetilde{S}(s) = \frac{5}{4} \cdot \frac{1}{1+s} - \frac{1}{4} \cdot \frac{5}{5+s}.$$

(ii) Determine the distribution of the production lead time and the mean production lead time.

One uses 3 hours as a norm for the production lead time. When the production lead time of a part exceeds this norm, it costs 100 dollar.

(iii) Calculate the mean cost per hour.

Exercise 7.10.10 *(Warehouse)*
In a warehouse for small items orders arrive according to a Poisson stream with a rate of 6 orders per hour. An order is a list with the quantities of products requested by a customer. The orders are picked one at a time by one order picker. For a quarter of the orders the pick time is exponentially distributed with a mean of 10 minutes and for the other orders the pick time is exponentially distributed with a mean of 5 minutes.

(i) Show that the Laplace-Stieltjes transform of the pick time in minutes of an arbitrary order is given by

$$\widetilde{B}(s) = \frac{1}{4} \cdot \frac{4 + 35s}{(1 + 10s)(1 + 5s)} \, .$$

(ii) Show that the Laplace-Stieltjes transform of the lead time (waiting time plus pick time) in minutes of an arbitrary order is given by

$$\widetilde{S}(s) = \frac{5}{32} \cdot \frac{3}{3 + 20s} + \frac{27}{32} \cdot \frac{1}{1 + 20s} \, .$$

(iii) Determine the fraction of orders for which the lead time is longer than half an hour.

(iv) Determine the mean lead time.

Exercise 7.10.11 *(Machine with breakdowns)*
Consider a machine processing parts. Per hour arrive according to a Poisson process on average 5 orders for the production of a part. The processing time of a part is exactly 10 minutes. During processing, however, tools can break down. When this occurs, the machine immediately stops, broken tools are replaced by new ones and then the machine resumes production again. Replacing tools takes exactly 2 minutes. The time that elapses between two breakdowns is exponentially distributed with a mean of 20 minutes. Hence, the total production time of a part, B, consists of the processing time of 10 minutes plus a random number of interruptions of 2 minutes to replace broken tools.

(i) Determine the mean and variance of the production time B.

(ii) Determine the mean lead time (waiting time plus production time) of an order.

Exercise 7.10.12 *(Arrival and departure distribution)*
Consider a queueing system in which customers arrive one by one and leave one by one. So the number of customers in the system only changes by $+1$ or -1. We wish to prove that $a_n = d_n$. Suppose that the system is empty at time $t = 0$.

(i) Show that if $L^d_{k+1} \leq n$, then $L^a_{k+n+1} \leq n$.

(ii) Show that if $L^a_{k+n+1} \leq n$, then $L^d_k \leq n$.

(iii) Show that *(i)* and *(ii)* give, for any $n \geq 0$,

$$\lim_{k\to\infty} P(L^d_k \leq n) = \lim_{k\to\infty} P(L^a_k \leq n).$$

Exercise 7.10.13 *(Number served in a busy period)*
Let the random variable N_{bp} denote the number of customers served in a busy period. Define the probabilities f_n by

$$f_n = P(N_{bp} = n).$$

We now wish to find its generating function

$$F_{N_{bp}}(z) = \sum_{n=1}^{\infty} f_n z^n.$$

(i) Show that $F_{N_{bp}}(z)$ satisfies

$$F_{N_{bp}}(z) = z\widetilde{B}(\lambda - \lambda F_{N_{bp}}(z)). \tag{7.24}$$

(ii) Show that the mean number of customers served in a busy period is given by

$$E(N_{bp}) = \frac{1}{1-\rho}.$$

(iii) Solve equation (7.24) for the $M/M/1$ queue.

(iv) Solve equation (7.24) for the $M/D/1$ queue (*Hint*: Use Lagrange inversion formula, see, e.g., [30]).

Chapter 8

$G/M/1$ queue

In this chapter we study the $G/M/1$ queue, which forms the dual of the $M/G/1$ queue. In this system customers arrive one by one with interarrival times identically and independently distributed according to an arbitrary distribution function $F_A(\cdot)$ with density $f_A(\cdot)$. The mean interarrival time is equal to $1/\lambda$. The service times are exponentially distributed with mean $1/\mu$. For stability we again require that the occupation rate $\rho = \lambda/\mu$ is less than one.

The state of the $G/M/1$ queue can be described by the pair (n, x) where n denotes the number of customers in the system and x the elapsed time since the last arrival. So we need a complicated two-dimensional state description. However, like for the $M/G/1$ queue, the state description is much easier at special points in time. If we look at the system on arrival instants, then the state description can be simplified to n only, because $x = 0$ at an arrival. Denote by L_k^a the number of customers in the system just before the kth arriving customer. In the next section we will determine the limiting distribution

$$a_n = \lim_{k\to\infty} P(L_k^a = n).$$

From this distribution we will be able to calculate the distribution of the sojourn time.

8.1 Arrival distribution

In this section we will determine the distribution of the number of customers found in the system just before an arriving customer when the system is in equilibrium.

We first derive a relation between the random variables L_{k+1}^a and L_k^a. Defining the random variable D_{k+1} as the number of customers served between the arrival of the kth and $k+1$th customer, it follows that

$$L_{k+1}^a = L_k^a + 1 - D_{k+1}.$$

From this equation it is immediately clear that the sequence $\{L_k^a\}_{k=0}^{\infty}$ forms a Markov chain. This Markov chain is called the $G/M/1$ *imbedded Markov chain*.

We must now calculate the associated transition probabilities

$$p_{i,j} = P(L^a_{k+1} = j | L^a_k = i).$$

Clearly $p_{i,j} = 0$ for all $j > i+1$ and $p_{i,j}$ for $j \le i+1$ is equal to the probability that exactly $i+1-j$ customers are served during the interarrival time of the $k+1$th customer. Hence the matrix P of transition probabilities takes the form

$$P = \begin{pmatrix} p_{0,0} & \beta_0 & 0 & \cdots & \\ p_{1,0} & \beta_1 & \beta_0 & 0 & \cdots \\ p_{2,0} & \beta_2 & \beta_1 & \beta_0 & 0 \\ p_{3,0} & \beta_3 & \beta_2 & \beta_1 & \beta_0 \\ \vdots & & & & \ddots \end{pmatrix},$$

where β_i denotes the probability of serving i customers during an interarrival time given that the server remains busy during this interval (thus there are more than i customers present). To calculate β_i we note that given the duration of the interarrival time, t say, the number of customers served during this interval is Poisson distributed with parameter μt. Hence, we have

$$\beta_i = \int_{t=0}^{\infty} \frac{(\mu t)^i}{i!} e^{-\mu t} f_A(t) dt. \tag{8.1}$$

Since the transition probabilities from state j should add up to one, it follows that

$$p_{i,0} = 1 - \sum_{j=0}^{i} \beta_j = \sum_{j=i+1}^{\infty} \beta_j.$$

The transition probability diagram is shown in figure 8.1.

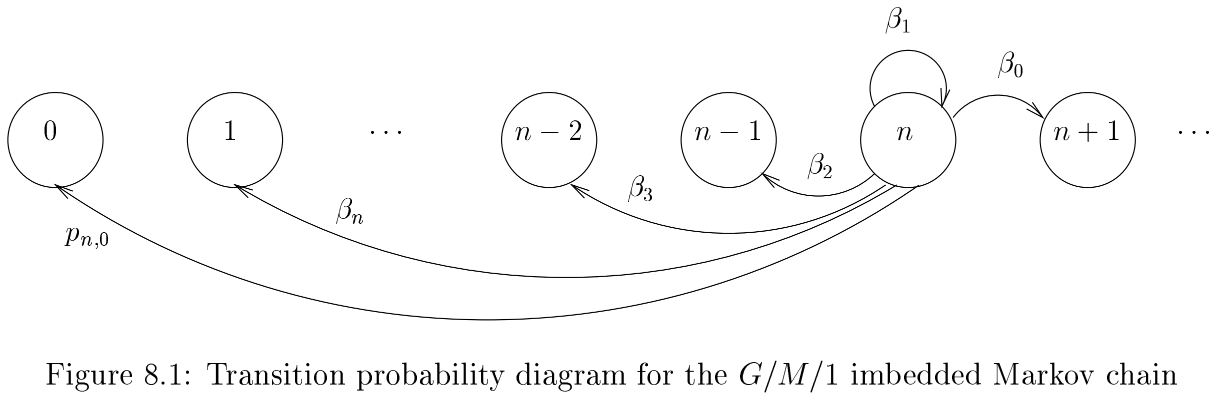

Figure 8.1: Transition probability diagram for the $G/M/1$ imbedded Markov chain

This completes the specification of the imbedded Markov chain. We now wish to determine its limiting distribution $\{a_n\}_{n=0}^{\infty}$. The limiting probabilities a_n satisfy the equilibrium equations

$$
\begin{aligned}
a_0 &= a_0 p_{0,0} + a_1 p_{1,0} + a_2 p_{2,0} + \cdots \\
&= \sum_{i=0}^{\infty} a_i p_{i,0} \qquad (8.2) \\
a_n &= a_{n-1}\beta_0 + a_n\beta_1 + a_{n+1}\beta_2 + \cdots \\
&= \sum_{i=0}^{\infty} a_{n-1+i}\beta_i, \qquad n = 1, 2, \ldots \qquad (8.3)
\end{aligned}
$$

To find the solution of the equilibrium equations it appears that the generating function approach does not work here (verify). Instead we adopt the direct approach by trying to find solutions of the form

$$a_n = \sigma^n, \quad n = 0, 1, 2, \ldots \qquad (8.4)$$

Substitution of this form into equation (8.3) and dividing by the common power σ^{n-1} yields

$$\sigma = \sum_{i=0}^{\infty} \sigma^i \beta_i\,.$$

Of course we know that β_i is given by (8.1). Hence we have

$$
\begin{aligned}
\sigma &= \sum_{i=0}^{\infty} \sigma^i \int_{t=0}^{\infty} \frac{(\mu t)^i}{i!} e^{-\mu t} f_A(t)dt \\
&= \int_{t=0}^{\infty} e^{-(\mu-\mu\sigma)t} f_A(t)dt.
\end{aligned}
$$

The last integral can be recognised as the Laplace-Stieltjes transform of the interarrival time. Thus we arrive at the following equation

$$\sigma = \widetilde{A}(\mu - \mu\sigma). \qquad (8.5)$$

We immediately see that $\sigma = 1$ is a root of equation (8.5), since $\widetilde{A}(0) = 1$. But this root is not useful, because we must be able to normalize the solution of the equilibrium equations. It can be shown that (see exercise 8.4.1) as long as $\rho < 1$ equation (8.5) has a unique root σ in the range $0 < \sigma < 1$, and this is the root which we seek. Note that the remaining equilibrium equation (8.2) is also satisfied by (8.4) since the equilibrium equations are dependent. We finally have to normalize solution (8.4) yielding

$$a_n = (1 - \sigma)\sigma^n\,, \quad n = 0, 1, 2, \ldots \qquad (8.6)$$

Thus we can conclude that the queue length distribution found just before an arriving customer is *geometric* with parameter σ, where σ is the unique root of equation (8.5) in the interval $(0, 1)$.

Example 8.1.1 *(M/M/1)*
For exponentially distributed interarrival times we have

$$\widetilde{A}(s) = \frac{\lambda}{\lambda + s}.$$

Hence equation (8.5) reduces to

$$\sigma = \frac{\lambda}{\lambda + \mu - \mu\sigma},$$

so

$$\sigma(\lambda + \mu - \mu\sigma) - \lambda = (\sigma - 1)(\lambda - \mu\sigma) = 0.$$

Thus the desired root is $\sigma = \rho$ and the arrival distribution is given by

$$a_n = (1 - \rho)\rho^n, \quad n = 0, 1, 2, \ldots$$

Note that this distribution is exactly the same as the equilibrium distribution of the $M/M/1$; see (4.6). This is of course no surprise, because here we have Poisson arrivals.

Example 8.1.2 *($E_2/M/1$)*
Suppose that the interarrival times are Erlang-2 distributed with mean 2/3, so

$$\widetilde{A}(s) = \left(\frac{3}{3 + s}\right)^2.$$

Further assume that $\mu = 4$ (so $\rho = 3/2 \cdot 1/4 = 3/8 < 1$). Then equation (8.5) reduces to

$$\sigma = \left(\frac{3}{7 - 4\sigma}\right)^2.$$

Thus

$$\sigma(7 - 4\sigma)^2 - 9 = (\sigma - 1)(4\sigma - 9)(4\sigma - 1) = 0.$$

Hence the desired root is $\sigma = 1/4$ and

$$a_n = \frac{3}{4}\left(\frac{1}{4}\right)^n, \quad n = 0, 1, 2, \ldots$$

Example 8.1.3
Suppose that the interarrival time consist of two exponential phases, the first phase with parameter μ and the second one with parameter 2μ (so it is slightly more complicated than Erlang-2 where both phases have the same parameter), where μ is also the parameter of the exponential service time. The Laplace-Stieltjes transform of the interarrival time is given by

$$\widetilde{A}(s) = \frac{2\mu^2}{(\mu + s)(2\mu + s)}.$$

For this transform equation (8.5) reduces to

$$\sigma = \frac{2\mu^2}{(2\mu - \mu\sigma)(3\mu - \mu\sigma)} = \frac{2}{(2-\sigma)(3-\sigma)}.$$

This leads directly to

$$\sigma^3 - 5\sigma^2 + 6\sigma - 2 = (\sigma - 1)(\sigma - 2 - \sqrt{2})(\sigma - 2 + \sqrt{2}) = 0.$$

Clearly only the root $\sigma = 2 - \sqrt{2}$ is acceptable. Therefore we have

$$a_n = (\sqrt{2} - 1)(2 - \sqrt{2})^n, \quad n = 0, 1, 2, \ldots$$

8.2 Distribution of the sojourn time

Since the arrival distribution is geometric, it is easy to determine the distribution of the sojourn time. In fact, the analysis is similar to the one for for the $M/M/1$ queue (see section 4.4). With probability a_n an arriving customer finds n customers in the system. Then his sojourn time is the sum of $n + 1$ exponentially distributed service times, each with mean $1/\mu$. Hence,

$$\begin{aligned}
\widetilde{S}(s) &= E(e^{-sS}) \\
&= \sum_{n=0}^{\infty} a_n \left(\frac{\mu}{\mu + s}\right)^{n+1} \\
&= \sum_{n=0}^{\infty} (1-\sigma)\sigma^n \left(\frac{\mu}{\mu + s}\right)^{n+1} \\
&= \frac{\mu(1-\sigma)}{\mu + s} \sum_{n=0}^{\infty} \left(\frac{\mu\sigma}{\mu + s}\right)^{n} \\
&= \frac{\mu(1-\sigma)}{\mu(1-\sigma) + s}.
\end{aligned}$$

From this we can conclude that the sojourn time S is exponentially distributed with parameter $\mu(1 - \sigma)$, i.e.,

$$P(S \leq t) = 1 - e^{-\mu(1-\sigma)t}, \quad t \geq 0.$$

Clearly the sojourn time distribution for the $G/M/1$ is of the *same form* as for the $M/M/1$, the only difference being that ρ is replaced by σ.

Along the same lines it can be shown that the distributon of the waiting time W is given by (cf. (4.14))

$$P(W \leq t) = 1 - \sigma e^{-\mu(1-\sigma)t}, \quad t \geq 0.$$

Note that the probability that a customer does not have to wait is given by $1 - \sigma$ (and *not* by $1 - \rho$).

8.3 Mean sojourn time

It is tempting to determine the mean sojourn time directly by the mean value approach. For an arriving customer we have

$$E(S) = E(L^a)\frac{1}{\mu} + \frac{1}{\mu}, \tag{8.7}$$

where the random variable L^a denotes the number of customers in the system found on arrival. According to Little's law it holds that

$$E(L) = \lambda E(S). \tag{8.8}$$

Unfortunately, we do not have Poisson arrivals, so

$$E(L^a) \neq E(L).$$

Hence the mean value approach does not work here, since we end up with only two equations for three unknowns. Additional information is needed in the form of (8.6), yielding

$$E(L^a) = \sum_{n=0}^{\infty} n a_n = \sum_{n=0}^{\infty} n(1-\sigma)\sigma^n = \frac{\sigma}{1-\sigma}.$$

Then it follows from (8.7) and (8.8) that

$$E(S) = \frac{\sigma}{(1-\sigma)\mu} + \frac{1}{\mu} = \frac{1}{(1-\sigma)\mu}, \qquad E(L) = \frac{\lambda}{(1-\sigma)\mu} = \frac{\rho}{(1-\sigma)}.$$

8.4 Exercises

Exercise 8.4.1
We want to show that as long as $\rho < 1$ equation (8.5) has a unique root σ in the range $0 < \sigma < 1$. Set

$$f(\sigma) = \widetilde{A}(\mu - \mu\sigma).$$

(i) Prove that $f(\sigma)$ is stricly convex on the interval $[0, 1]$, i.e., its derivative is increasing on this interval.

(ii) Show that $f(0) > 0$ and that $f'(1) > 1$ provided $\rho < 1$.

(iii) Show that (i) and (ii) imply that the equation $\sigma = f(\sigma)$ has exactly one root in the interval $(0, 1)$.

Exercise 8.4.2
In a record shop customers arrive according to a hyperexponential arrival process. The interarrival time is with probability 1/3 exponentially distributed with a mean of 1 minute and with probability 2/3 it is exponentially distributed with a mean of 3 minutes. The service times are exponentially distributed with a mean of 1 minute.

(i) Calculate the distribution of the number of customers found in the record shop by an arriving customer.

(ii) Calculate the mean number of customers in the record shop found on arrival.

(iii) Determine $\widetilde{S}(s)$.

(iv) Determine the mean time a customer spends in the record shop.

(v) Calculate the mean number of customers in the record shop (now at an arbitrary point in time).

Exercise 8.4.3
The distribution of the interarrival time is given by

$$F_A(t) = \frac{13}{24}\left(1 - e^{-3t}\right) + \frac{11}{24}\left(1 - e^{-2t}\right), \quad t \geq 0.$$

The service times are exponentially distributed with a mean of 1/6.

(i) Determine the distribution of the number of customers in the system just before an arrival.

(ii) Determine the distribution of the waiting time.

Exercise 8.4.4
Determine the distribution of the sojourn time in case of exponentially distributed service times with mean 1 and hyperexponentially distributed interarrival times with distribution function

$$F_A(t) = \frac{1}{2}\left(1 - e^{-t/2}\right) + \frac{1}{2}\left(1 - e^{-t/4}\right), \quad t \geq 0.$$

Exercise 8.4.5
Consider a queueing system where the interarrival times are exactly 4 minutes. The service times are exponentially distributed with a mean of 2 minutes.

(i) Compute σ.

(ii) Determine the distribution of the sojourn time.

Exercise 8.4.6
At a small river cars are brought from the left side to the right side of the river by a ferry. On average 15 cars per hour arrive according to a Poisson process. It takes the ferry an exponentially distributed time with a mean of 3 minutes to cross the river and return. The capacity of the ferry is equal to 2 cars. The ferry only takes off when there are two or more cars waiting.

(i) What is the fraction of time that the ferry is on its way between the two river sides?

(ii) Determine the distribution of the number of cars that are waiting for the ferry.

(iii) Determine the mean waiting time of a car.

Chapter 9

Priorities

In this chapter we analyse queueing models with different types of customers, where one or more types of customers have priority over other types. More precisely we consider an $M/G/1$ queue with r types of customers. The type i customers arrive according to a Poisson stream with rate λ_i, $i = 1, \ldots, r$. The service time and residual service of a type i customer is denoted by B_i and R_i, respectively. The type 1 customers have the highest priority, type 2 customers the second highest priority and so on. We consider two kinds of priorities. For the *non-preemptive* priority rule higher priority customers may not interrupt the service time of a lower priority customer, but they have to wait till the service time of the low priority customer has been completed. For the *preemptive-resume* priority rule interruptions are allowed and after the interruption the service time of the lower priority customer resumes at the point where it was interrupted. In the following two sections we show how the mean waiting times can be found for these two kinds of priorities.

9.1 Non-preemptive priority

The mean waiting time of a type i customer is denoted by $E(W_i)$ and $E(L_i^q)$ is the number of type i customers waiting in the queue. Further define $\rho_i = \lambda_i E(B_i)$. For the highest priority customers it holds that

$$E(W_1) = E(L_1^q)E(B_1) + \sum_{j=1}^{r} \rho_j E(R_j).$$

According to Little's law we have

$$E(L_1^q) = \lambda_1 E(W_1)$$

Combining the two equations yields

$$E(W_1) = \frac{\sum_{j=1}^{r} \rho_j E(R_j)}{1 - \rho_1}. \tag{9.1}$$

The determination of the mean waiting time for the lower priority customers is more complicated. Consider type i customers with $i > 1$. The waiting time of a type i customer can be divided in a number of portions. The first portion is the amount of work associated with the customer in service and all customers with the same or higher priority present in the queue upon his arrival. Call this portion X_1. The second portion, say X_2, is the amount of higher priority work arriving during X_1. Subsequently the third portion X_3 is the amount of higher priority work arriving during X_2, and so on. A realization of the waiting time for a type 2 customer is shown in figure 9.1. The increments of the amount of work are the service times of the arriving type 1 customers.

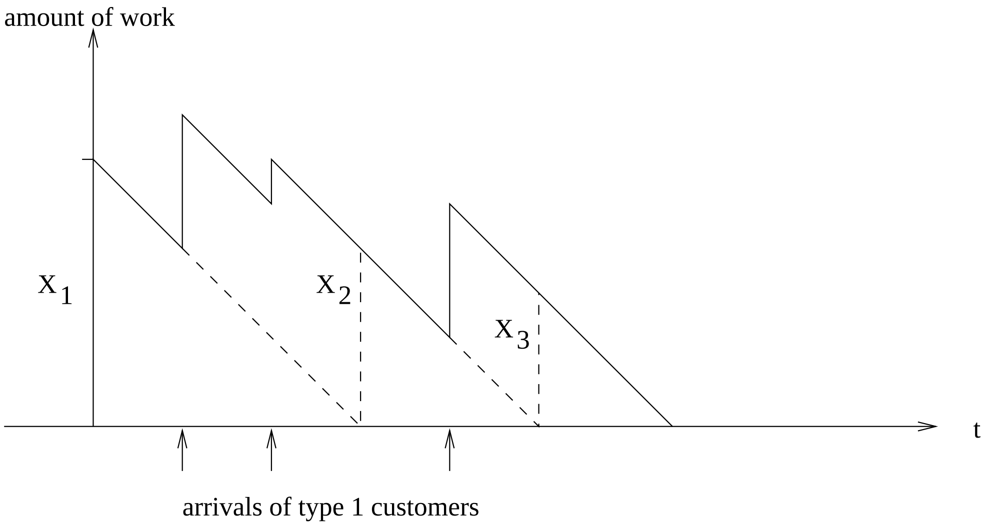

Figure 9.1: Realization of the waiting time of a type 2 customer

Hence the mean waiting time is given by

$$E(W_i) = E(X_1 + X_2 + X_3 + \cdots) = \sum_{k=1}^{\infty} E(X_k).$$

As mentioned above, the first portion of work an arriving type i customer has to wait for is the sum of the service times of all customers with the same or higher priority present in the queue plus the remaining service time of the customer in service. So

$$E(X_1) = \sum_{j=1}^{i} E(L_j^q)E(B_j) + \sum_{j=1}^{r} \rho_j E(R_j).$$

To determine $E(X_{k+1})$ note that X_{k+1} depends on X_k. We therefore condition on the length of X_k. Denote the density of X_k by $f_k(x)$. Then it follows that

$$\begin{aligned} E(X_{k+1}) &= \int_{x=0}^{\infty} E(X_{k+1}|X_k = x) f_k(x)dx \\ &= \int_{x=0}^{\infty} (\lambda_1 x E(B_1) + \cdots + \lambda_{i-1} x E(B_{i-1})) f_k(x)dx \\ &= (\rho_1 + \cdots + \rho_{i-1})E(X_k). \end{aligned}$$

Repeated application of the relation above yields

$$E(X_{k+1}) = (\rho_1 + \cdots + \rho_{i-1})^k E(X_1), \qquad k = 0, 1, 2, \ldots$$

Hence we find for $i = 2, \ldots, r$

$$E(W_i) = \frac{E(X_1)}{1 - (\rho_1 + \cdots + \rho_{i-1})} = \frac{\sum_{j=1}^{i} E(L_j^q)E(B_j) + \sum_{j=1}^{r} \rho_j E(R_j)}{1 - (\rho_1 + \cdots + \rho_{i-1})}, \tag{9.2}$$

An intuitive argument (which can be made rigorous) to directly obtain the above equation is by observing that the waiting time of a type i customer is equal to the first portion of work plus all the higher priority work arriving during his waiting time. So

$$E(W_i) = E(X_1) + \sum_{j=1}^{i-1} \lambda_j E(W_i) E(B_j),$$

from which equation (9.2) immediately follows. Substitution of Little's law

$$E(L_i^q) = \lambda_i E(W_i)$$

into equation (9.2) yields

$$\begin{aligned}(1 - (\rho_1 + \ldots + \rho_i))E(W_i) &= \sum_{j=1}^{i-1} E(L_j^q)E(B_j) + \sum_{j=1}^{r} \rho_j E(R_j) \\ &= (1 - (\rho_1 + \ldots + \rho_{i-2}))E(W_{i-1}).\end{aligned}$$

By multiplying both sides of this equality with $1 - (\rho_1 + \ldots + \rho_{i-1})$ we get the simple recursive relation

$$(1 - \sum_{j=1}^{i} \rho_j)(1 - \sum_{j=1}^{i-1} \rho_j)E(W_i) = (1 - \sum_{j=1}^{i-1} \rho_j)(1 - \sum_{j=1}^{i-2} \rho_j)E(W_{i-1}).$$

Repeatedly applying this relation and using (9.1) finally leads to

$$E(W_i) = \frac{\sum_{j=1}^{r} \rho_j E(R_j)}{(1 - (\rho_1 + \ldots + \rho_i))(1 - (\rho_1 + \ldots + \rho_{i-1}))}, \qquad i = 1, \cdots, r. \tag{9.3}$$

The mean sojourn time $E(S_i)$ of a type i customer follows from $E(S_i) = E(W_i) + E(B_i)$, yielding

$$E(S_i) = \frac{\sum_{j=1}^{r} \rho_j E(R_j)}{(1 - (\rho_1 + \ldots + \rho_i))(1 - (\rho_1 + \ldots + \rho_{i-1}))} + E(B_i),$$

for $i = 1, \cdots, r$.

9.2 Preemptive-resume priority

We will show that the results in case the service times may be interrupted easily follow from the ones in the previous section.

Consider a type i customer. For a type i customer there do not exist lower priority customers due to the preemption rule. So we henceforth assume that $\lambda_{i+1} = \cdots = \lambda_r = 0$.

The waiting time of a type i customer can again be divided into portions $X_1, X_2, \ldots$. Now X_1 is equal to the *total* amount of work in the system upon arrival, since we assumed that there are no lower priority customers. Observe that the total amount of work in the system does not depend on the order in which the customers are served. Hence, at each point in time, it is exactly the same as in the system where the customers are served according to the non-preemptive priority rule. So $X_1, X_2, \ldots$, and thus also W_i have the same distribution as in the system with non-preemptive priorities and, of course, with $\lambda_{i+1} = \cdots = \lambda_r = 0$. From (9.3) we then obtain

$$E(W_i) = \frac{\sum_{j=1}^{i} \rho_j E(R_j)}{(1-(\rho_1+\ldots+\rho_i))(1-(\rho_1+\ldots+\rho_{i-1}))}, \qquad i = 1, \cdots, r.$$

For the mean sojourn time we have to add the service time plus all the interruptions of higher priority customers during the service time. The mean of such a generalized service time can be found along the same lines as (9.2), yielding

$$\frac{E(B_i)}{1-(\rho_1+\cdots+\rho_{i-1})}.$$

So the mean sojourn time of a type i customer is given by

$$E(S_i) = \frac{\sum_{j=1}^{i} \rho_j E(R_j)}{(1-(\rho_1+\ldots+\rho_i))(1-(\rho_1+\ldots+\rho_{i-1}))} + \frac{E(B_i)}{1-(\rho_1+\cdots+\rho_{i-1})},$$

for $i = 1, \cdots, r$.

9.3 Shortest processing time first

In production systems one often processes jobs according to the shortest processing time first rule (SPTF). The mean production lead time in a single machine system operating according to the SPTF rule can be found using the results in section 9.1.

Consider an $M/G/1$ queue with arrival rate λ and service times B with density $f_B(x)$. Assume that $\rho = \lambda E(B) < 1$. The server works according to the SPTF rule. That is, after a service completion, the next customer to be served is the one with the shortest service time.

Define type x customers as the ones with a service time between x and $x + dx$. The mean waiting time of a type x customer is denoted by $E(W(x))$ and $\rho(x)dx$ is the fraction of time the server helps type x customers, so

$$\rho(x)dx = (\lambda f_B(x)dx)x = \lambda x f_B(x)dx. \qquad (9.4)$$

From (9.3) and by observing that the numerator in (9.3) corresponds to the mean amount of work at the server, which in the present situation is simply given by $\rho E(R)$, we obtain

$$\begin{aligned} E(W(x)) &= \frac{\rho E(R)}{(1 - \int_{y=0}^{y=x} \rho(y)dy)^2} \\ &= \frac{\rho E(R)}{(1 - \lambda \int_{y=0}^{y=x} y f_B(y)dy)^2} . \end{aligned}$$

Hence the mean overall waiting time is given by

$$\begin{aligned} E(W) &= \int_{x=0}^{\infty} E(W(x)) f_B(x)dx \\ &= \rho E(R) \int_{x=0}^{\infty} \frac{f_B(x)dx}{(1 - \lambda \int_{y=0}^{y=x} y f_B(y)dy)^2} . \end{aligned} \tag{9.5}$$

In table 9.1 we compare the SPTF rule with the usual first come first served (FCFS) rule for an $M/M/1$ system with mean service time 1. The mean waiting time for the SPTF rule is given by

$$E(W) = \rho \int_{x=0}^{\infty} \frac{e^{-x}dx}{(1 - \rho(1 - e^{-x} - xe^{-x}))^2}$$

and for the FCFS rule it satisfies

$$E(W) = \frac{\rho}{1-\rho} .$$

The results in table 9.1 show that considerable reductions in the mean waiting time are possible.

	$E(W)$	
ρ	FCFS	SPTF
0.5	1	0.713
0.8	4	1.883
0.9	9	3.198
0.95	19	5.265

Table 9.1: The mean waiting time for FCFS and SPTF in an $M/M/1$ with $E(B) = 1$

9.4 A conservation law

In this section we consider a single-server queue with r types of customers. The type i customers arrive according to a general arrival stream with rate λ_i, $i = 1, \ldots, r$. The mean service time and mean residual service time of a type i customer is denoted by $E(B_i)$ and $E(R_i)$, respectively. Define $\rho_i = \lambda_i E(B_i)$. We assume that

$$\sum_{i=1}^{r} \lambda_i E(B_i) < 1,$$

so that the server can handle the amount of work offered per unit of time. Customers enter service in an order independent of their service times and they may not be interrupted during their service. So, for example, the customers may be served according to FCFS, random or a non-preemptive priority rule. Below we derive a conservation law for the mean waiting times of the r types of customers, which expresses that a weighted sum of these mean waiting times is independent of the service discipline. This implies that an improvement in the mean waiting of one customer type owing to a service discipline will always degrade the mean waiting time of another customer type.

Let $E(V(P))$ and $E(L_i^q(P))$ denote the mean amount of work in the system and the mean number of type i customers waiting in the queue, respectively, for discipline P. The mean amount of work in the system is given by

$$E(V(P)) = \sum_{i=1}^{r} E(L_i^q(P))E(B_i) + \sum_{i=1}^{r} \rho_i E(R_i). \tag{9.6}$$

The first sum at the right-hand side is the mean amount of work in the queue, and the second one is the mean amount of work at the server. Clearly the latter does not depend on the discipline P.

The crucial observation is that the amount of work in the system does not depend on the order in which the customers are served. The amount of work decreases with one unit per unit of time independent of the customer being served and when a new customer arrives the amount of work is increased by the service time of the new customer. Hence, the amount of work does not depend on P. Thus from equation (9.6) and Little's law

$$E(L_i^q) = \lambda_i E(W_i(P)),$$

we obtain the following conservation law for the mean waiting times,

$$\sum_{i=1}^{r} \rho_i E(W_i(P)) = \text{constant with respect to service discipline } P.$$

Below we present two examples where this law is used.

Example 9.4.1 *($M/G//1$ with FCFS and SPTF)*
The SPTF rule selects customers in a way that is dependent of their service times. Nevertheless, the law above also applies to this rule. The reason is that the SPTF rule can be translated into a non-preemptive rule as explained in section 9.3. Below we check whether for the $M/G/1$ the weighted sum of the mean waiting times for SPTF is indeed the same as for FCFS.

In case the customers are served in order of arrival it holds that (see (7.14))

$$\rho E(W) = \frac{\rho^2 E(R)}{1-\rho}\,.$$

When the server works according to the SPTF rule we have (see (9.4) and (9.5))

$$\begin{aligned}
\int_{x=0}^{\infty} E(W(x))\rho(x)dx &= \int_{x=0}^{\infty} \frac{\rho E(R)\lambda x f_B(x)dx}{(1-\lambda\int_{y=0}^{y=x} y f_B(y)dy)^2} \\
&= \left.\frac{\rho E(R)}{1-\lambda\int_{y=0}^{y=x} y f_B(y)dy}\right|_{x=0}^{\infty} \\
&= \frac{\rho^2 E(R)}{1-\rho}\,,
\end{aligned}$$

which indeed is the same as for the FCFS rule.

Example 9.4.2 *($M/G/1$ with non-preemptive priority)*
Consider an $M/G/1$ queue with two types of customers. The type 1 customers have non-preemptive priority over the type 2 customers. The mean waiting time for the type 1 customers is given by (9.1). We now derive the mean waiting time for the type 2 customers by using the conservation law. According to this law it holds that

$$\rho_1 E(W_1) + \rho_2 E(W_2) = C, \tag{9.7}$$

where C is some constant independent of the service discipline. To determine C consider the FCFS discipline. For FCFS it follows from (7.14) that

$$E(W_1) = E(W_2) = \frac{\rho_1 E(R_1) + \rho_2 E(R_2)}{1-\rho_1-\rho_2}\,.$$

Hence,

$$C = (\rho_1+\rho_2)\,\frac{\rho_1 E(R_1) + \rho_2 E(R_2)}{1-\rho_1-\rho_2}\,. \tag{9.8}$$

By substituting (9.1) and (9.8) into equation (9.7) we retrieve formula (9.3) for the mean waiting time of the type 2 customers under the non-preemptive priority rule.

9.5 Exercises

Exercise 9.5.1
Customers arrive to a single-server queue according to a Poisson process with a rate of 10 customers per hour. Half of the customers has an exponential service time with a mean of 3 minutes. The other half has an exponential service time with a mean of 6 minutes. The first half are called type 1 customers, the second half type 2 customers. Determine in each of the following three cases the mean waiting time:

(i) Service in order of arrival (FCFS).

(ii) Non-preemptive priority for type 1 customers.

(ii) Non-preemptive priority for type 2 customers.

Exercise 9.5.2
Consider an $M/D/1$ queue with three types of customers. The customers arrive according to a Poisson process. Per hour arrive on average 1 type 1 customer, 2 type 2 customers and also 2 type 3 customers. Type 1 customers have preemptive resume priority over type 2 and 3 customers and type 2 customers have preemptive resume priority over type 3 customers. The service time of each customer is 10 minutes.
Calculate the mean sojourn time for each of the three types of customers.

Exercise 9.5.3
Consider a machine where jobs arrive according to a Poisson stream with a rate of 4 jobs per hour. Half of the jobs have a processing time of exactly 10 minutes, a quarter of the jobs have a processing time of exactly 15 minutes and the remaining quarter have a processing time of 20 minutes. The jobs with a processing time of 10 minutes are called type 1 jobs, the ones with a processing time of 15 minutes type 2 jobs and the rest type 3 jobs. The jobs are processed in order of arrival.

(i) Determine the mean sojourn time (waiting time plus processing time) of a type 1, 2 and 3 job and also of an arbitrary job.

One decides to process smaller jobs with priority. So type 1 orders have highest priority, type 2 orders second highest priority and type 3 orders lowest priority.
Answer question (i) for the following two cases:

(ii) Jobs processed at the machine may not be interrupted.

(iii) Type 1 and type 2 jobs may interrupt the processing of a type 3 job. Type 1 jobs may not interrupt the processing of a type 2 job.

Exercise 9.5.4
A machine produces a specific part type. Orders for the production of these parts arrive according to a Poisson stream with a rate of 10 orders per hour. The number of parts that has to be produced for an order is equal to n with probability 0.5^n, $n = 1, 2, \ldots$ The production of one part takes exactly 2 minutes. Orders are processed in order of arrival.

(i) Denote by B the production time in minutes of an arbitrary order. Show that

$$E(B) = 4, \qquad \sigma^2(B) = 8.$$

(ii) Determine the mean production lead time (waiting time plus production time) of an arbitrary order.

The management decides to give orders for the production of 1 part priority over the other orders. But the production of an order may not be interrupted.

(iii) Determine the mean production lead of an order for the production of 1 part, and of an order for the production of at least 2 parts.

(iv) Determine the mean production lead time of an arbitrary order.

Exercise 9.5.5
Consider a machine for mounting electronic components on printed circuit boards. Per hour arrive according to a Poisson process on average 30 printed circuit boards. The time required to mount all components on a printed circuit board is uniformly distributed between 1 and 2 minutes.

(i) Calculate the mean sojourn time of an arbitrary printed circuit board.

One decides to give printed circuit boards with a mounting time between 1 and x ($1 \leq x \leq 2$) minutes non-preemptive priority over printed circuit boards the mounting time of which is greater than x minutes.

(ii) Determine the mean sojourn time of an arbitrary printed circuit board as a function of x.

(iii) Determine the value of x for which the mean sojourn time of an arbitrary printed circuit board is minimized, and calculate for this specific x the relative improvement with respect to the mean sojourn time calculated in (i).

Exercise 9.5.6
A machine produces 2 types of products, type A and type B. Production orders (for one product) arrive according to a Poisson process. For the production of a type A product arrive on average 105 orders per day (8 hours), and for a type B product 135 orders per day. The processing times are exponentially distributed. The mean processing time for a type A order is 1 minute and for a type B order it is 2 minutes. Orders are processed in order of arrival.

(i) Calculate the mean production lead time for a type A product and for a type B product.

(ii) Determine the Laplace-Stieltjes transform of the waiting time.

(iii) Determine the Laplace-Stieltjes transform of the production lead time of a type A product and determine the distribution of the production lead time of a type A product.

For the production lead time of a type A product one uses a norm of 10 minutes. Each time the production lead time of a type A product exceeds this norm, a cost of 100 dollar is charged.

(iv) Calculate the average cost per day.

To reduce the cost one decides to give type A products preemptive resume priority over type B products.

(v) Determine the mean production lead time for a type A product and for a type B product.

(vi) Calculate the average cost per day.

Exercise 9.5.7
A machine mounts electronic components on three different types of printed circuit boards, type A, B and C boards say. Per hour arrive on average 60 type A boards, 18 type B boards and 48 type C boards. The arrival streams are Poisson. The mounting times are exactly 20 seconds for type A, 40 seconds for type B and 30 seconds for type C. The boards are processed in order of arrival.

(i) Calculate for each type of printed circuit board the mean waiting time and also calculate the mean overall waiting time.

Now suppose that the printed circuit boards are processed according to the SPTF rule.

(ii) Calculate for each type of printed circuit board the mean waiting time and also calculate the mean overall waiting time.

Chapter 10

Variations of the $M/G/1$ model

In this chapter we treat some variations of the $M/G/1$ model and we demonstrate that the mean value technique is a powerful technique to evaluate mean performance characteristics in thes models.

10.1 Machine with setup times

Consider a single machine where jobs are being processed in order of arrival and suppose that it is expensive to keep the machine in operation while there are no jobs. Therefore the machine is turned off as soon as the system is empty. When a new job arrives the machine is turned on again, but it takes some setup time till the machine is ready for processing. So turning off the machine leads to longer production leadtimes. But how much longer? This will be investigated for some simple models in the following subsections.

10.1.1 Exponential processing and setup times

Suppose that the jobs arrive according to a Poisson stream with rate λ and that the processing times are exponentially distributed with mean $1/\mu$. For stability we have to require that $\rho = \lambda/\mu < 1$. The setup time of the machine is also exponentially distributed with mean $1/\theta$. We now wish to determine the mean production lead time $E(S)$ and the mean number of jobs in the system. These means can be determined by using the mean value technique.

To derive an equation for the mean production lead time, i.e. *the arrival relation*, we evaluate what is seen by an arriving job. We know that the mean number of jobs in the system found by an arriving job is equal to $E(L)$ and each of them (also the one being processed) has an exponential (residual) processing time with mean $1/\mu$. With probability $1 - \rho$ the machine is not in operation on arrival, in which case the job also has to wait for the (residual) setup phase with mean $1/\theta$. Further the job has to wait for its own processing time. Hence

$$E(S) = (1-\rho)\frac{1}{\theta} + E(L)\frac{1}{\mu} + \frac{1}{\mu}$$

and together with Little's law,

$$E(L) = \lambda E(S)$$

we immediately find

$$E(S) = \frac{1/\mu}{1-\rho} + \frac{1}{\theta}\,.$$

So the mean production lead time is equal to the one in the system where the machine is always on, plus an extra delay caused by turning off the machine when there is no work. In fact, it can be shown that the extra delay is exponentially distributed with mean $1/\theta$.

10.1.2 General processing and setup times

We now consider the model with generally distributed processing times and generally distributed setup times. The arrivals are still Poisson with rate λ. The first and second moment of the processing time are denoted by $E(B)$ and $E(B^2)$ respectively, $E(T)$ and $E(T^2)$ are the first and second moment of the setup time. For stability we require that $\rho = \lambda E(B) < 1$. Below we demonstrate that also in this more general setting the mean value technique can be used to find the mean production lead time. We first determine the mean waiting time. Then the mean production lead time is found afterwards by adding the mean processing time.

The mean waiting time $E(W)$ of a job satisfies

$$\begin{aligned} E(W) &= E(L^q)E(B) + \rho E(R_B) \\ &+ P(\text{Machine is off on arrival})E(T) \\ &+ P(\text{Machine is in setup phase on arrival})E(R_T), \end{aligned} \tag{10.1}$$

where $E(R_B)$ and $E(R_T)$ denote the mean residual processing and residual setup time, so (see (7.15))

$$E(R_B) = \frac{E(B^2)}{2E(B)}\,, \qquad E(R_T) = \frac{E(T^2)}{2E(T)}\,.$$

To find the probability that on arrival the machine is off (i.e. not working *and* not in the setup phase), note that by PASTA, this probability is equal to the fraction of time that the machine is off. Since a period in which the machine is not processing jobs consists of an interarrival time followed by a setup time, we have

$$P(\text{Machine is off on arrival}) = (1-\rho)\frac{1/\lambda}{1/\lambda + E(T)}.$$

Similarly we find

$$P(\text{Machine is in setup phase on arrival}) = (1-\rho)\frac{E(T)}{1/\lambda + E(T)}.$$

Substituting these relations into (10.1) and using Little's law stating that

$$E(L^q) = \lambda E(W)$$

we finally obtain that

$$E(W) = \frac{\rho E(R_B)}{1-\rho} + \frac{1/\lambda}{1/\lambda + E(T)}E(T) + \frac{E(T)}{1/\lambda + E(T)}E(R_T).$$

Note that the first term at the right-hand side is equal to the mean waiting time in the $M/G/1$ without setup times, the other terms express the extra delay due to the setup times. Finally, the mean production lead time $E(S)$ follows by simply adding the mean processing time $E(B)$ to $E(W)$.

10.1.3 Threshold setup policy

A natural extension to the setup policy in the previous section is the one in which the machine is switched on when the number of jobs in the system reaches some threshold value, N say. This situation can be analyzed along the same lines as in the previous section. The mean waiting time now satisfies

$$\begin{aligned} E(W) &= E(L^q)E(B) + \rho E(R_B) \\ &+ \sum_{i=1}^{N} P(\text{Arriving job is number } i)\left(\frac{N-i}{\lambda} + E(T)\right) \\ &+ P(\text{Machine is in setup phase on arrival})E(R_T), \end{aligned} \tag{10.2}$$

The probability that an arriving job is the i-th one in a cycle can be determined as follows. A typical production cycle is displayed in figure 10.1.

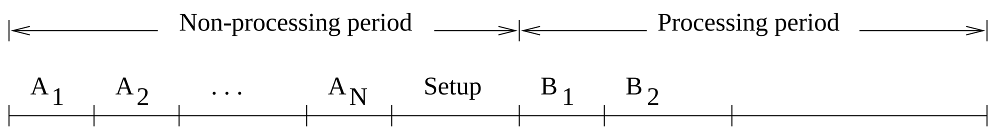

Figure 10.1: Production cycle in case the machine is switched on when there are N jobs

The probability that a job arrives in a non-processing period is equal to $1-\rho$. Such a period now consists of N interarrival times followed by a setup time. Hence, the probability that a job is the i-th one, given that the job arrives in a non-processing period, is equal to $1/\lambda$ divided by the mean length of a non-processing period. So

$$P(\text{Arriving job is number } i) = (1-\rho)\frac{1/\lambda}{N/\lambda + E(T)}, \qquad i = 1, \ldots, N,$$

and similarly,

$$P(\text{Machine is in setup phase on arrival}) = (1-\rho)\frac{E(T)}{N/\lambda + E(T)}.$$

Substituting these relations into (10.2) we obtain, together with Little's law, that

$$E(W) = \frac{\rho E(R_B)}{1-\rho} + \frac{N/\lambda}{N/\lambda + E(T)}\left(\frac{N-1}{2\lambda} + E(T)\right) + \frac{E(T)}{1/\lambda + E(T)}E(R_T).$$

10.2 Unreliable machine

In this section we consider an unreliable machine processing jobs. The machine can break down at any time it is operational, even though it is not processing jobs. What is the impact of these breakdowns on the production leadtimes? To obtain some insight in the effects of breakdowns we formulate and study some simple models in the following subsections.

10.2.1 Exponential processing and down times

Suppose that jobs arrive according to a Poisson stream with rate λ. The processing times are exponentially distributed with mean $1/\mu$. The machine is successively up and down. The up and down times of the machine are also exponentially distributed with means $1/\eta$ and $1/\theta$, respectively.

We begin with formulating the condition under which the machine can handle the amount of work offered per unit time. Let ρ_U and ρ_D denote the fraction of time the machine is up and down, respectively. So

$$\rho_U = \frac{1/\eta}{1/\eta + 1/\theta} = \frac{1}{1+\eta/\theta}, \qquad \rho_D = 1 - \rho_U = \frac{1}{1+\theta/\eta}.$$

Then we have to require that

$$\frac{\lambda}{\mu} < \rho_U. \tag{10.3}$$

We now proceed to derive an equation for the mean production lead time. An arriving job finds on average $E(L)$ jobs in the system and each of them has an exponential processing time with mean $1/\mu$. So in case of a perfect machine his mean sojourn time is equal to $(E(L)+1)/\mu$. However, it is not perfect. Breakdowns occur according to a Poisson process with rate η. So the mean number of breakdowns experienced by our job is equal to $\eta(E(L)+1)/\mu$, and the mean duration of each breakdown is $1/\theta$. Finally, with probability ρ_D the machine is already down on arrival, in which case our job has an extra mean delay of $1/\theta$. Summarizing we have

$$E(S) \;=\; (E(L)+1)\frac{1}{\mu} + \eta(E(L)+1)\frac{1}{\mu}\cdot\frac{1}{\theta} + \rho_D\frac{1}{\theta}$$

$$
\begin{aligned}
&= (1+\frac{\eta}{\theta})(E(L)+1)\frac{1}{\mu} + \frac{\rho_D}{\theta} \\
&= (E(L)+1)\frac{1}{\mu\rho_U} + \frac{\rho_D}{\theta}.
\end{aligned}
$$

Then, with Little's law stating that

$$E(L) = \lambda E(S),$$

we immediately obtain

$$E(S) = \frac{1/(\mu\rho_U) + \rho_D/\theta}{1 - \lambda/(\mu\rho_U)}.$$

In table 10.1 we investigate the impact of the variability in the availability of the machine on the mean production leadtime. The mean production time is 1 hour ($\mu = 1$). The average number of jobs that arrive in a week (40 hours) is 32, so $\lambda = 0.8$ jobs per hour. The fraction of time the machine is available is kept constant at 90%, so $\rho_U = 0.9$ (and thus $\eta/\theta = 1/9$). The rate η at which the machine breaks down is varied from (on the average) every 10 minutes till once a week. In the former case the mean down time is 1.1 minute, in the latter case it is more dramatic, namely nearly half a day (4.4 hours). The results show that the variation in the availability of the machine is essential to the behavior of the system. Note that as η and θ both tend to infinity such that $\eta/\theta = 1/9$, then $E(S)$ tends to 10, which is the mean sojourn time in an $M/M/1$ with arrival rate 0.8 and service rate 0.9.

η	θ	$E(S)$
6	54	10.02
3	27	10.03
1	9	10.1
0.125	1.125	10.8
0.0625	0.5625	11.6
0.025	0.225	14

Table 10.1: The mean production leadtime in hours as a function of the break-down rate η per hour for fixed availability of 90%

10.2.2 General processing and down times

We now consider the model with general processing times and general down times. The arrivals are still Poisson with rate λ. The first and second moment of the processing time are denoted by $E(B)$ and $E(B^2)$ respectively. The time between two breakdowns is exponentially distributed with mean $1/\eta$ and $E(D)$ and $E(D^2)$ are the first and second moment of the down time. For stability we require that (cf. (10.3))

$$\lambda E(B) < \frac{1}{1+\eta E(D)} .$$

Below we first determine the mean waiting time by using the mean value technique. The mean production leadtime is determined afterwards.

We start with introducing the generalized processing time, which is defined as the processing time plus the down times occurring in that processing time. Denote the generalized processing time by G. Then we have

$$G = B + \sum_{i=1}^{N(B)} D_i,$$

where $N(B)$ is the number of break-downs during the processing time B and D_i is the i-th down time. For the mean of G we get by conditioning on B and $N(B)$ that

$$\begin{aligned}
E(G) &= \int_{x=0}^{\infty} \sum_{n=0}^{\infty} E(G|B=x, N(B)=n) e^{-\eta x} \frac{(\eta x)^n}{n!} f_B(x) dx \\
&= \int_{x=0}^{\infty} \sum_{n=0}^{\infty} (x + nE(D)) e^{-\eta x} \frac{(\eta x)^n}{n!} f_B(x) dx \\
&= \int_{x=0}^{\infty} \sum_{n=0}^{\infty} (x + x\eta E(D)) e^{-\eta x} \frac{(\eta x)^n}{n!} f_B(x) dx \\
&= E(B) + E(B)\eta E(D),
\end{aligned}$$

and similarly,

$$E(G^2) = E(B^2)(1+\eta E(D))^2 + E(B)\eta E(D^2).$$

A typical production cycle is shown in figure 10.2. The non-processing period consists of cycles which start with an up time. When the machine is up two things can happen: a job arrives, in which case the machine starts to work, or the machine goes down and has to be repaired. Hence, an up time is exponentially distributed with mean $1/(\lambda+\eta)$ and it is followed by a processing period (i.e. a job arrives) with probability $\lambda/(\lambda+\eta)$ or otherwise, it is followed by a down time. During a processing period the machine works on jobs with generalized processing times (until all jobs are cleared). For the mean waiting time it holds that

$$\begin{aligned}
E(W) &= E(L^q)E(G) + \rho_G E(R_G) \\
&\quad + P[\text{Arrival in a down time in a non-processing period}] E(R_D),
\end{aligned} \tag{10.4}$$

where ρ_G is the fraction of time the machine works on generalized jobs and $E(R_G)$ and $E(R_D)$ denote the mean residual generalized processing time and the mean residual down time, so

$$\rho_G = \lambda E(G), \qquad E(R_G) = \frac{E(G^2)}{2E(G)}, E(R_D) = \frac{E(D^2)}{2E(D)} .$$

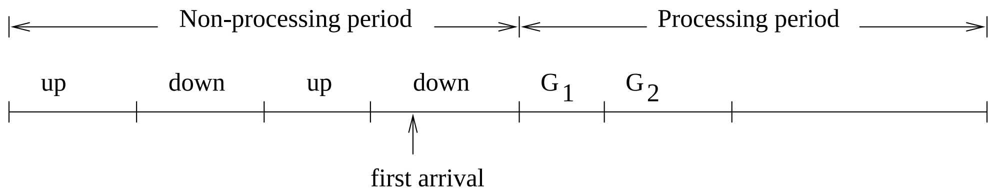

Figure 10.2: Production cycle of an unreliable machine

The probability that a job arriving in a non-processing period finds that the machine is down, is given by

$$\frac{E(D)\eta/(\lambda+\eta)}{1/(\lambda+\eta)+E(D)\eta/(\lambda+\eta)} = \frac{E(D)\eta}{1+E(D)\eta},$$

Hence, since a job arrives with probability $1-\rho_G$ in a non-processing period, we have

$$P(\text{Arrival in a down time in a non-processing period}) = (1-\rho_G)\frac{E(D)\eta}{1+E(D)\eta}.$$

Substitution of this relation into (10.4) and using Little's law yields

$$E(W) = \frac{\rho_G E(R_G)}{1-\rho_G} + \frac{E(D)\eta}{1+E(D)\eta}E(R_D),$$

and the mean production lead time finally follows from

$$E(S) = E(W) + E(G).$$

10.3 $M/G/1$ queue with an exceptional first customer in a busy period

A servers' life in an $M/G/1$ queue is an alternating sequence of periods during which no work is done, the so-called idle periods, and periods during which the server helps customers, the so-called busy periods. In some applications the service time of the first customer in a busy period is different from the service times of the other customers served in the busy period. For example, consider the setup problem in section 10.1 again. This problem may, alternatively, be formulated as an $M/G/1$ queue in which the first job in a busy period has an exceptional service time, namely the setup time plus his actual processing time. The problem in section 10.2 can also be described in this way. Here we can take the generalized processing times as the service times. However, on arrival of the first job in a busy period the machine can be down. Then the (generalized) servicing of that job cannot immediately start, but one has to wait till the machine is repaired. This repair time can be included in the service time of the first job. In this case, however, the determination of the distribution, or the moments of the service time of the first job is more difficult then in the setup problem. Below we show how, in the general setting of an

$M/G/1$ with an exceptional first customer, the mean sojourn time can be found by using the mean-value approach.

Let B_f and R_f denote the service time and residual service time, respectively, of the first customers in a busy period. For the mean waiting time we have

$$E(W) = E(L^q)E(B) + \rho_f E(R_f) + \rho E(R), \tag{10.5}$$

where ρ_f is the fraction of time the server works on first customers, and ρ is the fraction of time the server works on other (ordinary) customers. Together with Little's law we then obtain from (10.5) that

$$E(W) = \frac{\rho_f E(R_f) + \rho E(R)}{1 - \lambda E(B)},$$

and for the mean sojourn time we get

$$E(S) = E(W) + (1 - \rho_f - \rho)E(B_f) + (\rho_f + \rho)E(B).$$

It remains to determine ρ_f and ρ. The probability that an arriving customer is the first one in a busy cycle is given by $1 - \rho_f - \rho$. Hence, the number of first customers arriving per unit of time is equal to $\lambda(1 - \rho_f - \rho)$. Thus ρ_f and ρ satisfy

$$\begin{aligned} \rho_f &= \lambda(1 - \rho_f - \rho)E(B_f), \\ \rho &= \lambda(\rho_f + \rho)E(B), \end{aligned}$$

from which it follows that

$$\rho_f = \frac{\lambda E(B_f)(1 - \lambda E(B))}{1 + \lambda E(B_f) - \lambda E(B)}, \qquad \rho = \frac{\lambda E(B_f)\lambda E(B))}{1 + \lambda E(B_f) - \lambda E(B)}.$$

10.4 $M/G/1$ queue with group arrivals

In this section we consider the $M/G/1$ queue where customers do not arrive one by one, but in groups. These groups arrive according to a Poisson process with rate λ. The group size is denoted by the random variable G with probability distribution

$$g_k = P(G = k), \qquad k = 0, 1, 2, \ldots$$

Note that we also admit zero-size groups to arrive. Our interest lies in the mean waiting time of a customer, for which we can write down the following equation.

$$E(W) = E(L^q)E(B) + \rho E(R) + \sum_{k=1}^{\infty} r_k(k-1)E(B)\,, \tag{10.6}$$

where ρ is the server utilization, so

$$\rho = \lambda E(G)E(B),$$

and r_k is the probability that our customer is the kth customer served in his group. The first two terms at the right-hand side of (10.6) correspond to the mean waiting time of the whole group. The last one indicates the mean waiting time due to the servicing of members in his own group.

To find r_k we first determine the probability h_n that our customer is a member of a group of size n (cf. section 7.7). Since it is more likely that our customer belongs to a large group than to a small one, it follows that h_n is proportional to the group size n as well as the frequency of such groups. Thus we can write

$$h_n = Cng_n,$$

where C is a constant to normalize this distribution. So

$$C^{-1} = \sum_{n=1}^{\infty} ng_n = E(G).$$

Hence

$$h_n = \frac{ng_n}{E(G)}, \qquad n = 1, 2, \ldots$$

Given that our customer is a member of a group of size n, he will be with probability $1/n$ the kth customer in his group going into service (of course, $n \geq k$). So we obtain

$$r_k = \sum_{n=k}^{\infty} h_n \cdot \frac{1}{n} = \frac{1}{E(G)} \sum_{n=k}^{\infty} g_n,$$

and for the last term in (10.6) it immediately follows that

$$\begin{aligned}
\sum_{k=1}^{\infty} r_k(k-1)E(B) &= \frac{1}{E(G)} \sum_{k=1}^{\infty} \sum_{n=k}^{\infty} g_n(k-1)E(B) \\
&= \frac{1}{E(G)} \sum_{n=1}^{\infty} \sum_{k=1}^{n} g_n(k-1)E(B) \\
&= \frac{1}{E(G)} \sum_{n=1}^{\infty} \frac{1}{2} n(n-1) g_n E(B) \\
&= \frac{E(G^2) - E(G)}{2E(G)} E(B). \qquad (10.7)
\end{aligned}$$

From (10.6) and (10.7) and Little's law stating that

$$E(L^q) = \lambda E(G)E(W),$$

we finally obtain

$$E(W) = \frac{\rho E(R)}{1-\rho} + \frac{(E(G^2) - E(G))E(B)}{2E(G)(1-\rho)}.$$

The first term at the right-hand side is equal to the mean waiting time in the system where customers arrive one by one according to a Poisson process with rate $\lambda E(G)$. Clearly, the second term indicates the extra mean delay due to clustering of arrivals.

Example 10.4.1 *(Uniform group sizes)*
In case the group size is uniformly distributed over $1, 2, \ldots, n$, so

$$g_k = \frac{1}{n}, \qquad k = 1, \ldots, n,$$

we find

$$E(G) = \sum_{k=1}^{n} \frac{k}{n} = \frac{n+1}{2}, \qquad E(G^2 - G) = \sum_{k=1}^{n} \frac{k^2 - k}{n} = \frac{(n-1)(n+1)}{3}.$$

Hence

$$E(W) = \frac{\rho E(R)}{1-\rho} + \frac{(n-1)E(B)}{3(1-\rho)}.$$

10.5 Exercises

Exercise 10.5.1
Consider an $M/G/1$ queue where at the end of each busy period the server leaves for a fixed period of T units of time. Determine the mean sojourn time $E(S)$.

Exercise 10.5.2
In a factory a paternoster elevator (see figure 10.3) is used to transport rectangular bins from the ceiling to the floor. This is a chain-elevator with product carriers (each carrier can hold at most one bin) at a certain fixed distance. The number of carriers between the ceiling and the floor is 4. Every time a carrier reaches the ceiling (and, simultaneously, another one reaches the floor) the elevator stops 2 seconds. This is exactly the time a bin needs to enter or leave the carrier. Subsequently the elevator starts to move again till the next carrier reaches the ceiling. This takes exactly 3 seconds. Then it waits for 2 seconds, moves again, and so on. The process of waiting and moving goes on continuously. Note that a bin, finding on arrival an empty carrier positioned in front of it, cannot enter the carrier, because the residual time till the carrier start to move again is less than 2 seconds. The bin has to wait till the next carrier arrives.

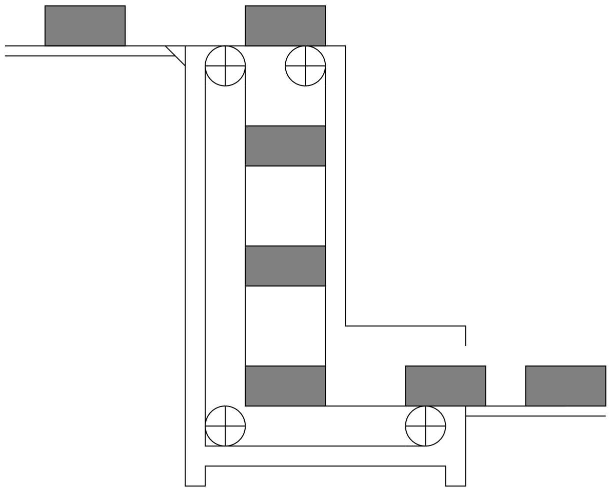

Figure 10.3: Paternoster elevator

Bins arrive according to a Poisson process with a rate of 10 bins per minute.

(i) Determine the mean waiting time of a bin in front of the elevator.

(ii) Determine the mean time that elapses from the arrival of a bin till the moment the bin leaves the elevator downstairs.

Exercise 10.5.3
Consider a machine which is turned off when there is no work. It is turned on and restarts work when enough orders, say N, arrived to the machine. The setup times are negligible.

The processing times are exponentially distributed with mean $1/\mu$ and the average number of orders arriving per unit time is $\lambda(< \mu)$.

(i) Determine the mean number of orders in the system.

(ii) Determine the mean production lead time.

Suppose that $\lambda = 20$ orders per hour, $1/\mu = 2$ minutes and that the setup cost is 54 dollar. In operation the machine costs 12 dollar per minute. The waiting cost is 1 dollar per minute per order.

(iii) Determine, for given threshold N, the average cost per unit time.

(iv) Compute the threshold N minimizing the average cost per unit time.

Exercise 10.5.4
At a small river pedestrians are brought from the left side to the right side of the river by a ferry. On average 40 pedestrians per hour arrive according to a Poisson process. It takes the ferry exactly 2 minutes to cross the river and return. The capacity of the ferry is sufficiently big, so it is always possible to take all waiting pedestrians to the other side of the river. The ferry travels continuously back and forth, also when there are no waiting pedestrians.

(i) Determine the mean waiting time and the mean number of pedestrians waiting for the ferry.

The ferry now only takes off when there are one or more pedestrians. In case there are no pedestrians the ferry waits for the first one to arrive, af ter which it immediately takes off (the ferry never waits at the other side of the river).

(ii) Determine the probability that an arriving pedestrian finds the ferry waiting to take off.

(iii) Determine the mean waiting time and the mean number of pedestrians waiting for the ferry.

Exercise 10.5.5
A machine produces products in two phases. The first phase is standard and the same for all products. The second phase is customer specific (the finishing touch). The first (resp. second) phase takes an exponential time with a mean of 10 (resp. 2) minutes. Orders for the production of one product arrive according to a Poisson stream with a rate of 3 orders per hour. Orders are processed in order of arrival.

(i) Determine the mean production lead time (waiting time plus production time) of an order.

The machine is switched off when the system is empty and it is switched on again as soon as the first order arrives. A fixed cost of 20$ is incurred each time the machine is switched on (the time needed to switch the machine on or off is negligible).

(ii) Determine the average switch-on cost per hour.

To reduce the production lead time one decides to start already with the production of phase 1 when the system is empty. If upon completion of phase 1 no order has arrived yet, the production stops and the machine is switched off. When the first order arrives the machine is switched on again and can directly start with phase 2.

(iii) Determine the reduction in the mean production lead time.

(iv) Determine the average switch-on cost per hour.

Exercise 10.5.6
In a bank there is a special office for insurances. Here arrive according to a Poisson process on average 8 customers per hour. The service times are exponentially distributed with a mean of 5 minutes. As soon as all customers are served and the office is empty again, the clerck also leaves to get some coffee. He returns after an exponential time with a mean of 5 minutes and then starts servicing the waiting customers (if there are any) or patiently waits for the first customer to arrive.

(i) What is the mean sojourn time and the mean number of customers in the office?

Now suppose that when the clerck finds an empty office upon his return, he immediately leaves again to get another cup of coffee. This takes again an exponential time with a mean of 5 minutes. The clerck repeats to leave until he finds a waiting customer upon return.

(ii) How many cups of coffee the clerck drinks on average before he starts servicing again?

(iii) Determine the mean sojourn time and the mean number of customers in the office.

Exercise 10.5.7
Consider a machine processing orders. These orders arrive according to a Poisson process with a rate of 1 order per hour. When there are no orders, the machine is switched off. As soon as a new order arrives, the machine is switched on again, which takes exactly T hours. The processing time of an order is exponentially distributed with a mean of 30 minutes.

(i) Determine as a function of T:

  (a) The mean numbers of orders processed in a production cycle.
  (b) The mean duration of a production cycle.
  (c) The mean production lead time of an order.

Suppose that it costs 17 dollar each time the machine is switched on again and that the waiting cost per hour per order are 1 dollar.

(ii) Show that the average cost per hour are minimal for $T = 3$ hour.

Exercise 10.5.8
Consider a queueing system where on average 3 groups of customers arrive per hour. The mean group size is 10 customers. The service time is exactly 1 minute for each customer. Determine the mean sojourn time of the first customer in a group, the last customer in a group and an arbitrary one in the following two cases:

(i) the group size is Poisson distributed;

(ii) the groups size is geometrically distributed.

Exercise 10.5.9
Customers arrive in groups at a server. The groups consist of 1 or 3 customers, both with equal probability, and they arrive according to a Poisson stream with a rate of 2 groups per hour. Customers are served one by one, and they require an exponential service time with a mean of 10 minutes.

(i) Determine the mean sojourn time of an arbitrary customer.

(ii) Determine the mean number of customers in the system.

Exercise 10.5.10
Passengers are brought with small vans from the airport to hotels nearby. At one of those hotels on average 6 vans per hour arrive according to a Poisson process. With probability 1/4 a van brings 2 guests for the hotel, with probability 1/4 only one guest and with probability 1/2 no guests at all. At the reception of the hotel there is always one receptionist present. It takes an exponential time with a mean of 5 minutes to check in a guest.

(i) Determine the distribution of the number of guests at the reception.

(ii) Determine the mean waiting time of an arbitrary guest at the reception.

Exercise 10.5.11
Customers arrive at a server according to a Poisson stream with a rate of 6 customers per hour. As soon as there are no customers the server leaves to do something else. He returns when there are 2 customers waiting for service again. It takes exactly 5 minutes for him to return. The service time of a customer is uniformly distributed between 5 and 10 minutes.

(i) Determine the mean duration of a busy period, i.e., a period during which the server is servicing customers without interruptions.

(ii) Determine the mean number of customers served in a busy period.

(iii) Determine the mean sojourn time of a customer.

Chapter 11

Insensitive systems

In this chapter we study some queueing systems for which the queue length distribution is *insensitive* to the distribution of the service time, but only depends on its mean.

11.1 $M/G/\infty$ queue

In this model customers arrive according to a Poisson process with rate λ. Their service times are independent and identically distributed with some general distribution function. The number of servers is infinite. So there is always a server available for each arriving customer. Hence, the waiting time of each customer is zero and the sojourn time is equal to the service time. Thus by Little's law we immediately obtain that

$$E(L) = \rho,$$

where $\rho = \lambda E(B)$ denotes the mean amount of work that arrives per unit time. In the remainder of this section we want to also determine the distribution of L, i.e., the probabilities p_n that there are n customers in the system.

Example 11.1.1 *($M/M/\infty$)*
In this model the service times are exponentially distributed with mean $1/\mu$.

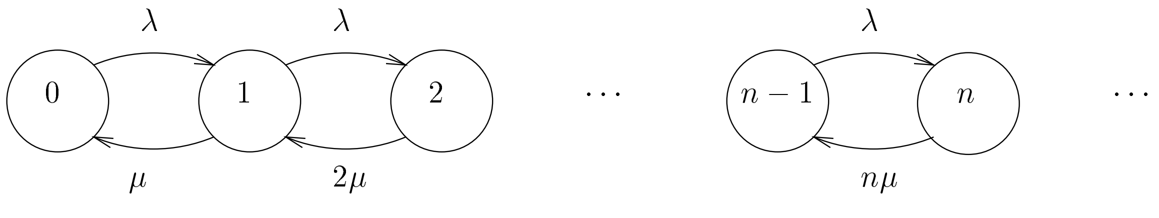

Figure 11.1: Flow diagram for the $M/M/\infty$ model

From figure 11.1 we obtain by equating the flow from state $n-1$ to n and the flow from n to $n-1$ that

$$p_{n-1}\lambda = p_n n\mu.$$

Thus

$$p_n = \frac{\lambda}{n\mu} p_{n-1} = \frac{\rho}{n} p_{n-1} = \frac{\rho^2}{n(n-1)} p_{n-2} = \cdots = \frac{\rho^n}{n!} p_0.$$

Since the probabilities p_n have to add up to one, it follows that

$$p_0^{-1} = \sum_{n=0}^{\infty} \frac{\rho^n}{n!} = e^{\rho}.$$

Summarizing, we have found that

$$p_n = \frac{\rho^n}{n!} e^{-\rho}. \tag{11.1}$$

Thus the number of customers in the system has a Poisson distribution with mean ρ.

Example 11.1.2 *($M/D/\infty$)*
Let b denote the constant service time. The probability $p_n(t)$ that there are exactly n customers in the system at time t is equal to the probability that between time $t-b$ and t exactly n customers arrived. Since the number of customers arriving in a time interval of length b is Poisson distributed with mean λb, we immediately obtain

$$p_n(t) = \frac{\lambda b}{n!} e^{-\lambda b} = \frac{\rho}{n!} e^{-\rho},$$

which is valid for all $t > b$, and thus also for the limiting distribution.

Example 11.1.3 *(Discrete service time distribution)*
Suppose that the service time distribution is discrete, i.e., there are nonnegative numbers b_i and probabilities q_i such that

$$P(B = b_i) = q_i, \qquad i = 1, 2, \ldots$$

In this case it is also easy to determine the probabilities p_n. We split the Poisson stream with rate λ into the countably many independent Poisson streams numbered $1, 2, \ldots$. The intensity of stream i is λq_i and the customers of this stream have a constant service time b_i. The number of servers is infinite and thus we immediately obtain from the previous example that the number of type i customers in the system is Poisson distributed with mean $\lambda q_i b_i$ and this number is of course independent of the other customers in the system. Since the sum of independent Poisson random variables is again Poisson (see exercise 2.7.11) it follows that the total number of customers in the system is Poisson distributed with mean $\lambda q_1 b_1 + \lambda q_2 b_2 + \cdots = \rho$.

The examples above suggest that also in the $M/G/\infty$ the number of customers in the system is Poisson distributed with ρ. In fact, since each distribution function can be approximated arbitrary close by a discrete distribution function, this follows from example 11.1.3 (see also exercise 11.4.1). Summarizing we may conclude that in the $M/G/\infty$ it holds that

$$p_n = \frac{\rho}{n!} e^{-\rho}, \qquad n = 0, 1, 2, \ldots,$$

where $\rho = \lambda E(B)$. Note that this is true regardless of the form of the distribution function $F_B(\cdot)$ of the service time.

11.2 $M/G/c/c$ queue

In this model customers also arrive according to a Poisson process with rate λ. Their service times are independent and identically distributed with some general distribution function. There are c servers available. Each newly arriving customer immediately goes into service if there is a server available, and that customer is lost if all servers are occupied. This system is therefore also referred to as the *$M/G/c$ loss system*.

In this section we want to find the probabilities p_n of n customers in the system. Of special interest is the probability p_c, which, according to the PASTA property, describes the fraction of customers that are lost.

Example 11.2.1 *($M/M/c/c$)*
In this model the service times are exponentially distributed with mean $1/\mu$.

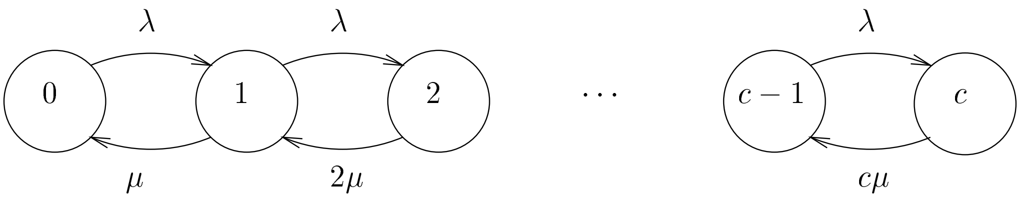

Figure 11.2: Flow diagram for the $M/M/c/c$ model

From figure 11.1 we immediately obtain that

$$p_{n-1}\lambda = p_n n\mu, \qquad n = 1, 2, \ldots, c.$$

Hence, it is readily verified that

$$p_n = \frac{(\lambda/\mu)^n/n!}{\sum_{n=0}^{c}(\lambda/\mu)^n/n!} = \frac{\rho^n/n!}{\sum_{n=0}^{c}\rho^n/n!}, \qquad n = 0, 1, \ldots, c,$$

where $\rho = \lambda/\mu$.

It can be proved that also for a general service time distribution the probabilities p_n are given by (see e.g. [7])

$$p_n = \frac{\rho^n/n!}{\sum_{n=0}^{c}\rho^n/n!}, \qquad n = 0, 1, \ldots, c,$$

where $\rho = \lambda E(B)$. Hence, the so-called *blocking probability* $B(c, \rho)$ is given by

$$B(c, \rho) = p_c = \frac{\rho^c/c!}{\sum_{n=0}^{c}\rho^n/n!}. \tag{11.2}$$

The name given to this formula is *Erlang's loss formula*. Note that by Little's law we obtain that

$$E(L) = \rho(1 - B(c, \rho))$$

Example 11.2.2 *(Parking lot)*
Customers arrive according to a Poisson process at a parking lot near a small shopping center with a rate of 60 cars per hour. The mean parking time is 2.5 hours and the parking lot offers place to 150 cars. When the parking lot is full, an arriving customer has to park his car somewhere else. Now we want to know the fraction of customers finding all places occupied on arrival.

The parking lot can be described by a $M/G/150/150$ model with $\rho = 60 \cdot 2.5 = 150$. Hence the fraction of customers finding all places occupied on arrival is given by

$$B(150, 150) = \frac{150^{150}/150!}{\sum_{n=0}^{150} 150^n/n!}.$$

It will be clear that the computation of $B(150, 150)$ gives rise to a serious problem!

11.3 Stable recursion for $B(c, \rho)$

Fortunately it is easy to derive a simple and numerically stable recursion for the blocking probabilities $B(c, \rho)$. From (11.2) we have

$$B(c, \rho) = \frac{\rho^c/c!}{\sum_{n=0}^{c-1} \rho^n/n! + \rho^c/c!}.$$

Dividing the numerator and denominator of this expression by $\sum_{n=0}^{c-1} \rho^n/n!$ yields

$$B(c, \rho) = \frac{\rho B(c-1, \rho)/c}{1 + \rho B(c-1, \rho)/c} = \frac{\rho B(c-1, \rho)}{c + \rho B(c-1, \rho)}. \tag{11.3}$$

Starting with $B(0, \rho) = 1$ we can use relation (11.3) to subsequently compute the blocking probabilities $B(c, \rho)$ for $c = 1, 2, 3, \ldots$.

c	$B(c, 150)$
150	0.062
155	0.044
160	0.028
165	0.017
170	0.009

Table 11.1: Blocking probability $B(c, \rho)$ for $\rho = 150$ and several values of c

Example 11.3.1 *(Parking lot)*
Using the recursion (11.3) it is easy to compute $B(c, 150)$ for $c = 150$. In table 11.1 we list $B(c, 150)$ for several values of c. Clearly in the present situation with 150 parking places 6% of the arriving customers find all places occupied. With 20 additional places this percentage drops below 1%.

Remark 11.3.2 *(Delay probability in the $M/M/c$ queue)*
It will be clear from (5.1) that the formula for delay probability Π_W in the $M/M/c$ suffers from the same numerical problems as (11.2). Luckely there is a simple relation between the queueing probability Π_W in the $M/M/c$ with server utilization ρ and the blocking probability $B(c, c\rho)$. From (5.1) we immediately obtain

$$\Pi_W = \frac{(c\rho)^c/c!}{(1-\rho)\sum_{n=0}^{c-1}(c\rho)^n/n! + (c\rho)^c/c!} = \frac{\rho B(c-1, c\rho)}{1-\rho+\rho B(c-1, c\rho)}\,.$$

By first computing $B(c-1, c\rho)$ from the recursion (11.3) we can use the relation above to determine Π_W.

11.4 Exercises

Exercise 11.4.1
Prove that the fact that (11.1) holds for discrete service time distributions implies that it also holds for general service time distributions.
(*Hint*: Approximate the service time distribution from below and from above by discrete distributions and then consider the probability that there are n or more customers in the system.)

Exercise 11.4.2
In a small restaurant there arrive according to a Poisson proces on average 5 groups of customers. Each group can be accommodated at one table and stays for an Erlang-2 distributed time with a mean of 36 minutes. Arriving groups who find all tables occupied leave immediately.
How many tables are required such that at most 7% of the arriving groups is lost?

Exercise 11.4.3
For a certain type of article there is a stock of at most 5 articles to satisfy customer demand directly from the shelf. Customers arrive according to a Poisson process at a rate of 2 customers per week. Each customer demands 1 article. The ordering policy is as follows. Each time an article is sold to a customer, an order for 1 article is immediately placed at the supplier. The lead time (the time that elapses from the moment the order is placed until the order arrives) is exponentially distributed with a mean of 1 week. If on arrival of a customer the shelf is empty, the customer demand will be lost. The inventory costs are 20 guilders per article per week. Each time an article is sold, this yields a reward of 100 guilders.

(i) Calculate the probability distribution of the number of outstanding orders.

(ii) Determine the mean number of articles on stock.

(iii) What is the average profit (reward - inventory costs) per week?

Exercise 11.4.4
In our library there are 4 VUBIS terminals. These terminals can be used to obtain information about the available literature. If all terminals are occupied when someone wants information, then that person will not wait but leave immediately (to look for the required information somewhere else). A user session on a VUBIS terminal takes on average 2.5 minutes. Since the number of potential users is large, it is reasonable to assume that users arrive according to a Poisson stream. On average 72 users arrive per hour.

(i) Determine the probability that i terminals are occupied, $i = 0, 1, \ldots, 4$.

(ii) What is the fraction of arriving users finding all terminals occupied?

(iii) How many VUBIS terminals are required such that at most 5% of the arriving users find all terminals occupied?

Exercise 11.4.5
Consider a machine continuously processing parts (there is always raw material available). The processing time of a part is exponentially distributed with a mean of 20 seconds. A finished part is transported immediately to an assembly cell by an automatic conveyor system. The transportation time is exactly 3 minutes.

(i) Determine the mean and variance of the number of parts on the conveyor.

To prevent that too many parts are simultaneously on the conveyor one decides to stop the machine as soon as there are N parts on the conveyor. The machine is turned on again as soon as this number is less than N.

(ii) Determine the throughput of the machine as a function of N.

(iii) Determine the smallest N for which the throughput is at least 100 parts per hour.

Exercise 11.4.6
A small company renting cars has 6 cars available. The costs (depreciation, insurance, maintenance, etc.) are 60 guilders per car per day. Customers arrive according to a Poisson process with a rate of 5 customers per day. A customer rents a car for an exponential time with a mean of 1.5 days. Renting a car costs 110 guilders per day. Arriving customers for which no car is available are lost (they will go to another company).

(i) Determine the fraction of arriving customers for which no car is available.

(ii) Determine the mean profit per day.

The company is considering to buy extra cars.

(iii) How many cars should be bought to maximize the mean profit per day?

Bibliography

[1] M. Abramowitz, I.A. Stegun, *Handbook of mathematical functions*, Dover, 1965.

[2] I.Adan, Y.Zhao, *Analyzing $GI/E_r/1$ queues*, Opns. Res. Lett., 19 (1996), pp. 183–190.

[3] I.J.B.F. Adan, W.A. van de Waarsenburg, J. Wessels, *Analyzing $E_k|E_r|c$ queues*, EJOR, 92 (1996), pp. 112–124.

[4] N.G. de Bruijn, *Asymptotic methods*, Dover, 1981.

[5] B.D. Bunday, *An introduction to queueing theory*, Arnold, London, 1996.

[6] J.A. Buzacott, J.G. Shanthikumar, *Stochastic models of manufacturing systems*, Prentice Hall, Englewood Cliffs, 1993.

[7] J.W. Cohen, *On regenerative processes in queueing theory*, Springer, Berlin, 1976.

[8] J.W. Cohen, *The single server queue*, North-Holland, Amsterdam, 1982.

[9] J.H. Dshalalow (editor), *Advances in Queueing: Theory, Methods and Open Problems*, CRC Press, Boca Raton, 1995.

[10] D. Gross, C.M. Harris, *Fundamentals of queueing theory*, Wiley, Chichester, 1985.

[11] M.C. van der Heijden, *Performance analysis for reliability and inventory models*, Thesis, Vrije Universiteit, Amsterdam, 1993.

[12] D.P. Heyman, M.J. Sobel, *Stochastic models in operations research*, McGraw-Hill, London, 1982.

[13] M.A. Johnson, *An emperical study of queueing approximations based on phase-type approximations*, Stochastic Models, 9 (1993), pp. 531–561.

[14] L. Kleinrock, *Queueing Systems, Vol. I: Theory.* Wiley, New York, 1975.

[15] L. Kleinrock, *Queueing Systems, Vol. II: Computer Applications.* Wiley, New York, 1976.

[16] A.M. Lee, *Applied queuing theory*, MacMillan, London, 1968.

[17] J.D. Little, *A proof of the queueing formula $L = \lambda W$*, Opns. Res., 9 (1961), pp. 383–387.

[18] R.A. Marie, *Calculating equilibrium probabilities for $\lambda(n)/C_k/1/N$ queue*, in: Proceedings Performance' 80, Toronto, (May 28–30, 1980), pp. 117–125.

[19] G.F. Newell, *Applications of queuing theory*, Chapman and Hall, London, 1971.

[20] S.M. Ross, *Introduction to probability models*, 6th ed., Academic Press, London, 1997.

[21] M. Rubinovitch, *The slow server problem*, J. Appl. Prob., 22 (1985), pp. 205–213.

[22] M. Rubinovitch, *The slow server problem: a queue with stalling*, J. Appl. Prob., 22 (1985), pp. 879–892.

[23] R.S. Schassberger, *On the waiting time in the queueing system $GI/G/1$*, Ann. Math. Statist., 41 (1970), pp. 182–187.

[24] R.S. Schassberger, *Warteschlangen,* Springer-Verlag, Berlin, 1973.

[25] S. Stidham, *A last word on $L = \lambda W$*, Opns. Res., 22 (1974), pp. 417–421.

[26] L. Takacs, *Introduction to the theory of queues*, Oxford, 1962.

[27] H.C. Tijms, *Stochastic modelling and analysis: a computational approach*, John Wiley & Sons, Chichester, 1990.

[28] H.C. Tijms, *Stochastic models: an algorithmic approach*, John Wiley & Sons, Chichester, 1994.

[29] W. Whitt *Approximating a point process by a renewal process I: two basic methods*, Opns. Res., 30 (1986), pp. 125–147.

[30] E.T. Whittaker, G.N. Watson, *Modern analysis*, Cambridge, 1946.

[31] R.W. Wolff, *Poisson arrivals see time averages*, Opns. Res., 30 (1982), pp. 223–231.

[32] R.W. Wolff, *Stochastic modeling and the theory of queues*, Prentice-Hall, London, 1989.

Index