# Introduction to Queueing Theory and Stochastic Teletraffic Models

**Moshe Zukerman**

**EE Department,**

**City University of Hong Kong**

email: moshezu@gmail.com



## Preface

The aim of this textbook is to provide students with basic knowledge of stochastic models that may apply to telecommunications research areas, such as traffic modelling, resource provisioning and traffic management. These study areas are often collectively called *teletraffic*. This book assumes prior knowledge of a programming language and mathematics normally taught in an electrical engineering bachelor program.

The book aims to enhance intuitive and physical understanding of the theoretical concepts it introduces. The famous mathematician Pierre-Simon Laplace is quoted to say that "Probability is common sense reduced to calculation" [14]; as the content of this book falls under the field of applied probability, Laplace's quote very much applies. Accordingly, the book aims to link intuition and common sense to the mathematical models and techniques it uses. It mainly focuses on steady-state analyses used mainly in applications associated with dimensioning of networks and systems and avoids discussions of time-dependent analyses.

A unique feature of this book is the considerable attention given to guided projects involving computer simulations and analyzes. By successfully completing the programming assignments, students learn to simulate and analyze stochastic models, such as queueing systems and networks, and by interpreting the results, they gain insight into the queueing performance effects and principles of telecommunications systems modelling. Although the book, at times, provides intuitive explanations, it still presents the important concepts and ideas required for the understanding of teletraffic, queueing theory fundamentals and related queueing behavior of telecommunications networks and systems. These concepts and ideas form a strong base for the more mathematically inclined students who can follow up with the extensive literature on probability models and queueing theory. A small sample of it is listed at the end of this book.

The first two chapters provide background on probability and stochastic processes topics relevant to the queueing and teletraffic models of this book. These two chapters provide a summary of the key topics with relevant homework assignments that are especially tailored for understanding the queueing and teletraffic models discussed in later chapters. The content of these chapters is mainly based on [14, 25, 74, 79, 80, 81]. Students are encouraged to study also the original textbooks that include far more explanations, illustrations, discussions, examples and homework assignments.

Chapter 3 discusses general queueing notation and concepts and it should be studied well. Chapter 4 aims to assist the student to perform simulations of queueing systems. Simulations are useful and important in the many cases where exact analytical results are not available. An important learning objective of this book is to train students to perform queueing simulations. Chapter 5 provides analyses of deterministic queues. Many queueing theory books tend to exclude deterministic queues; however, the study of such queues is useful for beginners in that it helps them better understand non-deterministic queueing models. Chapters 6 – 14 provide analyses of a wide range of queueing and teletraffic models most of which fall under the category of continuous-time Markov-chain processes. Chapter 15 provides an example of a discrete-time queue that is modelled as a discrete-time Markov-chain. In Chapters 16 and 17, various aspects of a single server queue with Poisson arrivals and general service times are studied, mainly focussing on mean value results as in [13]. Then, in Chapter 18, some selected results of a single server queue with a general arrival process and general service times are provided. Next, in Chapter 19, we extend our discussion to queueing networks. Finally, in Chapter 20, stochastic processes that have been used as traffic models are discussed with special focus on their characteristics that affect queueing performance.

Throughout the book there is an emphasis on linking the theory with telecommunications applications as demonstrated by the following examples. Section 1.19 describes how properties of Gaussian distribution can be applied to link dimensioning. Section 6.6 shows, in the context of an M/M/1 queueing model, how optimally to set a link service rate such that delay requirements are met and how the level of multiplexing affects the spare capacity required to meet such delay requirement. An application of M/M/$\infty$ queueing model to a multiple access performance problem [13] is discussed in Section 7.6. In Sections 8.8 and 9.5, discussions on dimensioning and related utilization issues of a multi-channel system are presented. Especially important is the emphasis on the insensitivity property of models such as M/M/$\infty$, M/M/$k/k$, processor sharing and multi-service that lead to practical and robust approximations as described in Sections 7, 8, 13, and 14. Section 19.4 guides the reader to simulate a mobile cellular network. Section 20.6 describes a traffic model applicable to the Internet.

Last but not least, the author wishes to thank all the students and colleagues that provided comments and questions that helped developing and editing the manuscript over the years.

# Contents

# 1 Background on Relevant Probability Topics

Probability theory provides the foundation for queueing theory and stochastic teletraffic models, therefore it is important that the student masters the probability concepts required for the material that follows. Although the cover here is comprehensive in the sense that it discusses all the probability concepts and techniques used in later chapters, it does not include the many examples and exercises that are normally included in a probability textbook to help readers grasp the material better. Therefore, readers may be aided by additional probability texts, such as [14] and [80].

## 1.1 Events, Sample Space, and Random Variables

Consider an experiment with an uncertain outcome. The term "experiment" refers to any uncertain scenario, such as tomorrow's weather, tomorrow's share price of a certain company, or the result of flipping a coin. The *sample space* is a set of all possible outcomes of an experiment. An *event* is a subset of the sample space. Consider, for example, an experiment which consists of tossing a die. The sample space is $\{1, 2, 3, 4, 5, 6\}$, and an event could be the set $\{2, 3\}$, or $\{6\}$, or the empty set $\{\}$ or even the entire sample space $\{1, 2, 3, 4, 5, 6\}$. Events are called *mutually exclusive* if their intersection is the empty set. A set of events is said to be *exhaustive* if its union is equal to the sample space.

A *random variable* is a real valued function defined on the sample space. This definition appears somewhat contradictory to the wording "random variable" as a random variable is not at all random, because it is actually a deterministic function which assigns a real valued number to each possible outcome of an experiment. It is the outcome of the experiment that is random and therefore the name: random variable. If we consider the flipping a coin experiment, the possible outcomes are Head (H) and Tail (T), hence the sample space is $\{\text{H}, \text{T}\}$, and a random variable $X$ could assign $X = 1$ for the outcome H, and $X = 0$ for the outcome T.

If $X$ is a random variable then $Y = g(X)$ for some function $g(\cdot)$ is also a random variable. In particular, some functions of interest are $Y = cX$ for some constant $c$ and $Y = X^n$ for some integer $n$.

If $X_1, X_2, X_3, \ldots, X_n$ is a sequence of random variables, then $Y = \sum_{i=1}^{n} X_i$ is also a random variable.

### Homework 1.1

Consider the experiment to be tossing a coin. What is the Sample Space? What are the events associated with this Sample Space?

### Guide

Notice that although the sample space includes only the outcome of the experiments which are Head (H) and Tail (T), the events associated with this samples space includes also the empty set which in this case is the event $\{H \cap T\}$ and the entire sample space which in this case is the event $\{H \cup T\}$. $\square$

## 1.2 Probability, Conditional Probability and Independence

Consider a sample space $S$. Let $A$ be a subset of $S$, the probability of $A$ is the function on $S$ and all its subsets, denoted $\text{P}(A)$, that satisfies the following three axioms:

1. $0 \leq P(A) \leq 1$
2. $P(S) = 1$
3. The probability of the union of mutually exclusive events is equal to the sum of the probabilities of these events.

Normally, higher probability signifies higher likelihood of occurrence. In particular, if we conduct a very large number of experiments, and we generate the *histogram* by measuring how many times each of the possible occurrences actually occurred. Then we normalize the histogram by dividing all its values by the total number of experiments to obtain the relative frequencies. These measurable relative frequencies can provide intuitive interpretation to the theoretical concept of probability. Accordingly, the term *limiting relative frequency* is often used as interpretation of probability.

We use the notation $P(A \mid B)$ for the *conditional probability* of $A$ given $B$, which is the probability of the event $A$ given that we know that event $B$ has occurred. If we know that $B$ has occurred, it is our new sample space, and for $A$ to occur, the relevant experiments outcomes must be in $A \cap B$, hence the new probability of $A$, namely the probability $P(A \mid B)$, is the ratio between the probability of $A \cap B$ and the probability of $B$. Accordingly,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{1}$$

Since the event $\{A \cap B\}$ is equal to the event $\{B \cap A\}$, we have that

$$P(A \cap B) = P(B \cap A) = P(B \mid A)P(A),$$

so by (1) we obtain

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{2}$$

Eq. (2) is useful to obtain conditional probability of one event ($A$) given another ($B$) when $P(B \mid A)$ is known or easier to obtain then $P(A \mid B)$.

**Remark:** The intersection of $A$ and $B$ is also denoted by $A, B$ or $AB$ in addition to $A \cap B$.

If events $A$ and $B$ are *independent*, which means that if one of them occurs, the probability of the other to occur is not affected, then

$$P(A \mid B) = P(A) \tag{3}$$

and hence, by Eq. (1), if $A$ and $B$ are independent then,

$$P(A \cap B) = P(A)P(B). \tag{4}$$

Let $B_1,\ B_2,\ B_3,\ \ldots,\ B_n$ be a sequence of mutually exclusive and exhaustive events in $S$, and let $A$ be another event in $S$. Then,

$$A = \bigcup_{i=1}^{n} (A \cap B_i) \tag{5}$$

and since the $B_i$s are mutually exclusive, the events $A \cap B_i$s are also mutually exclusive. Hence,

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i). \tag{6}$$

Thus, by Eq. (1),

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \times P(B_i). \tag{7}$$

The latter is a very useful formula for deriving probability of a given event by conditioning and unconditioning on a set of mutually exclusive and exhaustive events. It is called *the Law of Total Probability*. Therefore, by Eqs. (7) and (1) (again), we obtain the following formula for conditional probability between two events:

$$P(B_1 \mid A) = \frac{P(A \mid B_1)P(B_1)}{\sum_{i=1}^{n} P(A \mid B_i) \times P(B_i)}. \tag{8}$$

The latter is known as Bayes' theorem (or Bayes' law or Bayes' rule).

Note that Eq. (1) is also referred to as Bayes' theorem.

## 1.3 Probability and Distribution Functions

Random variables are related to events. When we say that random variable $X$ takes value $x$, this means that $x$ represents a certain outcome of an experiment which is an event, so $\{X = x\}$ is an event. Therefore, we may assign probabilities to all possible values of the random variable. This function denoted $P_X(x) = P(X = x)$ will henceforth be called *probability function*. Other names used in the literature for a probability function include *probability distribution function*, *probability distribution*, or simply *distribution*. Because probability theory is used in many applications, in many cases, there are many alternative terms to describe the same thing. It is important that the student is familiar with these alternative terms, so we will use these terms alternately in this book.

The *cumulative distribution function*) of random variable $X$ is defined for all $x \in R$ ($R$ being the set of all real numbers), is defined as

$$F_X(x) = P(X \leq x). \tag{9}$$

Accordingly, the *complementary distribution function* $\bar{F}_X(x)$ is defined by

$$\bar{F}_X(x) = P(X > x). \tag{10}$$

Consequently, for any random variable, for every $x \in R$, $F(x) + \bar{F}(x) = 1$. As the complementary and the cumulative distribution functions as well as the probability function can be obtained from each other, we will use the terms *distribution function* when we refer to any of these functions without being specific.

We have already mentioned that if $X$ is a random variable, then $Y = g(X)$ is also a random variable. In this case, if $P_X(x)$ is the probability function of $X$ then the probability function of $Y$ is

$$P_Y(y) = \sum_{x:g(x)=y} P_X(x). \tag{11}$$

## 1.4 Joint Distribution Functions

In some cases, we are interested in the probability that two or more random variables are within a certain range. For this purpose, we define, *the joint distribution function* for $n$ random variables $X_1,\ X_2,\ \ldots,\ X_n$, as follows:

$$F_{X_1,\ X_2,\ \ldots,X_n}(x_1,\ x_2,\ \ldots,x_n) = P(X_1 \leq x_1,\ X_2 \leq x_2,\ \ldots,\ X_n \leq x_n). \tag{12}$$

Having the joint distribution function, we can obtain the distribution function of a single random variable, say, $X_1$, as

$$F_{X_1}(x_1) = F_{X_1,\ X_2,\ \ldots,\ X_n}(x_1,\ \infty,\ \ldots,\ \infty). \tag{13}$$

A random variable is called *discrete* if it takes at most a countable number of possible values. On the other hand, a *continuous* random variable takes an uncountable number of possible values.

When the random variables $X_1,\ X_2,\ \ldots,X_n$ are discrete, we can use their *joint probability function* which is defined by

$$P_{X_1,\ X_2,\ \ldots,\ X_n}(x_1,\ x_2,\ \ldots,\ x_n) = P(X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n). \tag{14}$$

The probability function of a single random variable can then be obtained by

$$P_{X_1}(x_1) = \sum_{x_2} \cdots \sum_{x_n} P_{X_1,\ X_2,\ \ldots,\ X_n}(x_1,\ x_2,\ \ldots,\ x_n). \tag{15}$$

In this section and in sections 1.5, 1.6, 1.7, when we mention random variables or their distribution functions, we consider them all to be discrete. Then, in Section 1.9, we will introduce the analogous definitions and notation relevant to their continuous counterparts.

We have already mention the terms probability function, distribution, probability distribution function and probability distribution. These terms apply to discrete as well as to continuous random variables. There are however additional terms that are used to describe probability function only for discrete random variable they are: *probability mass function*, and *probability mass*, and there are equivalent terms used only for continuous random variables – they are *probability density function*, *density function* and simply *density*.

## 1.5 Conditional Probability for Random Variables

The conditional probability concept, which we defined for events, can also apply to random variables. Because $\{X = x\}$, namely, the random variable $X$ takes value $x$, is an event, by the definition of conditional probability (1) we have

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}. \tag{16}$$

Let $P_{X|Y}(x \mid y) = P(X = x \mid Y = y)$ be the conditional probability of a discrete random variable $X$ given $Y$, we obtain by (16)

$$P_{X|Y}(x \mid y) = \frac{P_{X,Y}(x, y)}{P_Y(y)}. \tag{17}$$

Noticing that

$$P_Y(y) = \sum_x P_{X,Y}(x, y), \tag{18}$$

we obtain by (17)

$$\sum_x P_{X|Y}(x \mid y) = 1. \tag{19}$$

This means that if we condition on the event $\{Y = y\}$ for a specific $y$, the probability function of $X$ given $\{Y = y\}$ is a legitimate probability function. This is consistent with our discussion above. The event $\{Y = y\}$ is the new sample space and $X$ has a legitimate distribution function there. By (17)

$$P_{X,Y}(x, y) = P_{X|Y}(x \mid y)P_Y(y) \tag{20}$$

and by symmetry

$$P_{X,Y}(x, y) = P_{Y|X}(y \mid x)P_X(x) \tag{21}$$

so the latter and (18) gives

$$P_Y(y) = \sum_x P_{X,Y}(x, y) = \sum_x P_{Y|X}(y \mid x)P_X(x) \tag{22}$$

which is another version of the Law of Total Probability (7).

## 1.6 Independence between Random Variables

The definition of independence between random variables is very much related to the definition of independence between events because when we say that random variables $U$ and $V$ are independent, it is equivalent to say that the events $\{U = u\}$ and $\{V = v\}$ are independent for every $u$ and $v$. Accordingly, random variables $U$ and $V$ are said to be independent if

$$P_{U,V}(u, v) = P_U(u)P_V(v) \quad \text{for all } u, v. \tag{23}$$

Notice that by (20) and (23), we obtain an equivalent definition of independent random variables $U$ and $V$ which is

$$P_{U|V}(u \mid v) = P_U(u) \tag{24}$$

which is equivalent to $P(A \mid B) = P(A)$ which we used to define independent events $A$ and $B$.

## 1.7 Convolution

Consider independent random variables $V_1$ and $V_2$ that have probability functions $P_{V_1}(v_1)$ and $P_{V_2}(v_2)$, respectively, and their sum which is another random variable $V = V_1 + V_2$. Let us now derive the probability function $P_V(v)$ of $V$.

$$\begin{aligned} P_V(v) &= P(V_1 + V_2 = v) \\ &= \sum_{v_1} P(V_1 = v_1, V_2 = v - v_1) \\ &= \sum_{v_1} P_{V_1}(v_1) P_{V_2}(v - v_1). \end{aligned}$$

The latter is called the *convolution* of the probability functions $P_{V_1}(v_1)$ and $P_{V_2}(v_2)$.

Let us now extend the result from two to $k$ random variables. Consider $k$ independent random variables $X_i$, $i = 1, 2, 3, \ \ldots, \ k$. Let $P_{X_i}(x_i)$ be the probability function of $X_i$, for $i = 1, 2, 3, \ \ldots, \ k$, and let $Y = \sum_{i=1}^{k} X_i$. If $k = 3$, we first compute the convolution of $X_1$ and $X_2$ to obtain the probability function of $V = X_1 + X_2$ using the above convolution formula and then we use the formula again to obtain the probability function of $Y = V + X_3 = X_1 + X_2 + X_3$. Therefore, for an arbitrary $k$, we obtain

$$P_Y(y) = \sum_{x_2,\ x_3,\ \ldots,\ x_k:\ x_2 + x_3 + \ \ldots\ + x_k \leq y} \left( P_{X_1}(y - \Sigma_{i=2}^{k} x_i) \prod_{i=2}^{k} P_{X_i}(x_i) \right). \tag{25}$$

If all the random variable $X_i$, $i = 1, 2, 3, \ \ldots, \ k$, are independent and identically distributed (IID) random variables, with probability function $P_{X_1}(x)$, then the probability function $P_Y(y)$ is called the k-fold convolution of $P_{X_1}(x)$.

**Homework 1.2**

Consider again the experiment to be tossing a coin. Assume that

$$P(H) = P(T) = 0.5.$$

Illustrate each of the Probability Axioms for this case. □

**Homework 1.3**

Now consider an experiment involving three coin tosses. The outcome of the experiment is now a 3-long string of Heads and Tails. Assume that all coin tosses have probability 0.5, and that the coin tosses are independent events.

1. Write the sample space where each outcome of the experiment is an ordered 3-long string of Heads and Tails.
2. What is the probability of each outcome?
3. Consider the event

   $$A = \{Exactly\ one\ head\ occurs\}.$$

   Find $P(A)$ using the additivity axiom.

**Partial Answer:** $P(A) = 1/8 + 1/8 + 1/8 = 3/8$. □

**Homework 1.4**

Now consider again three coin tosses. Find the probability $P(A \mid B)$ where $A$ and $B$ are the events:
$A$ = more that one head came up
$B$ = 1st toss is a head.

**Guide:**

$P(B) = 4/8$; $P(A \cap B) = 3/8$; $P(A \mid B) = (3/8)/(4/8) = 3/4$. □

**Homework 1.5**

Consider a medical test for a certain disease. The medical test detects the disease with probability 0.99 and fails to detect the disease with probability 0.01. If the disease is not present, the test indicates that it is present with probability 0.02 and that it is not present with probability 0.98. Consider two cases:
Case a: The test is done on a randomly chosen person from the population where the occurrence of the disease is 1/10000.
Case b: The test is done on patients that are referred by a doctor that have a prior probability (before they do the test) of 0.3 to have the disease.

Find the probability of a person to have the disease if the test shows positive outcome in each of these cases.

**Guide:**

$A$ = person has the disease.
$B$ = test is positive.
$\bar{A}$ = person does not have the disease.
$\bar{B}$ = test is negative.
We need to find $P(A \mid B)$.

**Case a:**
We know: $P(A) = 0.0001$.
$P(\bar{A}) = 0.9999$.
$P(B \mid A) = 0.99$.
$P(B \mid \bar{A}) = 0.02$.

By the Law of Total Probability:

$$P(B) = P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A}).$$

$P(B) = 0.99 \times 0.0001 + 0.02 \times 0.9999 = 0.020097$.

Now put it all together and use Eq. (2) to obtain:

$P(A \mid B) = 0.004926108.$

**Case b:**
$P(A) = 0.3.$

Repeat the previous derivations to show that for this case $P(A \mid B) = 0.954983923$. □

**Homework 1.6**

In a multiple choice exam, there are 4 answers to a question. A student knows the right answer with probability 0.8 (Case 1), with probability 0.2 (Case 2), and with probability 0.5 (Case 3). If the student does not know the answer s/he always guesses with probability of success being 0.25. Given that the student marked the right answer, what is the probability he/she knows the answer.

**Guide:**

$A$ = Student knows the answer.
$B$ = Student marks correctly.
$\bar{A}$ = Student does not know the answer.
$\bar{B}$ = Student marks incorrectly.
We need to find $P(A \mid B)$.

**Case 1:**

We know: $P(A) = 0.8$.
$P(\bar{A}) = 0.2$.
$P(B \mid A) = 1$.
$P(B \mid \bar{A}) = 0.25$.

By the Law of Total Probability:

$$P(B) = P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A}).$$

$P(B) = 1 \times 0.8 + 0.25 \times 0.2 = 0.85.$

Now put it all together and use Eq. (2) to obtain:

$$P(A \mid B) = 0.941176471.$$

**Case 2:**

Repeat the previous derivations to obtain:

$P(A) = 0.2$

$P(B) = 0.4$

$P(A \mid B) = 0.5.$

**Case 3:**

Repeat the previous derivations to obtain:

$P(A) = 0.5$

$P(B) = 0.625$

$P(A \mid B) = 0.8.$ □

**Homework 1.7**

Watch the following youtube link on the problem known as the Monty Hall problem:
`https://www.youtube.com/watch?v=mhlc7peGlGg`
Then study the Monty Hall problem in:
`http://en.wikipedia.org/wiki/Monty_Hall_problem`
and understand the different solution approaches and write a computer simulation to estimate the probability of winning the car by always switching. □

## 1.8 Selected Discrete Random Variables

We present here several discrete random variables and their corresponding distribution functions. Although our cover here is non-exhaustive, we do consider all the discrete random variables mentioned later in this book.

### 1.8.1 Non-parametric

Discrete non-parametric random variable $X$ is characterized by a set of $n$ values that $X$ can take with nonzero probability: $a_1, a_2, \ldots, a_n$, and a set of $n$ probability values $p_1, p_2, \ldots p_n$, where $p_i = P(X = a_i), \quad i = 1, 2, \ldots n$. The distribution of $X$ in this case is called a *non-parametric distribution* because it does not depend on a mathematical function that its shape and range are determined by certain parameters of the distribution.

### 1.8.2 Bernoulli

We begin with the Bernoulli random variable. It represents an outcome of an experiment which has only two possible outcomes. Let us call them "success" and "failure". These two outcomes are mutually exclusive and exhaustive events. The Bernoulli random variable assigns the value $X = 1$ to the "success" outcome and the value $X = 0$ to the "failure" outcome. Let $p$ be the probability of the "success" outcome, and because "success" and "failure" are mutually exclusive and exhaustive, the probability of the "failure" outcome is $1 - p$. The probability function in terms of the Bernoulli random variable is:

$$P(X = 1) = p \qquad (26)$$
$$P(X = 0) = 1 - p.$$

### 1.8.3 Geometric

There are two types of geometric distributions. The first one is a distribution of a random variable $X$ that represents the number of independent Bernoulli trials, each of which with $p$ being the probability of success, required until the first success. For $X$ to be equal to $i$ we must have $i-1$ consecutive failures and then one success in $i$ independent Bernoulli trials. Therefore, we obtain

$$P(X = i) = (1 - p)^{i-1}p \quad \text{for} \;\; i = 1, 2, 3, \; \ldots \; . \qquad (27)$$

The complementary distribution function of the geometric random variable is

$$P(X > i) = (1 - p)^{i} \quad \text{for} \;\; i = 0, 1, 2, 3, \; \ldots \; .$$

The second type is a distribution of a random variable $Y$ that represents the number of failures occur *before* the first success and accordingly it is given by $Y = X - 1$ where $X$ is a random variable of the geometric distribution of the first type.

In this book, we use the name *geometric distribution* for the geometric distribution of the first type and the geometric distribution of the second type we call *geometric distribution of failures*. Then the corresponding random variables will be called the *geometric random variable* and the *geometric random variable of failures*, respectively.

A geometric random variable (of either type) possesses an important property called memorylessness. In particular, discrete random variable $X$ is *memoryless* if

$$P(X > m + n \mid X > m) = P(X > n), \quad m = 0, 1, 2, \ldots, \text{and} \;\; n = 0, 1, 2, \ldots \qquad (28)$$

The Geometric random variable is memoryless because it is based on independent Bernoulli trials, and therefore the fact that so far we had $m$ failures does not affect the probability that the next $n$ trials will be failures.

The two types of geometric random variables are the only discrete random variables that are memoryless.

### 1.8.4 Binomial

Assume that $n$ independent Bernoulli trials are performed. Let $X$ be a random variable representing the number of successes in these $n$ trials. Such random variable is called a binomial random variable with parameters $n$ and $p$. Its probability function is:

$$P(X = i) = \binom{n}{i} p^{i}(1 - p)^{n-i} \qquad i = 0, \; 1, \; 2, \; \ldots, \; n.$$

Notice that a Binomial random variable with parameters 1 and $p$ is a Bernoulli random variable. The Bernoulli and binomial random variables have many applications. In particular, it is used as a model for voice as well as data sources. Such sources alternates between two states "on" and "off". During the "on" state the source is active and transmits at the rate equal to the transmission rate of its equipment (e.g. a modem), and during the "off" state, the source is idle. If $p$ is the proportion of time that the source is active, and if we consider a superposition of $n$ independent identical sources, than the binomial distribution gives us the probability of the number of sources which are simultaneously active which is important for resource provisioning.

**Homework 1.8**

Consider a state with voter population $N$. There are two candidates in the state election for governor and the winner is chosen based on a simple majority. Let $N_1$ and $N_2$ be the total number of votes obtained by Candidates 1 and 2, respectively, from voters other than Johnny. Johnny just voted for Candidate 1, and he would like to know the probability that his vote affects the election results. Johnny realizes that the only way he can affect the result of the election is if the votes are equal (without his vote) or the one he voted for (Candidate 1) had (without his vote) one call less than Candidate 2. That is, he tries to find the probability of the event

$$0 \geq N_1 - N_2 \geq -1.$$

Assume that any other voter (excluding Johnny) votes independently for Candidates 1 and 2 with probabilities $p_1$ and $p_2$, respectively, and also that $p_1 + p_2 < 1$ to allow for the case that a voter chooses not to vote for either candidate. Derive a formula for the probability that Johnny's vote affects the election results and provide an algorithm and a computer program to compute it for the case $N = 2,000,000$ and $p_1 = p_2 = 0.4$.

**Guide**

By the definition of conditional probability,

$$P(N_1 = n_1, N_2 = n_2) = P(N_1 = n_1)P(N_2 = n_2 \mid N_1 = n_1)$$

so

$$P(N_1 = n_1, N_2 = n_2) = \binom{N-1}{n_1} p_1^{n_1}(1-p_1)^{N-n_1-1}\binom{N-n_1-1}{n_2} p_2^{n_2}(1-p_2)^{N-n_1-1-n_2}.$$

Then, as the probability of the union of mutually exclusive events is the sum of their probabilities, the required probability is

$$\sum_{k=0}^{\lfloor (N-1)/2 \rfloor} P(N_1 = k, N_2 = k) + \sum_{k=0}^{\lceil (N-1)/2 \rceil - 1} P(N_1 = k, N_2 = k+1).$$

where $\lfloor x \rfloor$ is the largest integer smaller or equal to x. and $\lceil x \rceil$ is the smallest integer greater or equal to x. □

Next, let us derive the probability distribution of the random variable $Y = X_1 + X_2$ where $X_1$ and $X_2$ are two independent Binomial random variables with parameters $(N_1, p)$ and $(N_2, p)$, respectively. This will require the derivation of the convolution of $X_1$ and $X_2$ as follows.

$$\begin{aligned}
P_Y(k) &= P(X_1 + X_2 = k) \\
&= \sum_{i=0}^{k} P(\{X_1 = i\} \cap \{X_2 = k - i\}) \\
&= \sum_{i=0}^{k} P_{X_1}(i) P_{X_2}(k - i) \\
&= \sum_{i=0}^{k} \binom{N_1}{i} p^i (1-p)^{N_1 - i} \binom{N_2}{k-i} p^{k-i} (1-p)^{N_2 - (k-i)} \\
&= p^k (1-p)^{N_1 + N_2 - k} \sum_{i=0}^{k} \binom{N_1}{i} \binom{N_2}{k-i} \\
&= p^k (1-p)^{N_1 + N_2 - k} \binom{N_1 + N_2}{k}.
\end{aligned}$$

We can conclude that the convolution of two independent binomial random variables with parameters $(N_1, p)$ and $(N_2, p)$ has a binomial distribution with parameters $N_1 + N_2, p$. This is not surprising. As we recall the binomial random variable represents the number of successes of a given number of independent Bernoulli trials. Thus, the event of having $k$ Bernoulli successes out of $N_1 + N_2$ trials is equivalent to the event of having some (or none) of the successes out of the $N_1$ trials and the remaining out of the $N_2$ trials.

**Homework 1.9**

In the last step of the above proof, we have used the equality

$$\binom{N_1 + N_2}{k} = \sum_{i=0}^{k} \binom{N_1}{i} \binom{N_2}{k-i}.$$

Prove this equality.

**Guide**

Consider the equality

$$(1 + \alpha)^{N_1 + N_2} = (1 + \alpha)^{N_1} (1 + \alpha)^{N_2}.$$

Then equate the binomial coefficients of $\alpha^k$ on both sides. □

### 1.8.5 Poisson

A Poisson random variable with parameter $\lambda$ has the following probability function:

$$P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!} \qquad i = 0,\ 1,\ 2,\ 3,\ \ldots\ . \tag{29}$$

To compute the values of $P(X = i)$, it may be convenient to use the recursion

$$P(X = i + 1) = \frac{\lambda}{i + 1} P(X = i) \tag{30}$$

with

$$P(X = 0) = e^{-\lambda}.$$

However, if the parameter $\lambda$ is large, there may be a need to set $\hat{P}(X = \lfloor \lambda \rfloor) = 1$, where $\lfloor x \rfloor$ is the largest integer smaller or equal to x, and to compute recursively, using (30) (applied to the $\hat{P}(X = i)$'s), a sufficient number of $\hat{P}(X = i)$ values for $i > \lambda$ and $i < \lambda$ such that $\hat{P}(X = i) > \epsilon$, where $\epsilon$ is chosen to meet a given accuracy requirement. Clearly the $\hat{P}(X = i)$'s do not sum up to one, and therefore they do not represent a probability distribution. To approximate the Poisson probabilities they will need to be normalized as follows.

Let $\alpha$ and $\beta$ be the lower and upper bounds, respectively, of the $i$ values for which $\hat{P}(X = i) > \epsilon$. Then, the probabilities $P(X = i)$ are approximated using the normalization:

$$P(X = i) = \begin{cases} \frac{\hat{P}(X=i)}{\sum_{i=\alpha}^{\beta} \hat{P}(X=i)} & \alpha \leq i \leq \beta \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

The importance of the Poisson random variable lies in its property to approximate the binomial random variable in case when $n$ is very large and $p$ is very small so that $np$ is not too large and not too small. In particular, consider a sequence of binomial random variables $X_n, \ \ n = 1, 2, \ldots$ with parameters $(n, p)$ where $\lambda = np$, or $p = \lambda/n$. Then the probability function

$$\lim_{n \to \infty} P(X_n = k)$$

is a Poisson probability function with parameter $\lambda$.

To prove this we write:

$$\lim_{n \to \infty} P(X_n = k) = \lim_{n \to \infty} \binom{n}{k} p^k (1 - p)^{n-k}.$$

Substituting $p = \lambda/n$, we obtain

$$\lim_{n \to \infty} P(X_n = k) = \lim_{n \to \infty} \frac{n!}{(n - k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

or

$$\lim_{n \to \infty} P(X_n = k) = \lim_{n \to \infty} \frac{n!}{(n - k)!n^k} \left(\frac{\lambda^k}{k!}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}.$$

Now notice that

$$\lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda},$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k}=1,$$

and

$$\lim_{n\to\infty}\frac{n!}{(n-k)!n^k}=1.$$

Therefore,

$$\lim_{n\to\infty}P(X_n=k)=\frac{\lambda^k e^{-\lambda}}{k!}.$$

In Subsection 1.15.1 this important limit will be shown using Z-transform. The Poisson random variable accurately models the number of calls arriving at a telephone exchange or Internet service provider in a short period of time, a few seconds or a minute, say. In this case, the population of customers (or flows) $n$ is large. The probability $p$ of a customer making a call within a given short period of time is small, and the calls are typically independent because they are normally generated by independent individual people from a large population. Therefore, models based on Poisson random variables have been used successfully for design and dimensioning of telecommunications networks and systems for many years. When we refer to items in a queueing system in this book, they will be called customers, jobs or packets, interchangeably.

Next, let us derive the probability distribution of the random variable $Y=X_1+X_2$ where $X_1$ and $X_2$ are two independent Poisson random variables with parameters $\lambda_1$ and $\lambda_2$, respectively. This will require the derivation of the convolution of $X_1$ and $X_2$ as follows.

$$\begin{aligned}
P_Y(k) &= P(X_1+X_2=k)\\
&= \sum_{i=0}^{k}P(\{X_1=i\}\cap\{X_2=k-i\})\\
&= \sum_{i=0}^{k}P_{X_1}(i)P_{X_2}(k-i)\\
&= \sum_{i=0}^{k}\frac{\lambda_1^i}{i!}e^{-\lambda_1}\frac{\lambda_2^{k-i}}{(k-i)!}e^{-\lambda_2}\\
&= e^{-(\lambda_1+\lambda_2)}\sum_{i=0}^{k}\frac{\lambda_1^i\lambda_2^{k-i}}{i!(k-i)!}\\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{i=0}^{k}\frac{k!\lambda_1^i\lambda_2^{k-i}}{i!(k-i)!}\\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{i=0}^{k}\binom{k}{i}\lambda_1^i\lambda_2^{k-i}
\end{aligned}$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^k}{k!}.$$

We have just seen that the random variable $Y$ has a Poisson distribution with parameter $\lambda_1 + \lambda_2$.

**Homework 1.10**

Consider a Poisson random variable $X$ with parameter $\lambda = 500$. Write a program that computes the probabilities $P(X = i)$ for $0 \leq i \leq 800$ and plot the function $P_X(x)$. □

**Homework 1.11**

Let $X_1$ and $X_2$ be two independent Poisson distributed random variables with parameters $\lambda_1$ and $\lambda_2$, respectively. Let $Y = X_1 + X_2$. Find the distribution of $(X_1 \mid Y)$. In particular, find for any given $k$, the conditional probability $P(X_1 = j \mid Y = k)$ for $j = 0, 1, 2, \ldots, k$.

**Guide**

$$P(X_1 = j \mid Y = k) = \frac{P(X_1 = j \cap Y = k)}{P(Y = k)}.$$

Now notice that

$$\{X_1 = j \cap Y = k\} = \{X_1 = j \cap X_2 = k - j\}.$$

When you claim that the two events $A$ and $B$ are equal, you must be able to show that $A$ implies $B$ and that $B$ implies $A$. Show both for the present case.

Because $X_1$ and $X_2$ are independent, we have that

$$P(X_1 = j \cap X_2 = k - j) = P(X_1 = j)P(X_2 = k - j).$$

Now recall that $X_1$ and $X_2$ are Poisson distributed random variables with parameters $\lambda_1$ and $\lambda_2$, respectively, and that $Y$ is the convolution of $X_1$ and $X_2$ and therefore has a Poisson distribution with parameter $\lambda_1 + \lambda_2$. Put it all together and show that $P(X_1 = j \mid Y = k)$ is a binomial probability function with parameters $\lambda_1/(\lambda_1 + \lambda_2)$ and $k$.

□

### 1.8.6 Pascal

The Pascal random variable $X$ with parameters $k$ (integer and $\geq 1$) and $p$ (real within (0,1]), represents a sum of $k$ geometric random variables each with parameter $p$. This is equivalent to having the $k$th successful Bernoulli trial at the $i$th trial. In other words, the Pascal random variable represents the number of independent Bernoulli trials until the $k$th success. Accordingly, for a Pascal random variable $X$ to be equal to $i$, the $i$th trial must be the $k$th successful trial associated with the $k$th geometric random variable. Then, there must also be exactly $k-1$ successes among the first $i-1$ trials.

The probability to have a success at the $i$th trial, as in any trial, is equal to $p$, and the probability of having $k-1$ successes among the first $i-1$ is equal to the probability of having a binomial random variable with parameters $p$ and $i-1$ equal to $k-1$, for $i \geq k \geq 1$, which is equal to

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

and since the two random variables here, namely, the Bernoulli and the Binomial are independent (because the underlying Bernoulli trials are independent), we obtained

$$P(X = i) = \binom{i-1}{k-1} p^{k}(1-p)^{i-k} \qquad i = k,\ k+1,\ k+2,\ \ldots\ . \tag{32}$$

An alternative formulation of the Pascal random variable $Y$ is defined as the number of failures required to have $k$ successes. In this case, the relationship between $Y$ and $X$ is given by $Y = X - k$ for which the probability mass function is given by

$$P(Y = j) = \binom{k+j-1}{k-1} p^{k}(1-p)^{j} \qquad j = 0,\ 1,\ 2,\ \ldots\ . \tag{33}$$

One important property of the latter is that the cumulative distribution function can be expressed as

$$F_Y(j) = P(Y \leq j) = 1 - I_p(j+1, k) \qquad j = 0,\ 1,\ 2,\ \ldots\ . \tag{34}$$

where $I_p(j+1, k)$ is the regularized incomplete beta function with parameters $p, j+1$ and $k$.

### 1.8.7 Discrete Uniform

The discrete uniform probability function with integer parameters $a$ and $b$ takes equal non-zero values for $x = a, a+1, a+2, \ldots, b$. Its probability function is given by

$$P_X(x) = \begin{cases} \frac{1}{b-a+1} & \text{if } x = a, a+1, a+2, \ldots, b \\ 0 & \text{otherwise.} \end{cases}$$

**Homework 1.12**

Consider a group people. Assume that their birthdays are independent and there are 365 days in a year. That is, their birthdays can be viewed as independent discrete uniform random variables with parameters 1 and 365. Let $A$ be the event that there are at least two people in a group with the same birthday. Let $P_N(A)$ be the probability of the event $A$ occurs given that the number of people in a group is $N$. Write a recursive formula to derive $P_N(A)$. Obtain the values of $P_2(A)$, $P_5(A)$ and $P_{10}(A)$. Find the smallest values of $N$, such that $P_N(A) \geq 0.5$ and $P_N(A) \geq 0.999$, respectively.

**Guide and answers**

Notice the following recursive relationship for the probability of not having two people with the same birthday among $N$ people:

$$1 - P_N(A) = \left(1 - \frac{N-1}{365}\right)(1 - P_{N-1}(A)), \quad \text{for} \quad N = 2, 3, 4, \ \ldots, \ 365,$$

with $P_1(A) = 0$, and $P_N(A) = 1$ for $N \geq 365$.

$P_2(A) = 0.00274$, $P_5(A) = 0.027$ and $P_{10}(A) = 0.117$

For $N = 23$, $P_{23}(A) = 0.5073$.

For $N = 70$, $P_{70}(A) = 0.99916$.

Are you surprised? □

## 1.9 Continuous Random Variables and their Distributions

Continuous random variables are related to cases whereby the set of possible outcomes is uncountable. A continuous random variable $X$ is a function that assigns a real number to outcome of an experiment, and is characterized by the existence of a function $f(\cdot)$ defined for all $x \in R$, which has the property that for any set $A \subset R$,

$$P(X \in A) = \int_A f(x)dx. \tag{35}$$

Such function is the *probability density function* (or simply the *density*) of $X$. Since the continuous random variable $X$ must take a value in $R$ with probability 1, $f$ must satisfy,

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \tag{36}$$

If we consider Eq. (35), letting $A = [a, b]$, we obtain,

$$P(a \leq X \leq b) = \int_a^b f(x)dx. \tag{37}$$

An interesting point to notice is that the probability of a continuous random variable taking a particular value is equal to zero. If we set $a = b$ in Eq. (37), we obtain

$$P(X = a) = \int_a^a f(x)dx = 0. \tag{38}$$

As a result, for a continuous random variable $X$, the cumulative distribution function $F_X(x)$ is equal to both $P(X \leq x)$ and to $P(X < x)$. Similarly, the complementary distribution function is equal to both $P(X \geq x)$ and to $P(X > x)$.

By Eq. (37), we obtain

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(s)ds. \tag{39}$$

Hence, the probability density function is the derivative of the distribution function.

An important concept which gives rise to a continuous version of the Law of Total Probability is the continuous equivalence of Eq. (14), namely, the joint distribution of continuous random variables. Let $X$ and $Y$ be two continuous random variables. The joint density of $X$ and $Y$ denoted $f_{X,Y}(x,y)$ is a nonnegative function that satisfies

$$P(\{X,Y\} \in A) = \iint_{\{X,Y\}\in A} f_{X,Y}(x,y)dxdy \tag{40}$$

for any set $A \subset R^2$.

The continuous equivalence of the first equality in (22) is:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx. \tag{41}$$

Another important concept is the conditional density of one continuous random variable on another. Let $X$ and $Y$ be two continuous random variables with joint density $f_{X,Y}(x,y)$. For any $y$, such that the density of $Y$ takes a positive value at $Y = y$ (i.e. such that $f_Y(y) > 0$), the conditional density of $X$ given $Y$ is defined as

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}. \tag{42}$$

For every given fixed $y$, it is a legitimate density because

$$\int_{-\infty}^{\infty} f_{X|Y}(x \mid y)dx = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)dx}{f_Y(y)} = \frac{f_Y(y)}{f_Y(y)} = 1. \tag{43}$$

Notice the equivalence between the conditional probability (1) and the conditional density (42). By (42)

$$f_{X,Y}(x,y) = f_Y(y)f_{X|Y}(x \mid y) \tag{44}$$

so

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\infty}^{\infty} f_Y(y)f_{X|Y}(x \mid y)dy. \tag{45}$$

Recall again that $f_{X,Y}(x,y)$ is defined only for $y$ values such that $f_Y(y) > 0$.

Let define event $A$ as the event $\{X \in A\}$ for $A \subset R$. Thus,

$$P(A) = P(X \in A) = \int_A f_X(x)dx = \int_A \int_{-\infty}^{\infty} f_Y(y)f_{X|Y}(x \mid y)dydx. \tag{46}$$

Hence,

$$P(A) = \int_{-\infty}^{\infty} f_Y(y) \int_A f_{X|Y}(x \mid y)dxdy \tag{47}$$

and therefore

$$P(A) = \int_{-\infty}^{\infty} f_Y(y)P(A \mid Y = y)dy \tag{48}$$

which is the continuous equivalence of the Law of Total Probability (7).

**Homework 1.13**

I always use two trains to go to work. After traveling on my first train and some walking in the interchange station, I arrive at the platform (in the interchange station) at a time which is uniformly distributed between 9.01 and 9.04 AM. My second train arrives at the platform (in interchange station) exactly at times 9.02 and 9.04 AM. Derive the density of my waiting time at the interchange station. Ignore any queueing effects, and assume that I never incur any queueing delay.

**Guide:**

Use the continuous equivalence of the Law of Total Probability. □

We will now discuss the concept of convolution as applied to continuous random variables. Consider independent random variables $U$ and $V$ that have densities $f_U(u)$ and $f_V(v)$, respectively, and their sum which is another random variable $X = U + V$. Let us now derive the density $f_X(x)$ of $X$.

$$\begin{aligned} f_X(x) &= P(U + V = x) \\ &= \int_u f(U = u, V = x - u)du \\ &= \int_u f_U(u)f_V(x - u)du. \end{aligned} \tag{49}$$

The latter is the *convolution* of the densities $f_U(u)$ and $f_V(v)$.

As in the discrete case the convolution $f_Y(y)$, of $k$ densities $f_{X_i}(x_i)$, $i = 1, 2, 3, \ \ldots, \ k$, of random variables $X_i$, $i = 1, 2, 3, \ \ldots, \ k$, respectively, is given by

$$f_Y(y) = \iint_{x_2, \ \ldots, \ x_k: \ x_2 + \ \ldots, + x_k \leq y} \left( f_{X_1}(y - \Sigma_{i=2}^{k} x_i) \prod_{i=2}^{k} f_{X_i}(x_i) \right). \tag{50}$$

And again, in the special case where all the random variable $X_i$, $i = 1, 2, 3, \ \ldots, \ k$, are IID, the density $f_Y$ is the k-fold convolution of $f_{X_1}$.

**Homework 1.14**

Consider the following joint density:

$$f_{X,Y}(x, y) = \begin{cases} 2 & 0 \leq x + y \leq 1, \quad x \geq 0, \quad y \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{51}$$

1. Show that this is a legitimate joint density by showing first that all relevant probabilities are nonnegative and that the 2-dimensional integral of this joint density over the entire state space is equal to 1.
2. Derive the marginal density $f_Y(y)$.
3. Derive the conditional density $f_{Y|X}(y \mid x)$.

**Guide**

To show that this is a legitimate density observe that the joint density is nonnegative and also

$$\int_0^1 \int_0^{1-x} 2dydx = 1.$$

$$f_Y(y) = \begin{cases} \int_0^{1-y} f_{X,Y}(x,y)dx = 2-2y & 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases} \tag{52}$$

$$f_{Y|X}(y \mid x) = \begin{cases} \frac{2}{2-2y} = \frac{1}{1-y} & 0 \le y \le 1-x \\ 0 & \text{otherwise.} \end{cases} \tag{53}$$

□

## 1.10 Selected Continuous Random Variables

We will now discuss several continuous random variables and their corresponding probability distributions: uniform, exponential, hyper-exponential, Erlang, hypo-exponential Gaussian, multivariate Gaussian and Pareto. These are selected because of their applicability in teletraffic and related queueing models and consequently their relevance to the material in this book.

### 1.10.1 Uniform

The probability density function of the uniform random variable takes nonnegative values over the interval $[a, b]$ and is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise.} \end{cases} \tag{54}$$

Of particular interest is the special case - the uniform (0,1) random variable. Its probability density function is given by

$$f(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases} \tag{55}$$

The uniform (0,1) random variable is very important in simulations. Almost all computers languages have a function by which we can generate uniform (0,1) random deviates. By a simple transformation such uniform (0,1) random deviates can be translated to sequence of random deviates of any distribution as follows. Let $U_1(0,1)$ be the first uniform (0,1) random deviate, and let $F(x)$ be a distribution function of an arbitrary random variable. Set,

$$U_1(0,1) = F(x_1) \tag{56}$$

so $x_1 = F^{-1}(U_1(0,1))$ is the first random deviate from the distribution $F(\cdot)$. Then generating the second uniform (0,1) random deviate, the second $F(\cdot)$ random number is obtained in the same way, etc.

This method of generating random deviates from any distribution is known by the following names: inverse transform sampling, inverse transformation method, inverse probability integral transform, and Smirnov transform.

To see why this method works, let $U$ be a uniform (0,1) random variable. Let $F(x)$ be an arbitrary cumulative distribution function. Let the random variable $Y$ be defined by: $Y = F^{-1}(U)$. That is, $U = F(Y)$. We will now show that the distribution of $Y$, namely $P(Y \leq x)$, is equal to $F(x)$. Notice that $P(Y \leq x) = P[F^{-1}(U) \leq x] = P[U \leq F(x)]$. Because $U$ is a uniform (0,1) random variable, then $P[U \leq F(x)] = F(x)$. Thus, $P(Y \leq x) = F(x)$. □

Please notice that if $U$ is a uniform (0,1) random variable than $1 - U$ is also a uniform (0,1) random variable. Therefore, instead of (56), we can write

$$U_1(0,1) = 1 - F(x_1) = \bar{F}(x_1).$$

In various cases, it is more convenient to use the complementary $\bar{F}(x_1)$ instead of the cumulative distribution function. One of these cases is the exponential distribution as illustrated in the next section.

### Homework 1.15

Let $X_1, X_2, X_3, \ldots X_k$ be a sequence of $k$ independent random variables having a uniform $(0, s)$ distribution. Let $X = \min\{X_1, X_2, X_3, \ldots, X_k\}$. Prove that

$$P(X > t) = \begin{cases} 1 & \text{for } t \leq 0 \\ (1 - \frac{t}{s})^k & \text{for } 0 < t < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{57}$$

**Hint:** $P(X > t) = P(X_1 > t)P(X_2 > t)P(X_3 > t) \cdots P(X_k > t)$. □

### Homework 1.16

Derive the convolution of two independent uniform (0,1) random variables.

### Guide

Since $U$ and $V$ are uniform (0,1) random variables, for the product $f_U(u)f_V(x - u)$ in Eq. (49) to be non-zero, $u$ and $x$ must satisfy:
$0 \leq u \leq 1$ and $0 \leq x - u \leq 1$,
or
$\max(0, x - 1) \leq u \leq \min(1, x)$
and $0 \leq x \leq 2$.

Therefore,

$$f_X(x) = \begin{cases} u\Big|_{\max(0,x-1)}^{\min(1,x)} & 0 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

or

$$f_X(x) = \begin{cases} \min(1,x) - \max(0,x-1) & 0 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

□

### 1.10.2 Exponential

The exponential random variable has one parameter $\mu$ and its probability density function is given by,

$$f(x) = \begin{cases} \mu e^{-\mu x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{58}$$

Its distribution function is given by

$$F(x) = \int_0^x \mu e^{-\mu s} ds = 1 - e^{-\mu x} \qquad x \geq 0. \tag{59}$$

A convenient and useful way to describe the exponential random variable is by its complementary distribution function. It is given by,

$$\bar{F}(x) = e^{-\mu x} \qquad x \geq 0. \tag{60}$$

An important application of the exponential random variable is the time until the next call (or connection request) arrives at a switch.

Interestingly, such time does not depend on how long ago was the last call that arrived. In other words, the exponential random variable is memoryless. In particular, a continuous random variable is called memoryless if for any $t \geq 0$ and $s \geq 0$,

$$P(X > s + t \mid X > t) = P(X > s). \tag{61}$$

If our lifetime were memoryless, then the probability we survive at least 80 years given that we have survived 70 years is equal to the probability that a newborn baby lives to be 10 years. Of course human lifetime is not memoryless, but, as mentioned above, inter-arrivals of phone calls at a telephone exchange are approximately memoryless. To show that exponential random variable is memoryless we show that Eq. (61) holds using the conditional probability definition together with the complementary distribution function of an exponential random variable as follows.

$$\begin{aligned} P(X > s + t \mid X > t) &= \frac{P(X > s + t \cap X > t)}{P(X > t)} \\ &= \frac{P(X > s + t)}{P(X > t)} \\ &= \frac{e^{-\mu(s+t)}}{e^{-\mu t}} \\ &= e^{-\mu s} = P(X > s). \end{aligned}$$

Not only is exponential random variable memoryless, it is actually, the only memoryless continuous random variable.

**Homework 1.17**

Show how to apply the Inverse transform sampling to generate exponential deviates.

**Guide**

As discussed, we can use the complementary distribution function instead of the cumulative distribution function. Accordingly, to generate deviates from an exponential distribution with parameter $\lambda$, follow the following steps.

1. Obtain a new uniform (0,1) deviate. Most computer program provide a function that enable you to do it. Call this deviate $U(0,1)$.
2. Write: $U(0,1) = e^{-\lambda x}$. Then isolating $x$ in the latter gives:
$$x = \frac{-\ln U(0,1)}{\lambda}.$$
Using this equation, you can obtain an exponential deviate $x$ from the uniform (0,1) deviate $U(0,1)$.
3. Repeat steps 1 and 2 for each required exponential deviate.

□

**Homework 1.17**

Assume that the computer language that you use provides a sequence of independent uniform (0,1) deviates. Write computer programs that generate a sequence of 10,000 independent random deviates from:

1. an exponential distribution with $\mu = 1$;
2. a discrete uniform distribution with $a = 1$ and $b = 10$;
3. a geometric distribution with $p = 0.4$. Plot histograms for to demonstrate correctness these three cases.

□

**Homework 1.18**

Write a computer program that generates a sequence of 100 independent random deviates from a discrete uniform distribution with $a = 1$ and $b = 10$. Again, assume that the computer language that you use provides a sequence of independent uniform (0,1) deviates. □

Let $X_1$ and $X_2$ be independent and exponentially distributed random variables with parameters $\lambda_1$ and $\lambda_2$. We are interested to know the distribution of $X = \min[X_1, X_2]$. In other words, we are interested in the distribution of the time that passes until the first one of the two random variables $X_1$ and $X_2$ occurs. This is as if we have a competition between the two and we are interested in the time of the winner whichever it is. Then

$$P(X > t) = P(\min[X_1, X_2] > t) = P(X_1 > t, X_2 > t) = e^{-\lambda_1 t} e^{-\lambda_2 t} = e^{-(\lambda_1 + \lambda_2)t}. \quad (62)$$

Thus, the distribution of $X$ is exponential with parameter $\lambda_1 + \lambda_2$.

Another interesting question related to the competition between two exponential random variables is what is the probability that one of them, say $X_1$, wins. That is, we are interested in the probability of $X_1 < X_2$. This is obtained using the continuous version of the Law of Total Probability (48) as follows:

$$P(X_1 < X_2) = \int_0^\infty (1 - e^{-\lambda_1 t})\lambda_2 e^{-\lambda_2 t} dt = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \tag{63}$$

To understand the latter, note that $X_2$ can obtain many values: $t_1, t_2, t_3, \ldots$, infinitely many values ...

All these values, that $X_2$ may take, lead to the events
$X_2 = t_1, X_2 = t_2, X_2 = t_3, \ldots$ that are mutually exclusive and exhaustive.

Then, using the continuous version of the Law of Total Probability, namely, integration of the product
$P(X_1 < t)$ times the density of $X_2$ at $t$, will give us the probability of $X_1 < X_2$.
By integrating over all $t$ we "add up" the probabilities of infinitely many mutually exclusive and exhaustive events that make up the event $X_1 < X_2$. And this is exactly what the Law of Total Probability does!

In the following table we point out the equivalence between the corresponding terms in the two equations (48) and (63).

| **term in (48)** | **equivalent term in (63)** |
|---|---|
| event $A$ | event $\{X_1 < X_2\}$ |
| random variable $Y$ | random variable $X_2$ |
| event $\{Y = y\}$ | event $\{X_2 = t\}$ |
| event $\{A \mid Y = y\}$ | event $\{X_1 < t\}$ |
| $P(A \mid Y = y)$ | $P(X_1 < t) = 1 - e^{-\lambda_1 t}$ |
| density $f_Y(y)$ | density $f_{X_2}(t) = \lambda_2 e^{-\lambda_2 t}$ |

In a similar way,

$$P(X_1 > X_2) = \frac{\lambda_2}{\lambda_1 + \lambda_2}. \tag{64}$$

As expected, $P(X_1 < X_2) + P(X_1 > X_2) = 1$. Notice that as $X_1$ and $X_2$ are continuous-time random variables, the probability that they are equal to each other is equal to zero.

### 1.10.3 Relationship between Exponential and Geometric Random Variables

We have learnt that the geometric random variable is the only discrete random variable that is memoryless. We also know that the only memoryless continuous random variable is the exponential random variable. These facts indicate an interesting relationship between the two. Let $X_{exp}$ be an exponential random variable with parameter $\lambda$ and let $X_{geo}$ be a geometric random variable with parameter $p$.

Let $\delta$ be an "interval" size used to discretize the continuous values that $X_{exp}$ takes, and we are interested to find $\delta$ such that

$$F_{X_{exp}}(n\delta) = F_{X_{geo}}(n), \quad n = 1, 2, 3, \ldots.$$

To find such a $\delta$, it is more convenient to consider the complementary distributions. That is, we aim to find $\delta$ that satisfies

$$P(X_{exp} > n\delta) = P(X_{geo} > n), \quad n = 1, 2, 3, \ldots,$$

or

$$e^{-\lambda n\delta} = (1-p)^n, \quad n = 1, 2, 3, \ldots,$$

or

$$e^{-\lambda\delta} = 1 - p.$$

Thus,

$$\delta = \frac{-\ln(1-p)}{\lambda} \quad \text{and} \quad p = 1 - e^{-\lambda\delta}.$$

We can observe that as the interval size $\delta$ approaches zero the probability of success $p$ also approaches zero, and under these conditions the two distributions approach each other.

### 1.10.4 Hyper-Exponential

Let $X_i$ for $i = 1, 2, 3, \ \ldots, \ k$ be $k$ independent exponential random variables with parameters $\lambda_i$, $i = 1, 2, 3, \ \ldots, \ k$, respectively. Let $p_i$ for $i = 1, 2, 3, \ \ldots, \ k$ be $k$ nonnegative real numbers such that $\sum_{i=1}^{k} p_i = 1$. A random variable $X$ that is equal to $X_i$ with probability $p_i$ is called Hyper-exponential. By the Law of total probability, its density is

$$f_X(x) = \sum_{i=1}^{k} p_i f_{X_i}(x). \tag{65}$$

### 1.10.5 Erlang

A random variable $X$ has Erlang distribution with parameters $\lambda$ (positive real) and $k$ (positive integer) if its density is given by

$$f_X(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}. \tag{66}$$

This distribution is named after the Danish mathematician Agner Krarup Erlang (1878 – 1929) who was the originator of queueing theory and teletraffic. The term *Erlang* is used for other concepts in the field as we will see in later chapters.

**Homework 1.19**

Let $X_i$, $i = 1, \ 2, \ \ldots, \ k$, be $k$ independent exponentially distributed random variables each with parameter $\lambda$, prove by induction that the random variable $X$ defined by the sum $X = \sum_{i=1}^{k} X_i$ has Erlang distribution with parameters $k$ and $\lambda$. In other words, $f_X(x)$ of (66) is a k-fold convolution of $\lambda e^{-\lambda x}$. $\square$

**Homework 1.20**

Let $X_1$ and $X_2$ be independent and Erlang distributed random variables with parameters $(k, \lambda_1)$ and $(k, \lambda_2)$, respectively. Find the probability of $P(X_1 < X_2)$.

**Guide**

Define the probability of success $p$ as

$$p = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

the probability of failure by $1 - p$. Then observe that the event $\{X_1 < X_2\}$ is equivalent to the event: having $k$ successes before having $k$ failures, so the required probability is the probability of a Pascal random variable $Y$ with parameters $p$ and $k$ to be less or equal to $k - 1$. This observation is explained as follows. Consider individual points that occur randomly on the time axis starting from $t = 0$. The points are of two types 1 and 2. The first type 1 point occurs at time $t_1(1)$ where $t_1(1)$ is exponentially distributed with parameter $\lambda_1$. The second type 1 point occurs at time $t_2(1)$ where $t_2(1) - t_1(1)$ is also exponentially distributed with parameter $\lambda_1$, and in general, the $n$th type 1 point occurs at time $t_n(1)$ where $t_n(1) - t_{n-1}(1)$ is exponentially distributed with parameter $\lambda_1$. Observe that $t_k(1)$ is a sum of $k$ exponentially distributed random variables with parameter $\lambda_1$ and therefore it follows an Erlang distribution with parameters $(k, \lambda_1)$ exactly as $X_1$. Equivalently, we can construct the time process of type 2 points where $t_1(2)$ and all the inter-point times $t_n(2) - t_{n-1}(2)$ are exponentially distributed with parameter $\lambda_2$. Then $t_k(2)$ follows an Erlang distribution with parameters $(k, \lambda_2)$ exactly as $X_2$.

Accordingly, the event $\{X_1 < X_2\}$ is equivalent to the event $\{t_k(1) < t_k(2)\}$. Now consider a traveler that travels on the time axis starting from time 0. This traveler considers type 1 points as successes and type 2 points as failures, where $p$ is a probability that the next point of type 1 (a success) and and $1 - p$ is the probability the next point of type 2 (a failure). The event $\{t_k(1) < t_k(2)\}$ is equivalent to having $k$ successes before having $k$ failures, which lead to the observation that $P(X_1 < X_2)$ is the probability of a Pascal random variable $Y$ with parameters $p$ and $k$ to be less or equal to $k - 1$.

Based on this observation the probability $P(X_1 < X_2)$ is obtained by (34) as

$$P(X_1 < X_2) = F_Y(j) = P(Y \leq k - 1) = 1 - I_p(k, k). \tag{67}$$

where $I_p(\cdot, \cdot)$ is the regularized incomplete beta function.

□

**Homework 1.21**

Again, consider the two independent and Erlang distributed random variables random variables $X_1$ and $X_2$ with parameters $(k, \lambda_1)$ and $(k, \lambda_2)$, respectively. Assume $\lambda_1 < \lambda_2$. Investigate the probability $P(X_1 < X_2)$ as $k$ approaches infinity. Use numerical, intuitive and rigorous approaches.

□

### 1.10.6 Hypo-Exponential

Let $X_i$, $i = 1, 2, \ldots, k$ be $k$ independent exponentially distributed random variables each with parameters $\lambda_i$, respectively. The random variable $X$ defined by the sum $X = \sum_{i=1}^{k} X_i$ is called hypo-exponential. In other words, the density of $X$ is a convolution of the $k$ densities $\lambda_i e^{-\lambda_i x}$, $i = 1, 2, \ldots, k$. The Erlang distribution is a special case of hypo-exponential when all the $k$ random variables are identically distributed.

### 1.10.7 Gaussian

A continuous random variable, which commonly used in many applications, is the Gaussian (also called Normal) random variable. We say that the random variable $X$ has Gaussian distribution with parameters $m$ and $\sigma^2$ if its density is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/2\sigma^2} \qquad -\infty < x < \infty. \tag{68}$$

This density is symmetric and bell shaped.

The wide use of the Gaussian random variable is rooted in the so-called **The central limit theorem**. This theorem is the most important result in probability theory. Loosely speaking, it says that the sum of a large number of independent random variables (under certain conditions) has Gaussian (normal) distribution. This is also true if the distribution of these random variables are very different from Gaussian. This theorem explains why so many populations in nature and society have bell shaped Gaussian histograms, and justifies the use of the Gaussian distribution as their model. In Section 1.18 we will further discuss the central limit theorem and demonstrate its applicability to the telecommunication link dimensioning problem in Section 1.19.

### 1.10.8 Pareto

Another continuous random variable often used in telecommunication modelling is the **Pareto** random variable. This random variable, for a certain parameter range, it can be useful in modelling lengths of data bursts in data and multimedia networks [1]. We choose to define the Pareto random variable with parameters $\gamma$ and $\delta$ by its complementary distribution function which is given by

$$P(X > x) = \begin{cases} \left(\frac{x}{\delta}\right)^{-\gamma}, & x \geq \delta \\ 1, & \text{otherwise.} \end{cases}$$

Here $\delta > 0$ is the scale parameter representing a minimum value for the random variable, and $\gamma > 0$ is the shape parameter of the Pareto distribution.

**Homework 1.22**

Write a computer program that generates a sequence of 100 random deviates from a Pareto distribution with $\gamma = 1.2$ and $\delta = 4$. $\square$

## 1.11 Moments and Variance

### 1.11.1 Mean (or Expectation)

The *mean* (or the expectation) of a discrete random variable is defined by

$$E[X] = \sum_{\{n:P(n)>0\}} nP_X(n). \tag{69}$$

This definition can be intuitively explained by the concept of *limiting relative frequencies* illustrated by the following example. Let an outcome of an experiment be a person height. Let $p_i$ be the probability that a person's height is $i$ cm. Consider a sample of $N$ people. Each one of them reports his/her height, rounded to the nearest cm. Let $n_i$ be the number of people reported a height of $i$ cm. These $n_i$ values can be graphically presented in an histogram. Then the relative frequency $n_i/N$ approximates $p_i$. This approximation becomes more and more accurate as $N$ increases. This approximation is consistent with the requirement

$$\sum_i p_i = 1.$$

If we set $p_i = n_i/N$, then since $\sum_i n_i = N$, we obtain

$$\sum_i p_i = 1.$$

Then the average height is given by

$$\frac{\sum_i in_i}{N}.$$

For large $N$, we set

$$p_i = \frac{n_i}{N}$$

. Substituting in the above, we obtain

$$\text{The average height} = \sum_i ip_i$$

which is the definition of mean.

This is related to the Weak Law of Large Numbers discussed below in Section 1.18.

**Homework 1.23**

Consider a discrete random variable $X$ that takes the values 1, 2 and 3 with the following probabilities $P(X = 1) = 0.5$, $P(X = 2) = 0.3$, $P(X = 3) = 0.2$, and it takes all other values with probability zero. Find the mean of $X$.

**Answer**

$E[X] = 1.7$. □

**Homework 1.24**

In an American Roulette game, players place their bets, and a ball lends on one of 38 numbered pockets in a turning wheel with the same probability of 1/38. In all the questions about Roulette, we assume that the number that the ball lends on is independent of numbers it lends on in any other previous or future games (wheel turnings). A player has various betting options to choose from. Any of these options can be represented by a set of $n$ numbered pockets. If the player bet $B$ dollars and if the ball lends on any of the numbers in the chosen set, the payout given to the player is given by

$$\frac{1}{n}(36-n)B = \frac{36B}{n} - B \quad \text{dollars.}$$

If the ball does not lend in any of the chosen numbers the player looses the $B$ dollars bet.

For example, if $B = 1$ and $n = 18$, the player win one dollar payout. A roulette has 18 red pockets and 18 black pockets, and there are two additional green pockets for zero and double-zero. One way to bet on a chosen set on 18 numbers is to bet on the red (on all 18 red colored pockets).

Find the mean return for the player who bets $B = 1$ in one game for a relevant range of $n$ values.

For more information on Roulette and on the relevant range of $n$ values, and on the mean return, see:

http`http://en.wikipedia.org/wiki/Roulette`

and the following the youtube link:

http`https://www.youtube.com/watch?v=pYKcPL0N9QQ`

**Guide**

For the case $B = 1$ and and $n = 18$ with probability 18/38 the player wins one dollar and with probability 20/38, the player looses one dollar. Let the random variable $X$ represents the return in one game. Therefore, the mean return is given by

$$E[X] = (1)\frac{18}{38} + (-1)\frac{20}{38} = -0.052631579. \tag{70}$$

Repeat this calculation for various $n$ values. □

Equivalently, the mean of a continuous random variable is defined as

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx. \tag{71}$$

A useful expression for the mean of a continuous nonnegative random variable $Z$ (i.e. a random variable $Z$ with the property that its density $f(z) = 0$ for $z < 0$) is:

$$E[Z] = \int_{0}^{\infty} P(Z > z) dz = \int_{0}^{\infty} [1 - F_Z(z)] dz. \tag{72}$$

The discrete equivalence of the latter is:

$$E[Z] = \sum_{n=0}^{\infty} P(Z > n) = \sum_{n=0}^{\infty} [1 - F_Z(n)]. \tag{73}$$

**Homework 1.25**

Show (72) and (73).

**Guide**

For the case of discrete nonnegative random variables, by (69),

$$E[Z] = \sum_{n=1}^{\infty} nP_Z(n).$$

This can be written as

$$E[Z] = \begin{cases} P_Z(1) & +P_Z(2) & +P_Z(3) & +P_Z(4) & +\ldots \\ & +P_Z(2) & +P_Z(3) & +P_Z(4) & +\ldots \\ & & +P_Z(3) & +P_Z(4) & +\ldots \\ & & & +P_Z(4) & +\ldots \\ & & & & +\ldots \end{cases}$$

Therefore,

$$E[Z] = \sum_{n=1}^{\infty} P_Z(n) + \sum_{n=2}^{\infty} P_Z(n) + \sum_{n=3}^{\infty} P_Z(n) + \sum_{n=4}^{\infty} P_Z(n) + \quad \ldots .$$

and (73) follows.

Another way to write this proof is [90]

$$\begin{aligned} E[Z] &= \sum_{n=1}^{\infty} nP_Z(n) \\ &= \sum_{n=1}^{\infty} \left( \sum_{j=1}^{n} 1 \right) P_Z(n) \\ &= \sum_{j=1}^{\infty} \sum_{n=j}^{\infty} 1 P_Z(n) \\ &= \sum_{n=0}^{\infty} P(Z > n). \end{aligned}$$

To show (72) use an analogous approach (defining $y = f_Z(z)$) as follows [90]

$$\begin{aligned} E[Z] &= \int_0^\infty z f_Z(z)dz \\ &= \int_0^\infty \left(\int_0^z 1dy\right) f_Z(z)dz \\ &= \int_0^\infty 1 \left(\int_y^\infty f_Z(z)dz\right) dy \\ &= \int_0^\infty P(Z > y)dy \\ &= \int_0^\infty P(Z > z)dz. \end{aligned}$$

Check carefully and understand all the steps and operations that led to the above proofs.

□

**Homework 1.26**

Let $X_1, X_2, X_3, \ldots X_k$ be a sequence of $k$ independent random variables having a uniform (0,s) distribution. Let $X = \min\{X_1, X_2, X_3, \ldots, X_k\}$. Prove that

$$E[X] = \frac{s}{k+1}.$$

**Hint:** Use (57) and (72). □

As mentioned above, a function of a random variable is also a random variable. The mean of a function of random variables denoted $g(\cdot)$ by

$$E[g(X)] = \sum_{\{k:P_X(k)>0\}} g(k)P_X(k) \tag{74}$$

for a discrete random variable and

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx \tag{75}$$

for a continuous random variable. If $a$ and $b$ are constants then for a random variable $X$ (either discrete or continuous) we have:

$$E[aX] = aE[X], \tag{76}$$

$$E[X - b] = E[X] - b, \tag{77}$$

and

$$E[aX - b] = aE[X] - b. \tag{78}$$

### 1.11.2 Moments

The $n$**th moment** of the random variable $X$ is defined by $E[X^n]$. Substituting $g(X) = X^n$ in (74) and in (75), the $n$th moment of $X$ is given by:

$$E[X^n] = \sum_{\{k:P_X(k)>0\}} k^n P_X(k) \tag{79}$$

for a discrete random variable and

$$E[X^n] = \int_{-\infty}^{\infty} x^n f_X(x)dx \tag{80}$$

for a continuous random variable. Similarly, the $n$th central moment of random variable $X$ is defined by $E[(X - E[X])^n]$. Substituting $g(X) = (X - E[X])^n$ in (74) and in (75), the $n$**th central moment** of $X$ is given by:

$$E[(X - E[X])^n] = \sum_{\{k:P(k)>0\}} (k - E[X])^n P_X(k) \tag{81}$$

for a discrete random variable and

$$E[(X - E[X])^n] = \int_{-\infty}^{\infty} (x - E[X])^n f_X(x)dx \tag{82}$$

for a continuous random variable. By definition the first moment is the mean.

### 1.11.3 Variance

The second central moment is called the **variance**. It is defined as

$$Var[X] = E[(X - E[X])^2]. \tag{83}$$

The variance of a random variable $X$ is given by

$$Var[X] = \sum_{\{k:P(k)>0\}} (k - E[X])^2 P_X(k) \tag{84}$$

if $X$ is discrete, and by

$$Var[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x)dx \tag{85}$$

if it is continuous.

By (83) we obtain

$$Var[X] = E[(X - E[X])^2] = E[X^2 - 2XE[X] + (E[X])^2] = E[X^2] - (E[X])^2. \tag{86}$$

To understand the latter, notice that

$$E[X^2 - 2XE[X] + (E[X])^2] = E[X^2] - E[2XE[X]] + E[(E[X])^2] = E[X^2] - 2E[X]E[X] + (E[X])^2$$

**Homework 1.27**

Consider two independent and identically distributed random variables $X$ and $Y$ that obey the following probability function:

$$P(X = i) = \begin{cases} 1/5 & \text{for } \ i = -2, -1, 0, 1, \text{and } 2 \\ 0 & \text{otherwise.} \end{cases} \quad (87)$$

Let $Z = |X|$, $U = \max(X, Y)$ and $V = \min(X, Y)$. Find the probability function, the mean and variance on $Z, U$, and $V$. □

**Homework 1.28**

For random variable $X$ and constant $c$, show the following:

$$Var[X + c] = Var[X]$$

and

$$Var[cX] = c^2 Var[X].$$

**Guide**

$Var[X + c] = E[(X + c)^2] - (E[X + c])^2 = E[X^2 + 2Xc + c^2] - (E[X])^2 - 2cE[X] - c^2$
$= E[X^2] - (E[X])^2 = Var[X]$

and

$Var[cX] = E[(cX)^2] - (E[cX])^2 = c^2 E[X^2] - c^2(E[X])^2 = c^2 Var[X].$

□

While the mean provides the average according to the limiting relative frequencies concept, the variance is a measure of the level of variation of the possible values of the random variable. Another measure of such variation is the **standard deviation** denoted $\sigma_X$, or simply $\sigma$, and defined by

$$\sigma_X = \sqrt{Var[X]}. \quad (88)$$

Hence the variance is often denoted by $\sigma^2$.

Notice that the first central moment $E[x - E[X]]$ is not very useful because it is always equal to zero, the second central moment $E[(x - E[X])^2]$, which is the variance, and its square root, the standard deviation, are used for measuring the level of variation of a random variable.

The standard deviation of the random variable $X$, namely $\sigma_X$, has always the same units as the random variable $X$. For example, if the random variable $X$ represents the number of bytes in an IP packet then $\sigma_X$ is also in bytes. If $X$ represents the service time in seconds of an IP packet, then, $\sigma_X$ is also in seconds. Since $Var[X] = \sigma_X^2$, the units of the variance are in the units of $\sigma_X^2$ or $X^2$. Thus, for the above two examples, of being in bytes or in seconds, the units of variance are bytes$^2$ or in seconds$^2$, respectively.

The mean of sum of random variables is always the sum of their means, namely,

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] \tag{89}$$

but the variance of sum of random variables is not always equal to the sum of their variances. It is true for independent random variables. That is, if the random variables $X_1,\ X_2,\ X_3,\ \ldots,\ X_n$ are independent, then

$$Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} Var[X_i]. \tag{90}$$

Also, if $X_1,\ X_2,\ X_3,\ \ldots,\ X_n$ are independent, then

$$E[\Pi_{i=1}^{n} X_i] = \Pi_{i=1}^{n} E[X_i]. \tag{91}$$

### 1.11.4 Conditional Mean and the Law of Iterated Expectation

In many applications it is useful to use the concept of **Conditional Expectation (or Mean)** to derive moments of unknown distributions.

Let $E[X \mid Y]$ be the conditional expectation of random variable $X$ given the event $\{Y = y\}$ (random variable $Y$ is equal to $y$) for each relevant value of $y$.

The conditional expectation of two discrete random variables is defined by

$$E[X \mid Y = j] = \sum_{i} iP(X = i \mid Y = j). \tag{92}$$

If $X$ and $Y$ are continuous, their conditional expectation is defined as

$$E[X \mid Y = y] = \int_{x=-\infty}^{\infty} x f_{X|Y}(x \mid y) dx. \tag{93}$$

It is important to realize that $E[X \mid Y]$ is a random variable which is a function of the random variable $Y$. Therefore, if we consider its mean (in the case that $X$ and $Y$ are discrete) we obtain

$$\begin{aligned} E_Y[E[X \mid Y]] &= \sum_{j} E[X \mid Y = j] P(Y = j) \\ &= \sum_{j} \sum_{i} iP(X = i \mid Y = j) P(Y = j) \\ &= \sum_{i} i \sum_{j} P(X = i \mid Y = j) P(Y = j) \\ &= \sum_{i} iP(X = i) = E[X]. \end{aligned} \tag{94}$$

Thus, we have obtained the following formula for the mean $E[X]$

$$E[X] = E_Y[E[X \mid Y]]. \tag{95}$$

The latter is called *the law of iterated expectation*. It is also known by many other names including the law of iterated expectations, the law of total expectation, and Adam's law.

The law of iterated expectation also applies to continuous random variables. In this case we have:

$$\begin{aligned} E_Y[E[X \mid Y]] &= \int_{y=-\infty}^{\infty} E[X \mid Y=y] f_Y(y) dy \\ &= \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} x f_{X|Y}(x \mid y) dx f_Y(y) dy \\ &= \int_{x=-\infty}^{\infty} x \int_{y=-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y) dy dx \\ &= \int_{x=-\infty}^{\infty} x f_X(x) dx = E[X]. \end{aligned}$$

**Homework 1.29**

Show that $E[X] = E_Y[E[X \mid Y]]$ holds also for the case where $X$ is discrete and $Y$ is continuous and vice versa. □

Note that $P(X = x \mid Y = y)$ is itself a random variable that is a function of the values $y$ taken by random variable $Y$. Therefore, by definition $E_Y[P(X = x \mid Y = y)] = \sum_y P(X = x \mid Y = y)P(Y = y)$ which lead to another way to express the Law of Total Probability:

$$P_X(x) = E_Y[P(X = x \mid Y = y)]. \tag{96}$$

### 1.11.5 Conditional Variance and the Law of Total Variance

Define the **Conditional Variance** as

$$Var[X \mid Y] = E[(X - E[X \mid Y])^2 \mid Y]. \tag{97}$$

This gives rise to the following useful formula for the variance of a random variable known as *The Law of Total Variance*:

$$Var[X] = E[Var[X \mid Y]] + Var[E[X \mid Y]]. \tag{98}$$

Due to its form, it is also called *EVVE*, or *Eve's law*. In this book, we will refer to it as EVVE.

To show EVVE, we recall (86): $Var[X] = E[X^2] - (E[X])^2$, and (95): $E[X] = E_Y[E[X \mid Y]]$, we obtain

$$Var[X] = E[E[X^2 \mid Y]] - (E[E[X \mid Y]])^2. \tag{99}$$

Then using $E[X^2] = Var[X] + (E[X])^2$ gives

$$Var[X] = E[Var[X \mid Y] + (E[X \mid Y])^2] - (E[E[X \mid Y]])^2 \tag{100}$$

or

$$Var[X] = E[Var[X \mid Y]] + E[E[X \mid Y]]^2 - (E[E[X \mid Y]])^2. \tag{101}$$

Now considering again the formula $Var[X] = E[X^2] - (E[X])^2$, but instead of the random variable $X$ we put the random variable $E[X \mid Y]$, we obtain

$$Var[E[X \mid Y]] = E[E[X \mid Y]^2] - (E[E[X \mid Y]])^2, \tag{102}$$

observing that the right-hand side of (102) equals to the last two terms in the right-hand side of (101), we obtain EVVE.

To illustrate the use of conditional mean and variance, consider the following example. Every second the number of Internet flows that arrive at a router, denoted $\phi$, has mean $\phi_e$ and variance $\phi_v$. The number of packets in each flow, denoted $\varsigma$, has mean $\varsigma_e$ and variance $\varsigma_v$. Assume that the number of packets in each flow and the number of flows arriving per second are independent. Let $W$ the total number of packets arriving at the router per second which has mean $W_e$ and variance $W_v$. Assume $W = \varsigma\phi$. The network designer, aiming to meet certain quality of service (QoS) requirements, makes sure that the router serves the arriving packets at the rate of $s_r$ per second, such that $s_r = W_e + 4\sqrt{W_v}$. To compute $s_r$ one needs to have the values of $W_e$ and $W_v$. Because $\phi$ and $\varsigma$ are independent $E[W|\phi] = \phi\varsigma_e$ and by (95)

$$W_e = E[W] = E[E[W|\phi]] = E[\phi]E[\varsigma] = \phi_e\varsigma_e.$$

Note that the relationship

$$W_e = \phi_e\varsigma_e \tag{103}$$

is also obtained directly by (91). In fact, the above proves (91) for the case of two random variables.

Also by EVVE,

$$Var[W] = E[Var[W|\phi]] + Var[E[W|\phi]] = \varsigma_v E[\phi^2] + (\varsigma_e)^2 Var[\phi].$$

Therefore

$$W_v = \phi_v\varsigma_v + \varsigma_v\phi_e^2 + \phi_v\varsigma_e^2. \tag{104}$$

**Homework 1.30**

1. Provide detailed derivations of Equations (103) and (104) using (95) and (98).

**Guide**

Observe that because $W = \phi\varsigma$, derivation of $Var[W|\phi]$ is obtained by $Var[\phi\varsigma]$ when $\phi$ is a given constant, which is simplified by

$$Var[\phi\varsigma] = \phi^2 Var[\varsigma] = \phi^2\varsigma_v.$$

Therefore,

$$Var[W|\phi] = \phi^2\varsigma_v.$$

Now it is clear that $Var[W|\phi]$ is a random variable which is a function of the random variable $\phi$ (where $\varsigma_v$ is clearly a constant). Then we can obtain the mean of $Var[W|\phi]$, by

$$E[Var[W|\phi]] = E[\phi^2\varsigma_v] = \varsigma_v E[\phi^2].$$

The last equality is based on Equation (76).

To obtain $E[\phi^2]$, notice that

$$\phi_v = Var[\phi] = E[\phi^2] - (E[\phi])^2 = E[\phi^2] - \phi_e^2.$$

Therefore,

$$E[\phi^2] = \phi_v + \phi_e^2.$$

Similarly, the derivation of $E[W|\phi]$ is obtained by $E[\phi\varsigma]$ when $\phi$ is a given constant, which is simplified by

$$E[\phi\varsigma] = \phi E[\varsigma] = \phi\varsigma_e.$$

Accordingly,

$$E[W|\phi] = \phi\varsigma_e.$$

In a similar way to the previous elaboration, it is clear that $E[W|\phi]$ is a random variable which is a function of the random variable $\phi$ (where $\varsigma_e$ is a constant). Then, we can obtain the variance of $E[W|\phi]$, by

$$Var[E[W|\phi]] = Var[\phi\varsigma_e] = \varsigma_e^2 Var[\phi] = \varsigma_e^2\phi_v.$$

Now put it all together, and obtain Equation (104).

2. Derive Equations (103) and (104) in a different way, considering the independence of the number of packets in each flow and the number of flows arriving per second. □

**Homework 1.31**

A traveler considers buying a travel insurance and would like to assess the statistics of the cost that may be associated with unforseen insurable incident or incidents that may be associated with his/her planned travel. The probability that the traveler experiences incident(s) is $p_{inc}$. That is, the probability that nothing happens during travel is $1 - p_{inc}$ and the probability that some incidents (e.g. accident, theft, illness etc.) happen is $p_{inc}$. If incidents happen the total unforeseen cost of the incidents has mean of $M_{inc}$ and standard deviation of $\sigma_{inc}$. In this document, the term "incidents" refers to one or more incidents. What is the mean loss and the standard deviation of the loss incurred by the traveler?

**Guide**

Let $Y$ be a random variable that takes the value 1 if accidents happen, and 0 if no incidents happen. Accordingly, $Y$ is Bernoulli distributed with parameter $p_{inc}$. That is, $P(Y = 1) = p_{inc}$, and $P(Y = 0) = 1-p_{inc}$. Let $X$ be a random variable representing the total cost of the incidents. We know that $X$ depends on $Y$. We need to obtain $E[X]$ and $Var[X]$, which are obtainable by the law of iterated expectation and EVVE, respectively.

Therefore, according to these definitions and the given parameters, we obtain,

$$E_X[X \mid Y = 1] = M_{inc},$$

and

$$E_X[X \mid Y = 0] = 0,$$

which imply

$$E_X[X \mid Y] = M_{inc}Y.$$

The latter makes it very clear that $E_X[X \mid Y]$ is a random variable which is a function of the random variable $Y$.

In addition we obtain,

$$Var_X[X \mid Y = 1] = \sigma_{inc}^2,$$

and

$$Var_X[X \mid Y = 0] = 0.$$

Also, since $Y$ is a Bernoulli random variable with parameter $p_{inc}$, then its variance is equal to $(1 - p_{inc})p_{inc}$.

Then by the law of iterated expectation,

$$\begin{aligned} E[X] &= E_Y[E[X \mid Y]] \\ &= P(Y = 1)E_X[X \mid Y = 1] + P(Y = 0)E_X[X \mid Y = 0] \\ &= p_{inc}M_{inc} + 0 \\ &= p_{inc}M_{inc}, \end{aligned}$$

and by EVVE,

$$\begin{aligned} Var[X] &= E[Var[X \mid Y]] + Var[E[X \mid Y]] \\ &= P(Y = 1)Var[X \mid Y = 1] + P(Y = 0)Var[X \mid Y = 0] + Var[M_{inc}Y] \\ &= p_{inc}\sigma_{inc}^2 + 0 + M_{inc}^2 p_{inc}(1 - p_{inc}) \\ &= p_{inc}\sigma_{inc}^2 + M_{inc}^2 p_{inc}(1 - p_{inc}). \end{aligned}$$

□

**Homework 1.32**

Now consider the following special case for the previous problem: $p_{inc} = 0.04$, $\sigma_{inc} = 4000$ dollars, and $M_{inc} = 3000$ dollars. Find the mean, variance and standard deviation of the cost due to unforseen insurable incidents.

**Answers**

mean: 120 dollars; variance: 985600 [dollar$^2$]; standard deviation: 992.77 dollars (rounded to the nearest dollar). □

**Homework 1.33**

Consider the above two problems and assume that if incidents occur, and the traveler does not have insurance, then the traveler must pay the cost incurred as a result of the incidents. Otherwise, if the traveler has insurance, the insurance company pays it fully. Also assume that the event that incidents happen to a traveler is independent of whether or not the traveler buys insurance. In this way, if a traveler buys insurance, the cost to the traveller has a mean that is equal to the insurance premium and a zero variance. John is a traveler that considers buying a travel insurance. The insurance premium costs 200 dollars, but considering his time associated with claiming back the insurance, he is willing to buy travel insurance only if the mean plus half the standard deviation of the total costs of the insurable incidents is higher than 300 dollars. Will he buy or not buy the travel insurance?

**Guide**

To answer this question, notice that the mean cost is 120 dollars and half the standard deviation is $992.77/2 = 496.39$ dollars, so mean + $0.5 \times$ standard deviation is equal to 616.39 dollars which is more than 300 dollars, so he will buy the travel insurance.

□

**Homework 1.34**

Now consider an insurance company that has sold $N_p$ insurance policies as described in the above three problems. Assume $N_p = 1,000$. Also assume that the incidents that occur for different customers are independent. What is the mean and standard deviation of the total claims payouts of the insurance company? Compare them to $200N_p$ which is the income generated by the premium paid by the customers. What is the mean and standard deviation of the difference between the income generated by the premium paid by the customers, namely $200N_p$, and the total claims payouts of the insurance company? What do you think is the distribution of this difference? What is the probability that this difference is negative? What is the probability is less than half of its mean?

**Answers**

The mean of the total claims payouts is $p_{inc}M_{inc}N_p = 0.04 \times 3000 = 120,000$

The variance of the total claims payouts is $N_pVar[X] = 1,000 \times 985,600 = 985,600,000$.

The standard deviation of the total claims payouts is $\sqrt{985,600,000} = 31,394.267$ dollars.

Notice that the mean + $0.5\times$ standard deviation of the payouts is $120,000 + 31,394.267/2 = 135,697.13$ dollars is far smaller than the income of $200N_p = 200,000$ dollars.

The mean of this difference is $N_p(200 - E[X]) = 1,000(200 - 120) = 80,000$ dollars.

The variance of the difference is $N_p Var[X] = 1,000 \times 985,600 = 985,600,000$ dollars$^2$.

The standard deviation of the difference is $\sqrt{985,600,000} = 31,394.267$ dollars.

Notice that the results for the variance and standard deviation of the difference are the same as the results for those of the total claims. This is because

$$Var[X + c] = Var[X].$$

Based on the central limit theorem, the distribution of the difference between income and payouts is approximately normal with the above mean and standard deviation.

The probability that this difference is negative is approximately equal to 0.005413 using the Excel function: =NORM.DIST(0,80000,31394.267,TRUE).

The probability that this difference is less than 40,000 is approximately equal to 0.1013 using the Excel function: =NORM.DIST(40000,80000,31394.267,TRUE).

□

**Homework 1.35**

Now consider the case, $N_p = 10,000$, and answer all the questions of the previous Problem. Then compare the two cases and discuss economy of scale implications.

## Answers

The mean of this difference is $N_p(200 - E[X]) = 10,000(200 - 120) = 800,000$ dollars.

The variance of the difference is $N_p Var[X] = 10,000 \times 985,600 = 9,856,000,000$ dollars$^2$.

The standard deviation of the difference is $\sqrt{9,856,000,000} = 99,277.39$ dollars.

Based on the central limit theorem, again the distribution of the difference between income and payouts is approximately normal with the above mean and standard deviation.

The probability that this difference is negative is approximately equal to $3.9 \times 10^{-16}$ using the Excel function: =NORM.DIST(0,800000,99277.39,TRUE).

The probability that this difference is less than 40,000 is approximately equal to $2.8 \times 10^{-5}$ using the Excel function: =NORM.DIST(400000,800000,99277.39,TRUE).

We observe that for the larger insurance company with 10,000 customers, it is easier to guarantee that the difference between income and payouts will not be negative and will be over half its mean value - in this case, over 40 dollars per policy. Notice also that the payouts are mean $\pm$ a certain constant $\times$ the standard deviation. Therefore the standard deviation to mean ratio can be viewed as a measure of the risk or variability of income. Notice that in the case of $N_p = 1,000$, this ration is equal to 0.39, while in the case of $N_p = 10,000$, it is equal to approx 0.011. Which further demonstrate a far lower risk in the case of the larger company.

□

**Homework 1.36**

In insurance, the term *loss ratio* is the ratio of the total payouts in claims to the total income from premiums. Find the mean and standard deviations of the loss ratio in the two cases of $N_p = 1,000$ and $N_p = 10,000$.

**Answers**

In the case $N_p = 1,000$, the mean of the loss ratio is

$$\frac{1,000 \times 120}{1,000 \times 200} = 0.6;$$

the variance of the loss ratio is

$$\frac{985,600,000}{(1,000 \times 200)^2} = 0.02464;$$

and the standard deviation is

$$\sqrt{0.02464} = 0.157.$$

In the case $N_p = 10,000$, the mean of the loss ratio is

$$\frac{10,000 \times 120}{10,000 \times 200} = 0.6;$$

the variance of the loss ratio is

$$\frac{9,856,000,000}{(10,000 \times 200)^2} = 0.002464;$$

and the standard deviation is

$$\sqrt{0.002464} = 0.05.$$

□

## 1.12 Mean and Variance of Specific Random Variables

If $X$ is a Bernoulli random variable with parameter $p$, its mean and variance are obtained by

$$E[X] = 0 \times P(X = 0) + 1 \times P(X = 1) = 0(1 - p) + 1(p) = p$$

and

$$Var[X] = E[X^2] - (E[X])^2 = p - p^2 = p(1 - p),$$

or

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= (0 - p)^2 P(X = 0) + (1 - p)^2 P(X = 1) \\ &= p^2(1 - p) + (1 - p)^2 p \\ &= p(1 - p). \end{aligned}$$

Since the binomial random variable is a sum of $n$ IID Bernoulli random variables, its mean and its variance are $n$ times the mean and variance of these Bernoulli random variables, respectively. Notice also that by letting $p \to 0$, and $np \to \lambda$, both the mean and the variance of the binomial random variable approach $\lambda$, which is the value of both the mean and variance of the Poisson random variable.

**Homework 1.37**

Consider an experiment of tossing a die with 6 sides. Assume that the die is fair, i.e., each side has the same probability (1/6) to occur. Consider a random variable $X$ that takes the value $i$ if the outcome of the toss is $i$, for $i = 1, 2, 3, \cdots, 6$. Find $E[X]$, $Var[X]$ and $\sigma_X$.

**Answers**

$E[X] = 3.5$; $E[X^2] = 15.16666667$; $Var[X] = 2.916666667$; $\sigma_X = 1.707825128$. □

**Homework 1.38**

Consider the previous problem and plot the probability function, distribution function and the complementary distribution function of $X$. □

**Homework 1.39**

Let $X$ be a geometric random variable with parameter $p$. Derive its mean and variance using the law of iterated expectation and EVVE.

**Guide**

Consider a Bernoulli random variable $Y$ that takes the value 1 (success) with probability $p$ and the value 0 (failure) with probability $1 - p$. Let $Y$ be the first Bernoulli trial associated with the geometric random variable $X$. Therefore, $E[X|Y = 1] = 1$ and $E[X|Y = 0] = 1 + E[X]$. The former represents the case of having a success in the first trial, so $X = 1$, and the second represents a failure in the first trial, so due to the memoryless property of the geometric random variable the expected number of required trials will be one (the first failure) plus the mean of another independent geometric random variable that has mean of $E[X]$.

Then, $E[X|Y]$ can be written as

$$E[X|Y] = 1 + (1 - Y)E[X].$$

Accordingly,

$$E[X] = E_Y E[X|Y] = E[1 + (1 - E[Y])E[X]] = 1 + (1 - p)E[X].$$

Thus,

$$E[X] = \frac{1}{p}.$$

Also,

$$Var[X|Y] = \begin{cases} 0 & \text{if } Y = 1 \\ Var(X) & \text{if } Y = 0. \end{cases}$$

Thus,

$$Var[X|Y] = (1 - Y)Var[X],$$

so

$$E_Y[Var[X|Y]] = Var[X](1 - p).$$

Also,

$$Var[1 + (1 - Y)\frac{1}{p}] = \frac{1}{p^2}p(1 - p).$$

Then, by EVVE,

$$Var[X] = Var[X](1 - p) + \frac{p(1 - p)}{p^2}.$$

Hence,

$$Var[X] = \frac{1 - p}{p^2}.$$

□

Let $Y$ be a geometric random variable of failures with parameter $p$. Obtain its mean and variance.

**Guide**

We know that $Y = X - 1$, where

$$E[X] = \frac{1}{p}.$$

$$Var[X] = \frac{1 - p}{p^2}.$$

Therefore

$$E[Y] = E[X] - 1 = \frac{1}{p} - 1 = \frac{1 - p}{p}.$$

and

$$Var[Y] = Var[X] = \frac{1 - p}{p^2}.$$

□

**Homework 1.41**

Consider a discrete uniform probability function with parameters $a$ and $b$ that takes equal non-zero values for $x = a, a + 1, a + 2, \ldots, b$. Derive its mean $E[X]$ and variance $Var[X]$.

**Guide**

First, show that

$$E[X] = \frac{a+b}{2}.$$

To obtain $Var[X]$, notice first that the variance of this discrete uniform distribution over the $x$ values of $a, a+1, a+2, \ldots b$ has the same value as the variance of a discrete uniform distribution over the $x$ values of $1, 2, \ldots b-a+1$. Then for convenience set $n = b-a+1$ and derive the variance for the probability function

$$P_X(x) = \begin{cases} \frac{1}{n} & \text{if } x = 1, 2, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

Next, by induction on $n$ show that

$$E[X^2] = \frac{(n+1)(2n+1)}{6}.$$

Finally, use the equation $Var[X] = E[X^2] - (E[X])^2$ to show that

$$Var[X] = \frac{n^2-1}{12} = \frac{(b-a+1)^2-1}{12}.$$

□

**Homework 1.42**

Consider an exponential random variable with parameter $\lambda$. Derive its mean and Variance.

**Guide**

Find the mean by

$$E[X] = \int_0^\infty x\lambda e^{-\lambda x} dx.$$

Use integration by parts to show that:

$$E[X] = \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Now notice how (72) simplifies the derivation of the mean:

$$E[X] = \int_0^\infty e^{-\lambda x} dx = \left[\frac{1}{-\lambda} e^{-\lambda x}\right]_0^\infty = \frac{1}{\lambda}.$$

Then use integration by parts to derive the second moment. Understand and verify the following derivations:

$$
\begin{aligned}
E[X^2] &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\
&= \left[-x^2 e^{-\lambda x}\right]_0^\infty + 2\int_0^\infty x e^{-\lambda x} dx \\
&= \left[\left(-x^2 e^{-\lambda x} - \frac{2}{\lambda} x e^{-\lambda x} - \frac{2}{\lambda^2} e^{-\lambda x}\right)\right]_0^\infty \\
&= \frac{2}{\lambda^2}.
\end{aligned}
$$

$$Var[X] = E[X^2] - (E[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

□

**The mean of the Pareto random variable** is given by

$$E[X] = \begin{cases} \infty & \text{if } 0 < \gamma \leq 1 \\ \frac{\delta\gamma}{\gamma - 1} & \gamma > 1. \end{cases} \tag{105}$$

For $0 < \gamma \leq 2$, the variance $Var[X] = \infty$.

The following table provides the mean and the variance of some of the above-mentioned random variables.

| random variable | parameters | mean | variance |
|---|---|---|---|
| Bernoulli | $0 \leq p \leq 1$ | $p$ | $p(1-p)$ |
| geometric | $0 \leq p \leq 1$ | $1/p$ | $(1-p)/p^2$ |
| geometric (of failures) | $0 \leq p \leq 1$ | $(1-p)/p$ | $(1-p)/p^2$ |
| binomial | $n$ and $0 \leq p \leq 1$ | $np$ | $np(1-p)$ |
| Poisson | $\lambda > 0$ | $\lambda$ | $\lambda$ |
| discrete uniform | $a$ and $b$ | $(a+b)/2$ | $[(b-a+1)^2 - 1]/12$ |
| uniform | $a$ and $b$ | $(a+b)/2$ | $(b-a)^2/12$ |
| exponential | $\mu > 0$ | $1/\mu$ | $1/\mu^2$ |
| Gaussian | $m$ and $\sigma$ | $m$ | $\sigma^2$ |
| Pareto | $\delta > 0$ and $1 < \gamma \leq 2$ | $\delta\gamma/(\gamma-1)$ | $\infty$ |

## 1.13 Sample Mean and Sample Variance

If we are given a sample of $n$ realizations of a random variable $X$, denoted $X(1), X(2), \ldots, X(n)$ we will use the **Sample Mean** defined by

$$S_m = \frac{\sum_{i=1}^n X(i)}{n} \tag{106}$$

as an estimator for the mean of $X$. For example, if we run simulation of a queueing system and observe $n$ values of customer delays for $n$ different customers, the Sample Mean will be used to estimate a customer delay.

If we are given a sample of $n$ realizations of a random variable $X$, denoted $X(1), X(2), \ldots, X(n)$ we will use the *Sample Variance* defined by

$$S_v = \frac{\sum_{i=1}^{n}[X(i) - S_m]^2}{n-1} \qquad n > 1 \tag{107}$$

as an estimator for the variance of $X$. The sample standard deviation is then $\sqrt{S_v}$.

**Homework 1.43**

Generate 10 deviates from an exponential distribution of a given mean and compute the Sample Mean and Sample Variance. Compare them with the real mean and variance. Then increase the sample to 100, 1000, ..., 1,000,000. Observe the difference between the real mean and variance and the sample mean and variance. Repeat the experiment for a Pareto deviates of the same mean. Discuss differences. □

## 1.14 Covariance and Correlation

When random variables are positively dependent, namely, if when one of them obtains high values, the others are likely to obtain high value also, then the variance of their sum may be much higher than the sum of the individual variances. This is very significant for bursty multimedia traffic modeling and resource provisioning. For example, let time be divided into consecutive small time intervals, if $X_i$ is the amount of traffic arrives during the $i$th interval, and assume that we use a buffer that can store traffic that arrives in many intervals, the probability of buffer overflow will be significantly affected by the variance of the amount of traffic arrives in a time period of many intervals, which in turn is strongly affected by the dependence between the $X_i$s. Therefore, there is a need to define a quantitative measure for dependence between random variables. Such measure is called the **covariance**. The covariance of two random variables $X_1$ and $X_2$, denoted by $Cov[X_1, X_2]$, is defined by

$$Cov[X_1, X_2] = E[(X_1 - E[X_1])(X_2 - E[X_2])]. \tag{108}$$

Intuitively, by Eq. (108), if high value of $X_1$ implies high value of $X_2$, and low value of $X_1$ implies low value of $X_2$, the covariance is high. By Eq. (108),

$$Cov[X_1, X_2] = E[X_1X_2] - E[X_1]E[X_2]. \tag{109}$$

Hence, by (91), if $X_1$ and $X_2$ are independent then $Cov[X_1, X_2] = 0$. The variance of the sum of two random variables $X_1$ and $X_2$ is given by

$$Var[X_1 + X_2] = Var[X_1] + Var[X_2] + 2Cov[X_1, X_2]. \tag{110}$$

This is consistent with our comments above. The higher the dependence between the two random variables, as measured by their covariance, the higher the variance of their sum, and if they are independence, hence $Cov[X_1, X_2] = 0$, the variance of their sum is equal to the sum of their variances. Notice that the reverse is not always true: $Cov[X_1, X_2] = 0$ does not necessarily imply that $X_1$ and $X_2$ are independent.

Notice also that negative covariance results in lower value for the variance of their sum than the sum of the individual variances.

**Homework 1.44**

Prove that $Cov[X_1, X_2] = 0$ does not necessarily imply that $X_1$ and $X_2$ are independent.

**Guide**

The proof is by a counter example. Consider two random variables $X$ and $Y$ and assume that both have Bernoulli distribution with parameter $p$. Consider random variable $X_1$ defined by $X_1 = X+Y$ and another random variable $X_2$ defined by $X_2 = X-Y$. Show that $Cov[X_1, X_2] = 0$ and that $X_1$ and $X_2$ are not independent. □

Let the sum of the random variables $X_1, X_2, X_3, \ldots, X_k$ be denoted by

$$S_k = X_1 + X_2 + X_3 + \ldots + X_k.$$

Then

$$Var[S_k] = \sum_{i=1}^{k} Var[X_i] + 2\sum_{i<j} Cov[X_i, X_j] \tag{111}$$

where $\sum_{i<j} Cov[X_i, X_j]$ is a sum over all $Cov[X_i, X_j]$ such that $i$ and $j$ is a pair selected without repetitions out of $1, 2, 3, \ldots k$ so that $i < j$.

**Homework 1.45**

Prove Eq. (111).

**Guide**

First show that $S_k - E[S_k] = \sum_{i=1}^{k}(X_i - E[X_i])$ and that

$$(S_k - E[S_k])^2 = \sum_{i=1}^{k}(X_i - E[X_i])^2 + 2\sum_{i<j}(X_i - E[X_i])(X_j - E[X_j]).$$

Then take expectations of both sides of the latter. □

If we consider $k$ independent random variables denoted $X_1, X_2, X_3, \ldots, X_k$, then by substituting $Cov[X_i, X_j] = 0$ for all relevant $i$ and $j$ in (111), we obtain

$$Var[S_k] = \sum_{i=1}^{k} Var[X_i]. \tag{112}$$

**Homework 1.46**

Use Eq. (111) to explains the relationship between the variance of a Bernoulli random variable and a binomial random variable.

**Guide**

Notice that a binomial random variable with parameters $k$ and $p$ is a sum of $k$ independent Bernoulli random variables with parameter $p$. □

The covariance can take any value between $-\infty$ and $+\infty$, and in some cases, it is convenient to have a normalized dependence measure - a measure that takes values between -1 and 1. Such measure is the **correlation**. Noticing that the covariance is bounded by

$$Cov[X_1, X_2] \leq \sqrt{Var[X_1]Var[X_2]}, \tag{113}$$

the correlation of two random variables $X$ and $Y$ denoted by $Corr[X, Y]$ is defined by

$$Corr[X, Y] = \frac{Cov[X, Y]}{\sigma_X \sigma_Y}, \tag{114}$$

assuming $Var[X] \neq 0$ and $Var[Y] \neq 0$.

**Homework 1.47**

Prove that $\mid Corr[X, Y] \mid \leq 1$.

**Guide**

Let $C = Cov[X, Y]$, and show $C^2 - \sigma_X^2 \sigma_Y^2 \leq 0$, by noticing that $C^2 - \sigma_X^2 \sigma_Y^2$ is a discriminant of the quadratic $a^2\sigma_X^2 + 2aC + \sigma_Y^2$ which must be nonnegative because $E[a(X - E[X]) + (Y - E[Y])]^2$ is nonnegative. □

## 1.15 Transforms

Transforms are useful in analyses of probability models and queueing systems. We will first consider the following general definition [14] for a transform function $\Gamma$ of a random variable $X$:

$$\Gamma_X(\omega) = E[e^{\omega X}] \tag{115}$$

where $\omega$ is a complex scalar. Transforms have two important properties:

1. There is a one-to-one correspondence between transforms and probability distributions. This is why they are sometimes called *characteristics* functions. This means that for any distribution function there is a unique transform function that characterizes it and for each transform function there is a unique probability distribution it characterizes. Unfortunately it is not always easy to convert a transform to its probability distribution, and therefore we in some cases that we are able to obtain the transform but not its

probability function, we use it as means to characterize the random variable statistics instead of the probability distribution.

2. Having a transform function of a random variable we can generate its moments. This is why transforms are sometimes called *moment generating* functions. In many cases, it is easier to obtain the moments having the transform than having the actual probability distribution.

We will now show how to obtain the moments of a continuous random variable $X$ with density function $f_X(x)$ from its transform function $\Gamma_X(\omega)$. By definition,

$$\Gamma_X(\omega) = \int_{-\infty}^{\infty} e^{\omega x} f_X(x) dx. \tag{116}$$

Taking derivative with respect to $\omega$ leads to

$$\Gamma'_X(\omega) = \int_{-\infty}^{\infty} x e^{\omega x} f_X(x) dx. \tag{117}$$

Letting $\omega \to 0$, we obtain

$$\lim_{\omega \to 0} \Gamma'_X(\omega) = E[X], \tag{118}$$

and in general, taking the $n$th derivative and letting $\omega \to 0$, we obtain

$$\lim_{\omega \to 0} \Gamma^{(n)}_X(\omega) = E[X^n]. \tag{119}$$

**Homework 1.48**

Derive Eq. (119) using (116) – (118) completing all the missing steps. □

Consider for example the exponential random variable $X$ with parameter $\lambda$ having density function $f_X(x) = \lambda e^{-\lambda x}$ and derive its transform function. By definition,

$$\Gamma_X(\omega) = E[e^{\omega X}] = \lambda \int_{x=0}^{\infty} e^{\omega x} e^{-\lambda x} dx, \tag{120}$$

which gives after some derivations

$$\Gamma_X(\omega) = \frac{\lambda}{\lambda - \omega}. \tag{121}$$

**Homework 1.49**

Derive Eq. (121) from (120) □

Let $X$ and $Y$ be random variables and assume that $Y = aX + b$. The transform of $Y$ is given by

$$\Gamma_Y(\omega) = E[e^{\omega Y}] = E[e^{\omega(aX+b)}] = e^{\omega b} E[e^{\omega a X}] = e^{\omega b} \Gamma_X(\omega a). \tag{122}$$

Let random variable $Y$ be the sum of independent random variables $X_1$ and $X_2$, i.e., $Y = X_1 + X_2$. The transform of $Y$ is given by

$$\Gamma_Y(\omega) = E[e^{\omega Y}] = E[e^{\omega(X_1+X_2)}] = E[e^{\omega X_1}] E[e^{\omega X_2}] = \Gamma_{X_1}(\omega) \Gamma_{X_2}(\omega). \tag{123}$$

This result applies to a sum of $n$ independent random variables, so the transform of a sum of independent random variable equals to the product of their transform. If $Y = \sum_{i=1}^{n} X_i$ and all the $X_i$s are $n$ independent and identically distributed (IID) random variables, then

$$\Gamma_Y(\omega) = E[e^{\omega Y}] = [\Gamma_{X_1}(\omega)]^n. \tag{124}$$

Let us now consider a Gaussian random variable $X$ with parameters $m$ and $\sigma$ and density

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/2\sigma^2} \qquad -\infty < x < \infty. \tag{125}$$

Its transform is derived as follows

$$\begin{aligned}
\Gamma_X(\omega) &= E[e^{\omega X}] \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/2\sigma^2} e^{\omega x} \\
&= e^{(\sigma^2\omega^2/2)+m\omega} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/2\sigma^2} e^{\omega x} e^{-(\sigma^2\omega^2/2)-m\omega} \\
&= e^{(\sigma^2\omega^2/2)+m\omega} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m-\sigma^2\omega)^2/2\sigma^2} \\
&= e^{(\sigma^2\omega^2/2)+m\omega}.
\end{aligned}$$

Let us use the transform just derived to obtain the mean and variance of a Gaussian random variable with parameters $m$ and $\sigma$. Taking the first derivative and putting $\omega = 0$, we obtain

$$E[X] = \Gamma_X^{'}(0) = m. \tag{126}$$

Taking the second derivative and setting $\omega = 0$, we obtain

$$E[X^2] = \Gamma_X^{(2)}(0) = \sigma^2 + m^2. \tag{127}$$

Thus,

$$Var[X] = E[X^2] - (E[X])^2 = \sigma^2 + m^2 - m^2 = \sigma^2. \tag{128}$$

A Gaussian random variable with mean equal to zero and variance equal to one is called *standard Gaussian*. It is well known that if $Y$ is Gaussian with mean $m$ and standard deviation $\sigma$, then the random variable $X$ defined as

$$X = \frac{Y - m}{\sigma} \tag{129}$$

is standard Gaussian.

Substituting $\sigma = 1$ and $m = 0$ in the above transform of a Gaussian random variable, we obtain that

$$\Gamma_X(\omega) = e^{(\omega^2/2)} \tag{130}$$

is the transform of a standard Gaussian random variable.

**Homework 1.50**

Show the consistency between the results obtained for transform of a Gaussian random variable, (122), (129) and (130). □

Let $X_i$, $i = 1, 2, 3, \ldots, n$ be $n$ independent random variables and let $Y$ be a random variable that equals $X_i$ with probability $p_i$ for $i = 1, 2, 3, \ldots, N$. Therefore, by the Law of Total Probability,

$$P(Y = y) = \sum_{i=1}^{N} p_i P(X_i = y) \tag{131}$$

or for continuous densities

$$f_Y(y) = \sum_{i=1}^{n} p_i f_{X_i}(y). \tag{132}$$

Its transform is given by

$$\begin{aligned}
\Gamma_Y(\omega) &= E[e^{\omega Y}] \\
&= \int_{-\infty}^{\infty} f_Y(y) e^{\omega y} \\
&= \int_{-\infty}^{\infty} [\sum_{i=1}^{n} p_i f_{X_i}(y)] e^{\omega y} \\
&= \int_{-\infty}^{\infty} \sum_{i=1}^{n} p_i f_{X_i}(y) e^{\omega y} \\
&= \sum_{i=1}^{n} p_i \Gamma_{X_i}(\omega).
\end{aligned}$$

Notice that if the $X_i$ are exponential random variables then, by definition, $Y$ is hyper-exponential.

Particular transforms include the Z, the Laplace, and the Fourier transforms.

The **Z-transform** $\Pi_X(z)$ applies to integer valued random variable $X$ and is defined by

$$\Pi_X(z) = E[z^X].$$

This is a special case of (115) by setting $z = e^{\omega}$.

The **Laplace transform** applies to nonnegative valued random variable $X$ and is defined by

$$L_X(s) = E[e^{-sX}] \;\; \text{for} s \geq 0.$$

This is a special case of (115) by setting $\omega = -s$.

The **Fourier transform** applies to both nonnegative and negative valued random variable $X$ and is defined by

$$\Upsilon_X(s) = E[e^{i\theta X}],$$

where $i = \sqrt{-1}$ and $\theta$ is real. This is a special case of (115) by setting $\omega = i\theta$.

We will only use the Z and Laplace transforms in this book.

### 1.15.1 Z-transform

Consider a discrete and nonnegative random variable $X$, and let $p_i = P(X = i)$, $i = 0, 1, 2, \ldots$ with $\sum_{i=0}^{\infty} p_i = 1$. The Z-transform of $X$ is defined by

$$\Pi_X(z) = E[z^X] = \sum_{i=0}^{\infty} p_i z^i, \tag{133}$$

where $z$ is a real number that satisfies $0 \leq z \leq 1$. Note that in many applications the Z-transform is defined for complex $z$. However, for the purpose of this book, we will only consider real $z$ within $0 \leq z \leq 1$.

#### Homework 1.51

Prove the following properties of the Z-transform $\Pi_X(z)$:

1. $\lim_{z \to 1^-} \Pi_X(z) = 1$ ($z \to 1^-$ is defined as $z$ approaches 1 from below).
2. $p_i = \Pi_X^{(i)}(0)/i!$ where $\Pi_X^{(i)}(z)$ is the $i$th derivative of $\Pi_X(z)$.
3. $E[X] = \lim_{z \to 1^-} \Pi_X^{(1)}(z)$. □

For simplification of notation, in the following, we will use $\Pi_X^{(i)}(1) = \lim_{z \to 1^-} \Pi_X^{(i)}(z)$, but the reader must keep in mind that a straightforward substitution of $z = 1$ in $\Pi_X^{(i)}(z)$ is not always possible and the limit needs to be derived. An elegant way to show the 3rd property is to consider $\Pi_X(z) = E[z^X]$, and exchanging the operation of derivative and expectation, we obtain $\Pi_X^{(1)}(z) = E[Xz^{X-1}]$, so $\Pi_X^{(1)}(1) = E[X]$. Similarly,

$$\Pi_X^{(i)}(1) = E[X(X-1)\ldots(X-i+1)]. \tag{134}$$

#### Homework 1.52

Show that the variance $Var[X]$ is given by

$$Var[X] = \Pi_X^{(2)}(1) + \Pi_X^{(1)}(1) - (\Pi_X^{(1)}(1))^2. \quad \Box \tag{135}$$

#### Homework 1.53

Derive a formula for $E[X^i]$ using the Z-transform. □

As a Z-transform is a special case of the transform $\Gamma_Y(\omega) = E[e^{\omega Y}]$, the following results hold.

If random variables $X$ and $Y$ are related by $Y = aX + b$ for real numbers $a$ and $b$ then

$$\Pi_Y(z) = z^b \Pi_X(za). \tag{136}$$

Let random variable $Y$ be the sum of independent random variables $X_1$, $X_2$, $\ldots$, $X_n$ ($Y = \sum_{i=1}^{n} X_i$), The Z-transform of $Y$ is given by

$$\Pi_Y(z) = \Pi_{X_1}(z)\Pi_{X_2}(z)\Pi_{X_3}(z) \ \ldots \ \Pi_{X_n}(z). \tag{137}$$

If $X_1, X_2, \ldots, X_n$ are also identically distributed, then

$$\Pi_Y(z) = [\Pi_{X_1}(z)]^n. \tag{138}$$

Let us now consider several examples of Z-transforms of nonnegative discrete random variables. If $X$ is a Bernoulli random variable with parameter $p$, then its Z-transform is given by

$$\Pi_X(z) = (1-p)z^0 + pz^1 = 1 - p + pz. \tag{139}$$

Its mean is $E[X] = \Pi_X^{(1)}(1) = p$ and by (135) its variance is $p(1-p)$.

If $X$ is a Geometric random variable with parameter $p$, then its Z-transform is given by

$$\Pi_X(z) = p\sum_{i=1}^{\infty}(1-p)^{i-1}z^i = \frac{pz}{1-(1-p)z}. \tag{140}$$

Its mean is $E[X] = \Pi_X^{(1)}(1) = 1/p$ and by (135) its variance is $(1-p)/p^2$.

If $X$ is a Binomial random variable with parameter $p$, then we can obtain its Z-transform either by definition or by realizing that a Binomial random variable is a sum of $n$ IID Bernoulli random variables. Therefore its Z-transform is given by

$$\Pi_X(z) = (1-p+pz)^n = [1+(z-1)p]^n. \tag{141}$$

**Homework 1.54**

Verify that the latter is consistent with the Z-transform obtained using $\Pi_X(z) = \sum_{i=0}^{\infty} p_i z^i$.
□

The mean of $X$ is $E[X] = \Pi_X^{(1)}(1) = np$ and by (135) its variance is $np(1-p)$.

If $X$ is a Poisson random variable with parameter $\lambda$, then its Z-transform is given by

$$\Pi_X(z) = \sum_{i=0}^{\infty} p_i z^i = e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i z^i}{i!} = e^{(z-1)\lambda}. \tag{142}$$

Its mean is $E[X] = \Pi_X^{(1)}(1) = \lambda$ and by (135) its variance is also equal to $\lambda$.

We can now see the relationship between the Binomial and the Poisson random variables. If we consider the Z-transform of the Binomial random variable $\Pi_X(z) = (1-p+pz)^n$, and set $\lambda = np$ as a constant so that $\Pi_X(z) = (1+(z-1)\lambda/n)^n$ and let $n \to \infty$, we obtain

$$\lim_{n\to\infty}(1-p+pz)^n = \lim_{n\to\infty}[1+(z-1)\lambda/n]^n = e^{(z-1)\lambda} \tag{143}$$

which is exactly the Z-transform of the Poisson random variable. This proves the convergence of the binomial to the Poisson random variable if we keep $np$ constant and let $n$ go to infinity.

### 1.15.2 Laplace Transform

The Laplace transform of a non-negative random variable $X$ with density $f_X(x)$ is defined as

$$\mathcal{L}_X(s) = E[e^{-sX}] = \int_0^{\infty} e^{-sx} f_X(x)dx. \tag{144}$$

As it is related to the transform $\Gamma_X(\omega) = E[e^{\omega X}]$ by setting $\omega = -s$, similar derivations to those made for $\Gamma_X(\omega)$ above give the following.

If $X_1$, $X_2$, $\ldots$, $X_n$ are $n$ independent random variables then

$$\mathcal{L}_{X_1+X_2+\ \ldots\ +X_n}(s) = \mathcal{L}_{X_1}(s)\mathcal{L}_{X_2}(s)\ \ldots\ \mathcal{L}_{X_n}(s). \tag{145}$$

Let $X$ and $Y$ be random variables and $Y = aX + b$. The Laplace transform of $Y$ is given by

$$\mathcal{L}_Y(s) = e^{-sb}\mathcal{L}_X(sa). \tag{146}$$

The $n$th moment of random variable $X$ is given by

$$E[X^n] = (-1)^n\mathcal{L}_X^{(n)}(0) \tag{147}$$

where $\mathcal{L}_Y^{(n)}(0)$ is the $n$th derivative of $\mathcal{L}_X(s)$ at $s = 0$ (or at the limit $s \to 0$). Therefore,

$$Var[X] = E[X^2] - (E[X])^2 = (-1)^2\mathcal{L}_X^{(2)}(0) - ((-1)\mathcal{L}_X^{(1)}(0))^2 = \mathcal{L}_X^{(2)}(0) - (\mathcal{L}_X^{(1)}(0))^2. \tag{148}$$

Let $X$ be an exponential random variable with parameter $\lambda$. Its Laplace transform is given by

$$\mathcal{L}_X(s) = \frac{\lambda}{\lambda + s}. \tag{149}$$

**Homework 1.55**

Derive (145)–(149) using the derivations made for $\Gamma_X(\omega)$ as a guide. □

Now consider $N$ to be a nonnegative discrete (integer) random variable of a probability distribution that has the Z-transform $\Pi_N(z)$, and let $Y = X_1+X_2+\ \ldots\ +X_N$, where $X_1,\ X_2,\ \ldots,\ X_N$ are nonnegative IID random variables with a common distribution that has the Laplace transform $\mathcal{L}_X(s)$ (i.e., they are exponentially distributed). Let us derive the Laplace transform of $Y$. Conditioning and unconditioning on $N$, we obtain

$$\mathcal{L}_Y(s) = E[e^{-sY}] = E_N[E[e^{-s(X_1+X_2+\ \ldots\ +X_N|N)}]]. \tag{150}$$

Therefore, by independence of the $X_i$,

$$\mathcal{L}_Y(s) = E_N[E[e^{-sX_1} + E[e^{-sX_2} +\ \ldots\ + E[e^{-sX_N}]] = E_N[(\mathcal{L}_X(s))^N]. \tag{151}$$

Therefore

$$\mathcal{L}_Y(s) = \Pi_N[(\mathcal{L}_X(s))]. \tag{152}$$

An interesting example of (152) is the case where the $X_i$ are IID exponentially distributed each with parameter $\lambda$, and $N$ is geometrically distributed with parameter $p$. In this case, we already know that since $X$ is an exponential random variable, we have $\mathcal{L}_X(s) = \lambda/(\lambda + s)$, so

$$\mathcal{L}_Y(s) = \Pi_N\left(\frac{\lambda}{\lambda + s}\right). \tag{153}$$

We also know that $N$ is geometrically distributed, so $\Pi_N(z) = pz/[1-(1-p)z]$. Therefore, from (153), we obtain,

$$\mathcal{L}_Y(s) = \frac{\frac{p\lambda}{\lambda+s}}{1-\frac{(1-p)\lambda}{\lambda+s}} \tag{154}$$

and after some algebra we obtain

$$\mathcal{L}_Y(s) = \frac{p\lambda}{s+p\lambda}. \tag{155}$$

This result is interesting. We have shown that $Y$ is exponentially distributed with parameter $p\lambda$.

Note that if $N$ is a geometric random variable of failures then this result will not apply because $Y$ will have an atom at the zero and therefore it cannot be exponentially distributed.

### Homework 1.56

Let $X_1$ and $X_2$ be exponential random variables with parameters $\lambda_1$ and $\lambda_2$ respectively. Consider a random variable $Y$ defined by the following algorithm.

1. Initialization: $Y = 0$.
2. Conduct an experiment to obtain the values of $X_1$ and $X_2$. If $X_1 < X_2$ then $Y = Y + X_1$ and Stop. Else, $Y = Y + X_2$ and repeat 2.

Show that $Y$ is exponentially distributed with parameter $\lambda_1$.

### Hint

Notice that $Y$ is a geometric sum of exponential random variables. □

### Homework 1.57

Derive the density, the Laplace transform, the mean and the variance of $Y$ in the following three cases.

1. Let $X_1$ and $X_2$ be exponential random variables with parameters $\mu_1$ and $\mu_2$, respectively. In this case, $Y$ is a hyperexponential random variable with density $f_Y(y) = pf_{X_1}(y) + (1-p)f_{X_2}(y)$.
2. Let $X_1$ and $X_2$ be exponential random variables with parameters $\mu_1$ and $\mu_2$, respectively. The hypoexponential random variable $Y$ is defined by $Y = X_1 + X_2$.
3. Let $Y$ be an Erlang random variable, namely, $Y = \sum_{i=1}^{k} X_i$ where the $X_i$s are IID exponentially distributed random variables with parameter $\mu$.

Now plot the standard deviation to mean ratio for the cases of hyperexponential and Erlang random variables over a wide range of parameter values and discuss implications. For example, show that for Erlang(k) the standard deviation to mean ratio approaches zero as k approaches infinity. □

## 1.16 Multivariate Random Variables and Transform

A *multivariate random variable* is a vector $X = (X_1, X_2, \ \ldots, \ X_k)$ where each of the $k$ components is a random variable. A multivariate random variable is also known as *random vector*. These $k$ components of a random vector are related to events (outcomes of experiments) on the same sample space and they can be continuous or discrete. They also have a legitimate well defined joint distribution (or density) function. The distribution of each individual component $X_i$ of the random vector is its marginal distribution. A transform of a random vector $X = (X_1, X_2, \ \ldots, \ X_k)$ is called *multivariate transform* and is defined by

$$\Gamma_X(\omega_1, \omega_2, \ \ldots, \ \omega_k) = E[s^{\omega_1 X_1, \omega_2 X_2, \ \ldots, \ \omega_k X_k}]. \tag{156}$$

## 1.17 Probability Inequalities and Their Dimensioning Applications

In the course of design of telecommunications networks, a fundamental important problem is how much capacity a link should have. If we consider the demand as a non-negative random variable $X$ and the link capacity as a fixed scalar $C > 0$, we will be interested in the probability that the demand exceeds the capacity $P(X > C)$. The more we know about the distribution the more accurate out estimation of $P(X > C)$.

If we know only the mean, we use the so-called **Markov inequality**:

$$P(X > C) \leq \frac{E[X]}{C}. \tag{157}$$

**Homework 1.58**

Prove Eq. (157).

**Guide**

Define a new random variable $U(C)$ a function of $X$ and $C$ defined by: $U(C) = 0$ if $X < C$, and $U(C) = C$ if $X \geq C$. Notice $U(C) \leq X$, so $E[U(C)] \leq E[X]$. Also, $E[U(C)] = CP(U(C) = C) = CP(X \geq C)$, and Eq. (157) follows. □

If we know the mean and the variance of $X$, then we can use the so-called **Chebyshev inequality**:

$$P(\mid X - E[X] \mid > C) \leq \frac{Var[X]}{C^2}. \tag{158}$$

**Homework 1.59**

Prove Eq. (158).

**Guide**

Define a new random variable $(X - E[X])^2$ and apply the Markov inequality putting $C^2$ instead of C obtaining:

$$P((X - E[X])^2 \geq C^2) \leq \frac{E[(X - E[X])^2]}{C^2} = \frac{Var[X]}{C^2}.$$

Notice that the two events $(X - E[X])^2 \geq C^2$ and $\mid X - E[X] \mid \geq C$ are identical. □

Another version of Chebyshev inequality is

$$P(\mid X - E[X] \mid > C^*\sigma) \leq \frac{1}{(C^*)^2} \tag{159}$$

for $C^* > 0$.

**Homework 1.60**

Prove and provide interpretation to Eq. (159).

**Guide**

Observe that the right-hand side of (159) is equal to $\frac{Var[X]}{Var[X](C^*)^2}$. □

**Homework 1.61**

For a wide range of parameter values, study numerically how tight the bounds provided by Markov versus Chebyshev inequalities are. Discuss the differences and provide interpretations. □

A further refinement of the Chebyshev inequality is the following **Kolmogorov inequality**. Let $X_1, X_2, X_3, \ \ldots, \ X_k$ be a sequence of mutually independent random variables (not necessarily identically distributed) and let $S_k = X_1 + X_2 + X_3 + \ \ldots \ + X_k$ and $\sigma(S_k)$ be the standard deviation of $S_k$. Then for every $\epsilon > 0$,

$$P(\mid S_k - E[S_k] \mid < \theta\sigma(S_k) \ \text{ for all } \ k = 1, 2, \ldots, n) \geq 1 - \frac{1}{\theta^2}. \tag{160}$$

The interested reader may consult Feller [25] for the proof of the Kolmogorov inequality. We are however more interested in its teletraffic implication. If we let time be divided into consecutive intervals and we assume that $X_i$ is the number of packets arrive during the $i$th interval, and if the number of packets arrive during the different intervals are mutually independent, then it is rare that we will have within a period of $n$ consecutive intervals any period of $k$ consecutive intervals ($k \leq n$) during which the number of packets arriving is significantly more than the average.

## 1.18 Limit Theorems

Let $X_1, X_2, X_3, \ \ldots, \ X_k$ be a sequence of IID random variables with mean $\lambda$ and variance $\sigma^2$. Let $\bar{S}_k$ be the *sample mean* of these $k$ random variables defined by

$$\bar{S}_k = \frac{X_1 + X_2 + X_3 + \ \ldots \ + X_k}{k}.$$

This gives

$$E[\bar{S}_k] = \frac{E[X_1] + E[X_2] + E[X_3] + \ \ldots \ + E[X_k]}{k} = \frac{k\lambda}{k} = \lambda.$$

Recalling that the $X_i$s are independent, we obtain

$$Var[\bar{S}_k] = \frac{\sigma^2}{k}. \tag{161}$$

**Homework 1.62**

Prove Eq. (161). □

Applying Chebyshev's inequality, we obtain

$$P(| \ \bar{S}_k - \lambda \ | \geq \varepsilon) \leq \frac{\sigma^2}{k\varepsilon^2} \ \text{ for all } \ \varepsilon > 0. \tag{162}$$

Noticing that as $k$ approaches infinity, the right-hand side of (162) approaches zero which implies that the left-hand side approaches zero as well. This leads to the so-called **the weak law of large numbers** that states the following. Let $X_1, X_2, X_3, \ \ldots, \ X_k$ be $k$ IID random variables with common mean $\lambda$. Then

$$P\left(\left|\frac{X_1 + X_2 + X_3 + \ \ldots \ + X_k}{k} - \lambda\right| \geq \varepsilon)\right) \to 0 \ \text{ as } \ k \to \infty \ \text{ for all } \ \varepsilon > 0. \tag{163}$$

What the weak law or large number essentially says is that the sample mean approaches the mean as the sample size increases.

Next we state the central limit theorem that we have mentioned in Section 1.10.7. Let $X_1, X_2, X_3, \ \ldots, \ X_k$ be $k$ IID random variables with common mean $\lambda$ and variance $\sigma^2$. Let random variable $Y_k$ be defined as

$$Y_k = \frac{X_1 + X_2 + X_3 + \ \ldots \ + X_k - k\lambda}{\sigma\sqrt{k}}. \tag{164}$$

Then,

$$\lim_{k\to\infty} P(Y_k \leq y) = \Phi(y) \tag{165}$$

where $\Phi(\cdot)$ is the distribution function of a standard Gaussian random variable given by

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2} dt.$$

**Homework 1.63**

Prove that $E[Y_k] = 0$ and that $Var[Y_k] = 1$ from first principles without using the central limit theorem. □

As we mentioned in Section 1.10.7, the central limit theorem is considered the most important result in probability. Notice that it implies that the sum of $k$ IID random variable with common mean $\lambda$ and variance $\sigma^2$ is approximately Guassian with mean $k\lambda$ and variance $k\sigma^2$ *regardless* of the distribution of these variables.

Moreover, under certain conditions, the central limit theorem also applies in the case of sequences that are not identically distributed, provided one of a number of conditions apply. One of the cases where the central limit theorem also applies in the case of non-IID random variables is due to Lyapunov described as follows. Consider $X_1, X_2, X_3, \ldots, X_k$ to be a sequence of independent random variables. Let $\lambda_n = E[X_n],\ n = 1, 2, \ldots, k$ and $\sigma_n^2 = Var[X_n],\ n = 1, 2, \ldots, k$, and assume that all $\lambda_n$ and $\sigma_n^2$ are finite. Let

$$\hat{S}_n^2 = \sum_{i=1}^{n} \sigma_i^2,$$

$$\hat{R}_n^3 = \sum_{i=1}^{n} E[|X_i - \lambda_i|^3],$$

and assume that $\hat{S}_n^2$ and $\hat{R}_n^3$ are finite for all $n = 1, 2, \ldots, k$. Further assume that

$$\lim_{k\to\infty} \frac{\hat{R}}{\hat{S}} = 0.$$

The latter is called "Lyapunov condition".

If these conditions hold then the random variable $\sum_{i=1}^{k} X_i$ has Gaussian distribution with mean $\sum_{i=1}^{k} \lambda_i$ and variance $\sum_{i=1}^{k} \sigma_i^2$. This generalization of the central limit theorem to non IID random variables, based on Lyapunov condition, is called "Lyapunov's central limit theorem".

## 1.19 Link Dimensioning

Before we end this chapter on probability, let us demonstrate how the probability concepts discussed so far can be used to provide simple means for link dimensioning. We will consider several scenarios of sources (individuals or families) sharing a communication link. Each of the sources has certain requirements for capacity and the common link must be dimensioned in such a way that minimizes the cost for the telecommunications provider, but still meets the individual QoS requirements. The link dimensioning procedures that we described below apply to user requirements for capacity. These requirements apply to transmissions from the sources to the network as well as to downloads from the networks to the user or to combination of downloads and transmissions. We are not concerned with specific directions of transmission. We assume that the capacity of the common link can be used in either direction. When we say a source "transmits" it should always be read as "transmits and/or downloads".

### 1.19.1 Case 1: Homogeneous Individual Sources

Consider $N$ independent sources (end-terminals), sharing a transmission link of capacity $C$ [Mb/s]. Any of the sources transmits data in accordance with an on-off process. That is, a source alternates between two states: 1) the on state during which the source transmits at a rate $R$ [Mb/s], and 2) the off state during which the source is idle. Assume that the proportion of time the source is in the on-state is $p$, so it is in the off-state $1-p$ of the time. The question is how much capacity should the link have so it can serve all $N$ sources such that the probability that the demand exceeds the total link capacity is not higher than $\alpha$.

We first derive the distribution of the total traffic demanded by the $N$ sources. Without loss of generality, let us normalize the traffic generated by a source during an on-period by setting $R = 1$. Realizing that the demand generated by a single source is Bernoulli distributed with parameter $p$, we obtain that the demand generated by all $N$ sources has Binomial distribution with parameters $p$ and $N$. Accordingly, finding the desired capacity is reduced to finding the smallest $C$ such that

$$\sum_{i=C+1}^{N} \binom{N}{i} p^i (1-p)^{N-i} \leq \alpha. \tag{166}$$

Since the left-hand side of (166) increases as $C$ decreases, and since its value is zero if $C = N$, all we need to do to find the optimal $C$ is to compute the value of the left-hand side of (166) for $C$ values of $N-1$, $N-2$, ... until we find the first $C$ value for which the inequality (166) is violated. Increasing that $C$ value by one will give us the desired optimal $C$ value.

If $N$ is large we can use the central limit theorem and approximate the Binomial distribution by a Gaussian distribution. Accordingly, the demand can be approximated by a Gaussian random variable with mean $Np$ and variance $Np(1-p)$ and simply find $C_G$ such that the probability of our Gaussian random variable to exceed $C_G$ is $\alpha$.

It is well known that Gaussian random variables obey the so-called 68-95-99.7% Rule which means that the following apply to a random variable $X$ with mean $m$ and standard deviation $\sigma$.

$$\begin{aligned} P(m-\sigma \leq X \leq m+\sigma) &= 0.68 \\ P(m-2\sigma \leq X \leq m+2\sigma) &= 0.95 \\ P(m-3\sigma \leq X \leq m+3\sigma) &= 0.997. \end{aligned}$$

Therefore, if $\alpha = 0.0015$ then $C_G$ should be three standard deviations above the mean, namely,

$$C_G = Np + 3\sqrt{Np(1-p)}. \tag{167}$$

Note that $\alpha$ is a preassign QoS measure representing the proportion of time that the demand exceeds the supply and under the zero buffer approximation during that period some traffic is lost. If it is required that $\alpha$ is lower than 0.0015, then more than three standard deviations above the mean are required. Recall that for our original problem, before we introduced the Gaussian approximation, $C = N$ guarantees that there is sufficient capacity to serve all arriving traffic without losses. Therefore, we set our dimensioning rule for the optimal $C$ value as follows:

$$C_{opt} = \min\left[N, Np + 3\sqrt{Np(1-p)}\right]. \tag{168}$$

### 1.19.2 Case 2: Non-homogeneous Individual Sources

Here we generalize the above scenario to the case where the traffic and the peak rates of different sources can be different. Consider $N$ sources where the $i$th source transmits at rate $R_i$ with probability $p_i$, and at rate 0 with probability $1 - p_i$. In this case where the sources are non-homogeneous, we must invoke a generalization of the central limit theorem that allows for non IID random variables (i.e., the "Lyapunov's central limit theorem"). Let $R_X(i)$ be a random variable representing the rate transmitted by source $i$. We obtain:

$$E[R_X(i)] = p_i R_i.$$

and

$$Var[R_X(i)] = R_i^2 p_i - (R_i p_i)^2 = R_i^2 p_i (1 - p_i).$$

The latter is consistent with the fact that $R_X(i)$ is equal to $R_i$ times a Bernoulli random variable. We now assume that the random variable

$$\Sigma_R = \sum_{i=1}^{N} R_X(i)$$

has a Gaussian distribution with mean

$$E[\Sigma_R] = \sum_{i=1}^{N} E[R_X(i)] = \sum_{i=1}^{N} p_i R_i$$

and variance

$$Var[\Sigma_R] = \sum_{i=1}^{N} Var[R_X(i)] = \sum_{i=1}^{N} R_i^2 p_i (1 - p_i).$$

Notice that the allocated capacity should not be more than the total sum of the peak rates of the individual sources. Therefore, in this more general case, for the QoS requirement $\alpha = 0.0015$, our optimal $C$ value is set to:

$$C_{opt} = \min \left[ \sum_{i=1}^{N} R_i, E[\Sigma_R] + 3\sqrt{Var[\Sigma_R]} \right]. \tag{169}$$

**Homework 1.64**

There are 20 sources each transmits at a peak-rate of 10 Mb/s with probability 0.1 and is idle with probability 0.9, and there are other 80 sources each transmits at a peak-rate of 1 Mb/s with probability 0.05 and is idle with probability 0.95. A service provider aims to allocate the minimal capacity $C_{opt}$ such that no more than 0.0015 of the time, the demand of all these 100 sources exceeds the available capacity. Set $C_{opt}$ using above describe approach.

**Answer:** $C_{opt} = 64.67186$ Mb/s.

Notice the difference in contributions to the total variance of sources from the first group versus such contributions of sources from the second group.

Consider a range of examples where the variance is the dominant part of $C_{opt}$ versus examples where the variance is not the dominant part of $C_{opt}$. □

### 1.19.3 Case 3: Capacity Dimensioning for a Community

In many cases, the sources are actually a collection of sub-sources. A source could be a family of several members and at any given point in time, one or more of the family members are accessing the link. In such a case, we assume that source $i$, $i = 1, 2, 3, \ldots, N$, transmits at rate $R_j(i)$ with probability $p_{ij}$ for $j = 0, 1, 2, 3, \ldots, J(i)$. For all $i$, $R_0(i) \equiv 0$ and $R_{J(i)}(i)$ is defined to be the peak rate of source $i$. For each source (family) $i$, $R_j(i)$ and $p_{ij}$ for $j = 1, 2, 3, \ldots, J(i)-1$, are set based on measurements for the various rates reflecting the total rates transmitted by active family members and their respective proportion of time used. For example, for a certain family $i$, $R_1(i)$ could be the rate associated with one individual family member browsing the web, $R_2(i)$ the rate associated with one individual family member using Voice over IP, $R_3(i)$ the rate associated with one individual family member watching video, $R_4(i)$ the rate associated with one individual family member watching video and another browsing the web, etc. The $p_{ij}$ is the proportion of time during the busiest period of the day that $R_i(j)$ is used.

Again, defining $R_X(i)$ as a random variable representing the rate transmitted by source $i$, we have

$$E[R_X(i)] = \sum_{j=0}^{J(i)} p_{ij} R_j(i) \quad \text{for} \quad i = 1, 2, 3, \ldots, N.$$

and

$$Var[R_X(i)] = \sum_{j=0}^{J(i)} \{R_j(i)\}^2 p_{ij} - \{E[R_X(i)]\}^2 \quad \text{for} \quad i = 1, 2, 3, \ldots, N.$$

Again, assume that the random variable

$$\Sigma_R = \sum_{i=1}^{N} R_X(i)$$

has a Gaussian distribution with mean

$$E[\Sigma_R] = \sum_{i=1}^{N} E[R_X(i)]$$

and variance

$$Var[\Sigma_R] = \sum_{i=1}^{N} Var[R_X(i)].$$

Therefore, in this general case, for the QoS requirement $\alpha = 0.0015$, our optimal $C$ value is again set by

$$C_{opt} = \min\left[\sum_{i=1}^{N} R_{J(i)}(i), E[\Sigma_R] + 3\sqrt{Var[\Sigma_R]}\right]. \tag{170}$$

# 2 Relevant Background in Stochastic Processes

Aiming to understand behaviors of various natural and artificial processes, researchers often model them as collections of random variables where the mathematically defined statistical characteristics and dependencies of such random variables are fitted to those of the real processes. The research in the field of stochastic processes has therefore three facets:

**Theory:** mathematical explorations of stochastic processes models that aim to better understand their properties.

**Measurements:** taken on the real process in order to identify its statistical characteristics.

**Modelling:** fitting the measured statistical characteristics of the real process with those of a model and development of new models of stochastic processes that well match the real process.

This chapter provides background on basic theoretical aspects of stochastic processes which form a basis for queueing theory and teletraffic models discussed in the later chapters.

## 2.1 General Concepts

For a given *index set* $T$, a *stochastic process* $\{X_t,\ t \in T\}$ is an indexed collection of random variables. They may or may not be identically distributed. In many applications the index $t$ is used to model time. Accordingly, the random variable $X_t$ for a given $t$ can represent, for example, the number of telephone calls that have arrived at an exchange by time $t$.

If the index set $T$ is countable, the stochastic process is called a *discrete-time* process, or a *time series* [7, 16, 63]. Otherwise, the stochastic process is called a *continuous-time* process. Considering our previous example, where the number of phone calls arriving at an exchange by time $t$ is modelled as a continuous-time process $\{X_t,\ t \in T\}$, we can alternatively, use a discrete-time process to model, essentially, the same thing. This can be done by defining the discrete-time process $\{X_n,\ n = 1,\ 2,\ 3,\ \dots\}$, where $X_n$ is a random variable representing, for example, the number of calls arriving within the $n$th minute.

A stochastic process $\{X_t,\ t \in T\}$ is called *discrete space* stochastic process if the random variables $X_t$ are discrete, and it is called *continuous space* stochastic process if it is continuous. We therefore have four types of stochastic processes:

1. Discrete Time Discrete Space
2. Discrete Time Continuous Space
3. Continuous Time Discrete Space
4. Continuous Time Continuous Space.

A discrete-time stochastic process $\{X_n,\ n = 1,\ 2,\ 3,\ \dots\}$ is *strictly stationary* if for any subset of $\{X_n\}$, say, $\{X_{n(1)},\ X_{n(2)},\ X_{n(3)},\ \dots,\ X_{n(k)}\}$, for any integer $m$ the joint probability function $P(X_{n(1)},\ X_{n(2)},\ X_{n(3)},\ \dots,\ X_{n(k)})$, is equal to the joint probability function $P(X_{n(1)+m},\ X_{n(2)+m},\ X_{n(3)+m},\ \dots,\ X_{n(k)+m})$. In other words, $P(X_{n(1)+m},\ X_{n(2)+m},\ X_{n(3)+m},\ \dots,\ X_{n(k)+m})$ is independent of $m$. In this case, the probability structure of the process does not change with time. An equivalent definition for strict stationarity is applied also for a continuous-time process accept that in that case $m$ is non-integer.

Notice that for the process to be strictly stationary, the value of $k$ is unlimited as the joint probability should be independent of $m$ for any subset of $\{X_n,\ n = 1,\ 2,\ 3,\ \ldots\}$. If $k$ is limited to some value $k^*$, we say that the process is *stationary of order* $k^*$.

A equivalent definition applies to a continuous-time stochastic process. A continuous-time stochastic process $X_t$ is said to be strictly stationary if its statistical properties do not change with a shift of the origin. In other words the process $X_t$ statistically the same as the process $X_{t-d}$ for any value of $d$.

An important stochastic process is the *Gaussian Process* defined as a process that has the property that the joint probability function (density) associated with any set of times is multivariate Gaussian. The importance of the Gaussian process lies in its property to be an accurate model for superposition of many independent processes. This makes the Gaussian process a useful model for heavily multiplexed traffic which arrive at switches or routers deep in a major telecommunications network. Fortunately, the Gaussian process is not only useful, but it is also simple and amenable to analysis. Notice that for a multivariate Gaussian distribution, all the joint moments of the Gaussian random variables are fully determined by the joint first and second order moments of the variables. Therefore, if the first and second order moments do not change with time, the Gaussian random variables themselves are stationary. This implies that for a Gaussian process, stationarity of order two (also called *weak stationarity*) implies strict stationarity.

For a time series $\{X_n,\ n = 1,\ 2,\ 3,\ \ldots\ \}$, weak stationarity implies that, for all $n$, $E[X_n]$ is constant, denoted $E[X]$, independent of $n$. Namely, for all $n$,

$$E[X] = E[X_n]. \tag{171}$$

Weak stationarity (because it is stationarity of order two) also implies that the covariance between $X_n$ and $X_{n+k}$, for any $k$, is independent of $n$, and is only a function of $k$, denoted $U(k)$. Namely, for all $n$,

$$U(k) = Cov[X_n, X_{n+k}]. \tag{172}$$

Notice that, the case of $k = 0$ in Eq. (172), namely,

$$U(0) = Cov[X_n, X_n] = Var[X_n] \tag{173}$$

implies that the variance of $X_n$ is also independent of $n$. Also for all integer $k$,

$$U(-k) = U(k) \tag{174}$$

because $Cov[X_n, X_{n+k}] = Cov[X_{n+k}, X_n] = Cov[X_n, X_{n-k}]$. The function $U(k)$, $k = 0, 1, 2,\ \ldots$, is called the *autocovariance function*. The value of the autocovariance function at $k$, $U(k)$, is also called the autocovariance of lag $k$.

Important parameters are the so-called Autocovariance Sum, denoted $S$, and Asymptotic Variance Rate (AVR) denoted $v$ [4, 5]. They are defined by:

$$S = \sum_{i=1}^{\infty} U(i) \tag{175}$$

and

$$v = \sum_{i=-\infty}^{\infty} U(i). \tag{176}$$

Notice that

$$v = 2S + Var[X_n]. \tag{177}$$

Another important definition of the AVR which justifies its name is

$$v = \lim_{n\to\infty} \frac{Var[S_n]}{n}. \tag{178}$$

We will further discuss these concepts in Section 20.1.

**Homework 2.1**

Prove that the above two definitions are equivalent; namely, prove that

$$\lim_{n\to\infty} \frac{Var[S_n]}{n} = 2S + Var[X_n] \tag{179}$$

where

$$S_n = \sum_{i=1}^{n} X_i.$$

**Guide**

Define

$$S(k^*) = \sum_{i=1}^{k^*} U(i)$$

and notice that

$$\lim_{j\to\infty} S(j) = S.$$

Let

$$S_{k^*} = \sum_{i=1}^{k^*} X_i$$

and notice that

$$\sum_{i<j} Cov[X_i, X_j] = \sum_{n=1}^{k^*-1} \sum_{k=1}^{k^*-n} Cov[X_n, X_{n+k}] = \sum_{n=1}^{k^*-1} \sum_{k=1}^{k^*-n} U(k).$$

Noticing that by the weak stationarity property, we have that $Var[X_i] = Var[X_j]$ and $Cov[X_i, X_{i+k}] = Cov[X_j, X_{j+k}]$ for all pairs $i, j$ , and letting $k^* \to \infty$, by (111), we obtain

$$Var[S_k^*] = k^* Var[X_n] + 2k^* S$$

which leads to (179). □

The *autocorrelation function* at lag $k$, denoted $C(k)$, is the normalized version of the autocovariance function, and since by weak stationarity, for all $i$ and $j$, $Var[X_j] = Var[X_i]$, it is given by:

$$C(k) = \frac{U(k)}{Var[X_n]}. \tag{180}$$

A stochastic process is called *ergodic* if every realization contains sufficient information on the probabilistic structure of the process. For example, let us consider a process which can be in either one of two realization: either $X_n = 1$ for all $n$, or $X_n = 0$ for all $n$. Assume that each one of these two realizations occur with probability 0.5. If we observe any one of these realizations, regardless of the duration of the observations, we shall never conclude that $E[A] = 0.5$. We shall only have the estimations of either $E[A] = 0$ or $E[A] = 1$, depends on which realization we happen to observe. Such a process is not ergodic.

Assuming $\{X_n,\ n = 1,\ 2,\ 3,\ \ldots\ \}$ is ergodic and stationary, and we observe $m$ observations of this $\{X_n\}$ process, denoted by $\{\hat{A}_n,\ n = 1,\ 2,\ 3,\ \ldots,\ m\}$, then the mean of the process $E[A]$ can be estimated by

$$\hat{E}[A] = \frac{1}{m}\sum_{n=1}^{m}\hat{A}_n, \tag{181}$$

and the autocovariance function $U(k)$ of the process can be estimated by

$$\hat{U}(k) = \frac{1}{m-k}\sum_{n=k+1}^{m}(\hat{A}_{n-k} - E[A])(\hat{A}_n - E[A]). \tag{182}$$

## 2.2 Two Orderly and Memoryless Point Processes

In this section we consider a very special class of stochastic processes called *point* processes that also possess two properties: *orderliness* and *memorylessness*. After providing, somewhat intuitive, definitions of these concepts, we will discuss two processes that belong to this special class: one is discrete-time - called the *Bernoulli process* and the other is continuous-time - called the *Poisson process*.

We consider here a physical interpretation, where a *point process* is a sequence of events which we call *arrivals* occurring at random in points of time $t_i$, $i = 1, 2,\ \ldots\ , t_{i+1} > t_i$, or $i = \ldots,\ -2, -1, 0, 1, 2,\ \ldots, t_{i+1} > t_i$. The index set, namely, the time, or the set where the $t_i$ get their values from, can be continuous or discrete, although in most books the index set is considered to be the real line, or its non-negative part. We call our events arrivals to relate is to the context of queueing theory, where a point process typically corresponds to points of arrivals, i.e., $t_i$ is the time of the $i$th arrival that joints a queue. A point process can be defined by its *counting process* $\{N(t),\ t \geq 0\}$, where $N(t)$ is the number of arrivals occurred within $[0, t)$. A counting process $\{N(t)\}$ has the following properties:

1. $N(t) \geq 0$,
2. $N(t)$ is integer,
3. if $s > t$, then $N(s) \geq N(t)$ and $N(s) - N(t)$ is the number of occurrences within $(t, s]$.

Note that $N(t)$ is not an independent process because for example, if $t_2 > t_1$ then $N(t_2)$ is dependent on the number of arrivals in $[0, t_1)$, namely, $N(t_1)$.

Another way to define a point process is by the stochastic process of the inter-arrival times $\Delta_i$ where $\Delta_i = t_{i+1} - t_i$.

One important property of a counting process is the so-called *Orderliness* which means that the probability that two or more arrivals happen at once is negligible. Mathematically, for a continuous-time counting process to be *orderly*, it should satisfy:

$$\lim_{\Delta t \to 0} P(N(t+\Delta t) - N(t) > 1 \mid N(t+\Delta t) - N(t) \geq 1) = 0. \tag{183}$$

Another very important property is the *memorylessness*. A stochastic process is *memoryless* if at any point in time, the future evolution of the process is statistically independent of its past.

### 2.2.1 Bernoulli Process

The Bernoulli process is a discrete-time stochastic process made up of a sequence of IID Bernoulli distributed random variables $\{X_i,\ i = 0, 1, 2, 3,\ \ldots\}$ where for all $i$, $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. In other words, we divide time into consecutive equal time slots. Then, for each time-slot $i$, we conduct an independent Bernoulli experiment. If $X_i = 1$, we say that there was an *arrival* at time-slot $i$. Otherwise, if $X_i = 0$, we say that there was no arrival at time-slot $i$.

The Bernoulli process is both orderly and memoryless. It is orderly because, by definition, no more than one arrival can occur at any time-slot as the Bernoulli random variable takes values of more than one with probability zero. It is also memoryless because the Bernoulli trials are independent, so at any discrete point in time $n$, the future evolution of the process is independent of its past.

The counting process for the Bernoulli process is another discrete-time stochastic process $\{N(n), n \geq 0\}$ which is a sequence of Binomial random variables $N(n)$ representing the total number of arrivals occurring within the first $n$ time-slots. Notice that since we start from slot 0, $N(n)$ does not include slot $n$ in the counting. That is, we have

$$P[N(n) = i] = \binom{n}{i} p^i (1-p)^{n-i} \qquad i = 0,\ 1,\ 2,\ \ldots,\ n. \tag{184}$$

The concept of an inter-arrival time for the Bernoulli process can be explained as follows. Let us assume without loss of generality that that there was an arrival at time-slot $k$, the inter-arrival time will be the number of slots between $k$ and the first time-slot to have an arrival following $k$. We do not count time-slot $k$ but we do count the time-slot of the next arrival. Because the Bernoulli process is memoryless, the inter-arrival times are IID, so we can drop the index $i$ of $\Delta_i$, designating the $i$ inter-arrival time, and consider the probability function of the random variable $\Delta$ representing any inter-arrival time. Because $\Delta$ represents a number of independent Bernoulli trials until a success, it is geometrically distributed, and its probability function is given by

$$P(\Delta = i) = p(1-p)^{i-1} \qquad i = 1,\ 2,\ \ldots\ . \tag{185}$$

Another important statistical measure is the time it takes $n$ until the $i$th arrival. This time is a sum of $i$ inter-arrival times which is a sum of $i$ geometric random variables which we already know has a Pascal distribution with parameters $p$ and $i$, so we have

$$P[\text{the } i\text{th arrival occurs in time slot } n] = \binom{n-1}{i-1} p^i (1-p)^{n-i} \qquad i = i,\ i+1,\ i+2,\ \ldots\ . \tag{186}$$

The reader may notice that the on-off sources discussed in Section 1.19 could be modeled as Bernoulli processes where the on-periods are represented by consecutive successes of Bernoulli trials and the off-periods by failures. In this case, for each on-off process, the length of the on- and the off-periods are both geometrically distributed. Accordingly, the **superposition** of $N$ Bernoulli processes with parameter $p$ is another discrete-time stochastic process where the number of arrivals during the different slots are IID and binomial distributed with parameters $N$ and $p$.

**Homework 2.2**

Prove the last statement. □

Another important concept is **merging** of processes which is different from superposition. Let us use a sensor network example to illustrate it. Consider $N$ sensors that are spread around a country to detect certain events. Time is divided into consecutive fixed-length time-slots and a sensor is silent if it does not detect an event in a given time-slot and active (transmitting an alarm signal) if it does. Assume that time-slots during which the $i$th sensor is active follow a Bernoulli process with parameter $p_i$, namely, the probability that sensor $i$ detects an event in a given time-slot is equal to $p_i$, and that the probability of such detection is independent from time-slot to time-slot. We also assume that the $N$ Bernoulli processes associated with the $N$ servers are independent. Assume that an alarm is sound during a time-slot when at least one of the sensors is active. We are interested in the discrete-time process representing alarm sounds. The probability that an alarm is sound in a given time-slot is the probability that at least one of the sensors is active which is one minus the probability that they are all silent. Therefore the probability that the alarm is sound is given by

$$P_a = 1 - \prod_{i=1}^{N}(1 - p_i). \tag{187}$$

Now, considering the independence of the processes, we can realize that the alarms follow a Bernoulli process with parameter $P_a$.

In general, an arrival in the process that results from merging of $N$ Bernoulli processes is the process of time-slots during which at least one of the $N$ processes records an arrival. Unlike superposition in which we are interested in the total number of arrivals, in merging we are only interested to know if there was at least one arrival within a time-slot without any interest of how many arrivals there were in total.

Let us now consider **splitting**. Consider a Bernoulli process with parameter $p$ and then color each arrival, independently of all other arrivals, in red with probability $q$ and in blue with probability $1 - q$. Then in each time-slot we have a red arrival with probability $pq$ and a blue one with probability $p(1 - q)$. Therefore, the red arrivals follow a Bernoulli process with parameter $pq$ and the blue arrivals follow a Bernoulli process with parameter $p(1 - q)$.

### 2.2.2 Poisson Process

The Poisson process is a continuous-time point process which is also memoryless and orderly. It applies to many cases where a certain event occurs at different points in time. Such occurrences of the events could be, for example, arrivals of phone call requests at a telephone exchange. As mentioned above such a process can be described by its *counting process* $\{N(t),\ t \geq 0\}$ representing the total number of occurrences by time $t$.

A counting process $\{N(t)\}$ is defined as *a Poisson process* with rate $\lambda > 0$ if it satisfies the following three conditions.

1. $N(0) = 0$.
2. The number of occurrences in two non-overlapping intervals are independent. That is, for any $s > t > u > v > 0$, the random variable $N(s) - N(t)$, and the random variable $N(u) - N(v)$ are independent. This means that the Poisson process has what is called *independent increments*.
3. The number of occurrences in an interval of length $t$ has a Poisson distribution with mean $\lambda t$.

These three conditions will be henceforth called the *Three Poisson process conditions*.

By definition, the Poisson process $N(t)$ has what is called *stationary increments* [67, 79], that is, for any $t_2 > t_1$, the random variable $N(t_2)-N(t_1)$, and the random variable $N(t_2+u)-N(t_1+u)$ have the same distribution for any $u > 0$. In both cases, the distribution is Poisson with parameter $\lambda(t_2 - t_1)$. Intuitively, if we choose the time interval $\Delta = t_2 - t_1$ to be arbitrarily small (almost a "point" in time), then the probability of having an occurrence there is the same regardless of where the "point" is. Loosely speaking, every point in time has the same chance of having an occurrence. Therefore, occurrences are equally likely to happen at all times. This property is also called *time-homogeneity* [14].

Another important property of the Poisson process is that the inter-arrival times of occurrences is exponentially distributed with parameter $\lambda$. This is shown by considering $s$ to be an occurrence and $T$ the time until the next occurrence, noticing that $P(T > t) = P(N(t) = 0) = e^{-\lambda t}$, and recalling the properties of independent and stationary increments. As a result, the mean inter-arrival time is given by

$$E[T] = \frac{1}{\lambda}. \tag{188}$$

By the memoryless property of the exponential distribution, the time until the next occurrence is always exponentially distributed and therefore, at any point in time, not necessarily at points of occurrences, the future evolution of the Poisson process is independent of the past, and is always probabilistically the same. The Poisson process is therefore memoryless. Actually, the independence of the past can be explained also by the Poisson process property of *independent increments* [79], and the fact that the future evolution is probabilistically the same can also be explained by the stationary increments property.

An interesting paradox emerges when one considers the Poisson process. If we consider a random point in time, independent of a given Poisson process, the time until the next occurrence event has exponential distribution with parameter $\lambda$. Because the Poisson process in reverse is also a Poisson process, then at any point in time, the time passed from the last Poisson occurrence event also has exponential distribution with parameter $\lambda$. Therefore, if we pick a random point in time the mean length of the interval between two consecutive Poisson occurrences must be $1/\lambda + 1/\lambda = 2/\lambda$. How can we explain this phenomenon, if we know that the time between consecutive Poisson occurrences must be exponentially distributed with mean $1/\lambda$? The explanation is that if we pick a point of time at random we are likely to pick an interval that is longer than the average.

**Homework 2.3**

Demonstrate the above paradox as follows. Generate a Poisson process with rate $\lambda = 1$ for a period of time of length $T \geq 10,000$. Pick a point in time from a uniform distribution within the interval [1,10000]. Record the length of the interval (between two consecutive Poisson occurrences) that includes the chosen point in time. Repeat the experiment 1000 times. Compute the average length of the intervals you recorded. □

A **superposition** of a number of Poisson processes is another point process that comprises all the points of the different processes. Another important property of the Poisson process is that superposition of two Poisson processes with parameters $\lambda_1$ and $\lambda_2$ is a Poisson process with parameter $\lambda_1 + \lambda_2$. Notice that in such a case, at any point in time, the time until the next occurrence is a competition between two exponential random variables one with parameter $\lambda_1$ and the other with parameter $\lambda_2$. Let $T$ be the time until the winner of the two occurs, and let $T_1$ and $T_2$ be the time until the next occurrence of the first process and the second process, respectively. Then by (62)

$$P(T > t) = e^{-(\lambda_1+\lambda_2)t}. \tag{189}$$

Thus, the inter-arrival time of the superposition is exponentially distributed with parameter $\lambda_1 + \lambda_2$. This is consistent with the fact that the superposition of the two processes is a Poisson process with parameter $\lambda_1 + \lambda_2$.

**Homework 2.4**

Prove that a superposition of $N$ Poisson processes with parameters $\lambda_1, \lambda_2, \ldots, \lambda_N$, is a Poisson process with parameter $\lambda_1 + \lambda_2 +, \ldots, +\lambda_N$. □

Another interesting question related to superposition of Poisson processes is the question of what is the probability that the next event that occurs will be of a particular process. This is equivalent to the question of having say two exponential random variables $T_1$ and $T_2$ with parameters $\lambda_1$ and $\lambda_2$, respectively, and we are interested in the probability of $T_1 < T_2$. By (63),

$$P(T_1 < T_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \tag{190}$$

Before we introduce further properties of the Poisson process, we shall introduce the following definition: a function $g(\cdot)$ is $o(\Delta t)$ if

$$\lim_{\Delta t \to 0} \frac{g(\Delta t)}{\Delta t} = 0. \tag{191}$$

Examples of functions which are $o(\Delta t)$ are $g(x) = x^v$ for $v > 1$. Sum or product of two functions which are $o(\Delta t)$ is also $o(\Delta t)$, and a constant times a function which is $o(\Delta t)$ is $o(\Delta t)$.

If a counting process $\{N(t)\}$ is *a Poisson process* then, for a small interval $\Delta t$, we have:

1. $P(N(\Delta t) = 0) = 1 - \lambda \Delta t + o(\Delta t)$

2. $P(N(\Delta t) = 1) = \lambda \Delta t + o(\Delta t)$
3. $P(N(\Delta t) \geq 2) = o(\Delta t)$.

The above three conditions will henceforth be called *small interval conditions*. To show the first, we know that $N(\Delta t)$ has a Poisson distribution, therefore

$$P(N(\Delta t) = 0) = e^{-\lambda \Delta t} \tag{192}$$

and developing it into a series gives,

$$P(N(\Delta t) = 0) = 1 - \lambda \Delta t + o(\Delta t). \tag{193}$$

The second is shown by noticing that $P(N(\Delta t) = 1) = \lambda \Delta t P(N(\Delta t) = 0)$ and using the previous result. The third is obtained by $P(N(\Delta t) \geq 2) = 1 - P(N(\Delta t) = 1) - P(N(\Delta t) = 0)$. In fact, these three small interval conditions plus the stationarity and independence properties together with $N(0) = 0$, can serve as an alternative definition of the Poisson process. These properties imply that the number of occurrences per interval has a Poisson distribution.

**Homework 2.5**

Prove the last statement. Namely, show that the three small-interval conditions plus the stationarity and independence properties together with $N(0) = 0$ are equivalent to the Three Poisson Conditions.

**Guide**

Define

$$P_n(t) = P(N(t) = n)$$

Using the assumptions of stationary and independent increments show that

$$P_0(t + \Delta t) = P_0(t) P_0(\Delta t).$$

Therefore

$$\frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = P_0(t) \frac{P_0(\Delta t) - 1}{\Delta t}.$$

From the small interval conditions, we know that $P_0(\Delta t) = 1 - \lambda \Delta t + o(\Delta)$, so let $\Delta t \to 0$ in the above and obtain the differential equation:

$$P_0{}'(t) = -\lambda P_0(t).$$

Consider the boundary condition $P_0(0) = 1$ due to the condition $N(0) = 0$, and solve the differential equation to obtain

$$P_0(t) = e^{-\lambda t}.$$

This proves the Poisson distribution for the case $n = 0$. Now continue the proof for $n > 0$. This will be done by induction, but as a first step, consider $n = 1$ to show that

$$P_1(t) = \lambda t e^{-\lambda t}.$$

Notice the $P_n(t + \Delta t)$ can be obtained by conditioning and un-conditioning (using the Law of Total Probability) on the number of occurrences in the interval $(t, t + \Delta t)$. The interesting events are:

1. no occurrences with probability $1 - \lambda \Delta t + o(\Delta t)$,
2. one occurrence with probability $\lambda \Delta t + o(\Delta t)$,
3. two or more occurrences with probability $o(\Delta t)$.

Considering these events show that

$$P_n(t + \Delta t) = P_n(t)(1 - \lambda \Delta t) + P_{n-1}(t)\lambda \Delta t + o(\Delta t)$$

which leads to

$$\frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(\Delta t)}{\Delta t}.$$

Let $\Delta t \to 0$ in the above and obtain the differential equation:

$$P_n{}'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t).$$

Multiply both sides by $e^{\lambda t}$ and rearrange to obtain

$$\frac{d\{e^{\lambda t} P_n(t)\}}{dt} = \lambda e^{\lambda t} P_{n-1}(t). \quad (194)$$

Then use the result for $P_0(t)$ and the boundary condition of $P_1(0) = 0$ to obtain

$$P_1(t) = \lambda t e^{-\lambda t}.$$

To show that the Poisson distribution holds for any $n$, assume it holds for $n - 1$, i.e.,

$$P_{n-1}(t) = \frac{e^{-\lambda t}(\lambda t)^{n-1}}{(n-1)!}.$$

Now by the latter and (194) obtain

$$\frac{d\{e^{\lambda t} P_n(t)\}}{dt} = \frac{\lambda^n t^{n-1}}{(n-1)!}$$

Then use the latter plus the condition $P_n(0) = 0$ to obtain

$$P_n(t) = \frac{e^{-\lambda t}(\lambda t)^n}{(n)!}.$$

□

In many networking applications, it is of interest to study the effect of **splitting** of packet arrival processes. In particular, we will consider two types of splitting: *random splitting* and *regular splitting*. To explain the difference between the two, consider an arrival process of packets to a certain switch called Switch X. This packet arrival process is assumed to follow a Poisson process with parameter $\lambda$. Some of these packets are then forwarded to Switch A and the others to Switch B. We are interested in the process of packets arriving from Switch X to Switch A, designated *X-A Process*.

Under random splitting, every packet that arrives at Switch X is forwarded to A with probability $p$ and to B with probability $1-p$ independently of any other event associated with other packets. In this case, the packet stream from X to A follows a Poisson process with parameter $p\lambda$.

### Homework 2.6

Prove that under random splitting, the X-A Process is a Poisson process with parameter $p\lambda$.

### Guide

To show that the small interval conditions hold for the X-A Process, let $N_{X-A}(t)$ be the counting process of the X-A process, then

$P(N_{X-A}(\Delta t) = 0) = P(N(\Delta t) = 0) + (1-p)P(N(\Delta t) = 1) + o(\Delta t) = 1 - \lambda\Delta t + (1-p)\lambda\Delta t + o(\Delta t) = 1 - p\lambda\Delta t + o(\Delta t)$,

$P(N_{X-A}(\Delta t) = 1) = pP(N(\Delta t) = 1) + o(\Delta t) = p\lambda\Delta t + o(\Delta t)$,

$P(N_{X-A}(\Delta t) > 1) = o(\Delta t)$,

and the stationarity and independence properties together with $N(0) = 0$ follow from the same properties of the Poisson counting process $N(t)$. □

It may be interesting to notice that the inter-arrival times in the X-A Process are exponentially distributed because they are geometric sums of exponential random variables.

Under regular splitting, the first packet that arrives at Switch X is forwarded to A the second to B, the third to A, the fourth to B, etc. In this case, the packet stream from X to A (the X-A Process) will follows a stochastic process which is a point process where the inter-arrival times are Erlang distributed with parameter $\lambda$ and 2.

### Homework 2.7

1. Prove the last statement.
2. Derive the mean and the variance of the inter-arrival times of the X-A process in the two cases above: random splitting and regular splitting.
3. Consider now 3-way splitting. Derive and compare the mean and the variance of the inter-arrival times for the regular and random splitting cases.
4. Repeat the above for $n$-way splitting and let $n$ increase arbitrarily. What can you say about the burstiness/variability of regular versus random splitting. □

The properties of the Poisson process, namely, independence and time-homogeneity, make the Poisson process able to randomly inspect other continuous-time stochastic processes in a way that the sample it provides gives us enough information on what is called *time-averages*. In other words, its inspections are not biased. Examples of time-averages are the proportion of time a process $X(t)$ is in state $i$, i.e., the proportion of time during which $X(t) = i$. Or the overall mean of the process defined by

$$E[X(t)] = \frac{\int_0^T X(t)dt}{T} \tag{195}$$

for an arbitrarily large $T$. These properties that an occurrence can occur at any time with equal probability, regardless of times of past occurrences, gave rise to the expression a *pure chance process* for the Poisson process.

**Homework 2.8**

Consider a Poisson process with parameter $\lambda$. You know that there was exactly one occurrence during the interval [0,1]. Prove that the time of the occurrence is uniformly distributed within [0,1].

**Guide**

For $0 \leq t \leq 1$, consider

$$P(\text{occurrence within } [0,t) \mid \text{exacly one occurrence within } [0,1])$$

and use the definition of conditional probability. Notice that the latter is equal to:

$$\frac{P(\text{one occurrence within } [0,t) \text{ and no occurrence within } [t,1])}{P(\text{exacly one occurrence within } [0,1])}$$

or

$$\frac{P(\text{one occurrence within } [0,t))P(\text{no occurrence within } [t,1])}{P(\text{exacly one occurrence within } [0,1])}.$$

Then recall that the number of occurrences in any interval of size $T$ has Poisson distribution with parameter $\lambda T$. □

In addition to the Poisson process there are other processes, the so-called *mixing processes* that also has the property of inspections without bias. In particular, Baccelli et al. [8, 9] promoted the use of a point process where the inter-arrival times are IID Gamma distributed for probing and measure packet loss and delay over the Internet. Such a point-process is a mixing process and thus can "see time-averages" with no bias.

## 2.3 Markov Modulated Poisson Process

The stochastic process called Markov modulated Poisson process (MMPP) is a point process that behaves as a Poisson process with parameter $\lambda_i$ for a period of time that is exponentially distributed with parameter $\delta_i$. Then it moves to mode (state) $j$ where it behaves like a Poisson process with parameter $\lambda_j$ for a period of time that is exponentially distributed with parameter $\delta_j$. The parameters are called *mode duration parameters* [95][96],[97]. In general, the MMPP can have an arbitrary number of modes, so it requires a transition probability matrix as an additional set of parameters to specify the probability that it moves to mode $j$ given that it is in mode $i$. However, we are mostly interested in the simplest case of MMPP – the two mode MMPP denoted MMPP(2) and defined by only four parameters: $\lambda_0$, $\lambda_1$, $\delta_0$, and $\delta_1$. The MMPP(2) behaves as a Poisson process with parameter $\lambda_0$ for a period of time that is exponentially distributed with mode duration parameter $\delta_0$. Then moves to mode 1 where it behaves like a Poisson process with mode duration parameter $\lambda_1$ for a period of time that is exponentially distributed with parameter $\delta_1$. Then it switches back to mode 0, etc. alternating between the two modes 0 and 1.

## 2.4 Discrete-time Markov-chains

### 2.4.1 Definitions and Preliminaries

Markov-chains are certain discrete space stochastic processes which are amenable for analysis and hence are very popular for analysis, traffic characterization and modeling of queueing and telecommunications networks and systems. They can be classified into two groups: discrete-time Markov-chains discussed here and continues time Markov-chains discussed in the next section.

A discrete-time Markov-chain is a discrete-time stochastic process $\{X_n, n = 0,\ 1,\ 2,\ \ldots\}$ with the Markov property; namely, that at any point in time $n$, the future evolution of the process is dependent only on the state of the process at time $n$, and is independent of the past evolution of the process. The state of the process can be a scalar or a vector. In this section, for simplicity we will mainly discuss the case where the state of the process is a scalar, but we will also demonstrate how to extend the discussion to a multiple dimension case.

The discrete-time Markov-chain $\{X_n, n = 0,\ 1,\ 2,\ \ldots\}$ at any point in time may take many possible values. The set of these possible values is finite or countable and it is called the state space of the Markov-chain, denoted by $\Theta$. A *time-homogeneous Markov-chain* is a process in which

$$P(X_{n+1} = i \mid X_n = j) = P(X_n = i \mid X_{n-1} = j) \quad \text{for all } n.$$

We will only consider, in this section, Markov-chains which are time-homogeneous.

A discrete-time time-homogeneous Markov-chain is characterized by the property that, for any $n$, given $X_n$, the distribution of $X_{n+1}$ is fully defined regardless of states that occur before time $n$. That is,

$$P(X_{n+1} = j \mid X_n = i) = P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2},\ \ldots). \tag{196}$$

### 2.4.2 Transition Probability Matrix

A Markov-chain is characterized by the so-called *Transition Probability Matrix* $\mathbf{P}$ which is a matrix of one step transition probabilities $P_{ij}$ defined by

$$P_{ij} = P(X_{n+1} = j \mid X_n = i) \ \text{ for all } n. \tag{197}$$

We can observe in the latter that the event $\{X_{n+1} = j\}$ depends only on the state of the process at $X_n$ and the transition probability matrix $\mathbf{P}$.

Since the $P_{ij}s$ are probabilities and since when you transit out of state $i$, you must enter some state, all the entries in $\mathbf{P}$ are non-negatives, less or equal to 1, and the sum of entries in each row of $\mathbf{P}$ must add up to 1.

### 2.4.3 Chapman-Kolmogorov Equation

Having defined the one-step transition probabilities $P_{ij}$ in (197), let us define the $n$-step transition probability from state $i$ to state $j$ as

$$P_{ij}^{(n)} = P(X_n = j \mid X_0 = i). \tag{198}$$

The following is known as the Chapman-Kolmogorov equation:

$$P_{ij}^{(n)} = \sum_{k \in \Theta} P_{ik}^{(m)} P_{kj}^{(n-m)}, \tag{199}$$

for any $m$, such that $0 < m < n$.

Let $\mathbf{P}^{(n)}$ be the matrix that its entries are the $P_{ij}^{(n)}$ values.

**Homework 2.9**

First prove the Chapman-Kolmogorov equation and then use it to prove:

1. $\mathbf{P}^{(k+n)} = \mathbf{P}^{(k)} \times \mathbf{P}^{(n)}$
2. $\mathbf{P}^{(n)} = \mathbf{P}^n$. $\square$

### 2.4.4 Marginal Probabilities

Consider the marginal distribution $\pi_n(i) = P(X_n = i)$ of the Markov-chain at time $n$, over the different states $i \in \Theta$. Assuming that the process started at time 0, the initial distribution of the Markov-chain is $\pi_0(i) = P(X_0 = i)$, $i \in \Theta$. Then $\pi_n(i)$, $i \in \Theta$, can be obtained based on the marginal probability $\pi_{n-1}(i)$ as follows

$$\pi_n(j) = \sum_{k \in \Theta} P_{kj} \pi_{n-1}(k), \tag{200}$$

or based on the initial distribution by

$$\pi_n(j) = \sum_{k \in \Theta} P_{kj}^{(n)} \pi_0(k), \quad j \in \Theta \tag{201}$$

or, in matrix notation

$$\pi_n(j) = \sum_{k \in \Theta} P_{kj}^{(n)} \pi_0(k), \tag{202}$$

Let the vector $\Pi_n$ be defined by $\Pi_n = \{\pi_n(j),\ j = 0,\ 1,\ 2,\ 3,\ \ldots\}$ . The vector $\Pi_n$ can be obtained by

$$\Pi_n = \Pi_{n-1}\mathbf{P} = \Pi_{n-2}\mathbf{P}^2 = \ \ldots\ = \Pi_0\mathbf{P}^n. \tag{203}$$

### 2.4.5 Properties and Classification of States

One state $i$ is said to be *accessible* from a second state $j$ if there exists $n$, $n = 0, 1, 2, \ldots,$ such that

$$P_{ji}^{(n)} > 0. \tag{204}$$

This means that there is a positive probability for the Markov-chain to reach state $i$ at some time in the future if it is now in state $j$.

A state $i$ is said to *communicate* with state $j$ if $i$ is accessible from $j$ and $j$ is accessible from $i$.

**Homework 2.10**

Prove the following:

1. A state communicates with itself.
2. If state $a$ communicates with $b$, then $b$ communicates with $a$.
3. If state $a$ communicates with $b$, and $b$ communicates with $c$, then $a$ communicates with $c$. □

A *communicating class* is a set of states that every pair of states in it communicates with each other.

**Homework 2.11**

Prove that a state cannot belong to two different classes. In other words, two different classes must be disjoint. □

The latter implies that the state space $\Theta$ is divided into a number (finite of infinite) of communicating classes.

**Homework 2.12**

Provide an example of a Markov-chain with three communicating classes. □

A communicating class is said to be *closed* if no state outside the class is accessible from a state that belong to the class.

A Markov-chain is said to be *irreducible* if all the states in its state space are accessible from each other. That is, the entire state space is one communicating class.

A state $i$ has *period* $m$ if $m$ is the greatest common divisor of the set $\{n : P(X_n = i | X_0 = i) > 0\}$. In this case, the Markov-chain can return to state $i$ only in a number of steps that is a multiple of $m$. A state is said to be *aperiodic* if it has a period of one.

**Homework 2.13**

Prove that in a communicating class, it is not possible that there are two states that have different periods. □

Given that the Markov-chain is in state $i$, define *return time* as the random variable representing the next time the Markov-chain returns to state $i$. Notice that the return time is a random variable $R_i$, defined by

$$R_i = \min\{n : X_n = i \mid X_0 = i\}. \tag{205}$$

A state is called *transient* if, given that we start in it, there is a positive probability that we will never return back to it. In other words, state $i$ is transient if $P(R_i < \infty) < 1$. A state is called *recurrent* if it is not transient. Namely, if $P(R_i < \infty) = 1$.

Because in the case of a recurrent state $i$, the probability to return to state $i$ in finite time is one, the process will visit state $i$ infinitely many number of times. However, if $i$ is transient, then the process will visit state $i$ only a geometrically distributed number of times with parameter. (Notice that the probability of "success" is $1 - P(R_i < \infty)$.) In this case the number of visits in state $i$ is finite with probability 1.

**Homework 2.14**

Show that state $i$ is recurrent if and only if

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty.$$

**Guide**

This can be shown by showing that if the condition holds, the Markov-chain will visit state $i$ an infinite number of times, and if it does not hold, the Markov-chain will visit state $i$ a finite number of times. Let $Y_n = J_i(X_n)$, where $J_i(x)$ is a function defined for $x = 0, 1, 2, \ldots$, taking the value 1 if $x = i$, and 0 if $x \neq i$. Notice that $E[J_i(X_n) \mid X_0 = i] = P(X_n = i \mid X_0 = i)$, and consider summing up both sides of the latter. □

**Homework 2.15**

Prove that if state $i$ is recurrent then all the states in a class that $i$ belongs to are recurrent. In other words, prove that recurrence is a class property.

**Guide**

Consider $m$ and $n$, such that $P_{ji}^{(m)} > 0$ and $P_{ij}^{(n)} > 0$, and argue that $P_{ji}^{(m)} P_{ii}^{(k)} P_{ij}^{(n)} > 0$ for some $m, k, n$. Then use the ideas and result of the previous proof. □

**Homework 2.16**

Provide an example of a Markov-chain where $P(R_i < \infty) = 1$, but $E[R_i] = \infty$. □

State $i$ is called *positive recurrent* if $E[R_i]$ is finite. A recurrent state that is not positive recurrent is called *null recurrent*. In a finite state Markov-chain, there are no null recurrent states, i.e., all recurrent states must be positive recurrent. We say that a Markov-chain is *stable* if all its states are positive recurrent. This notion of stability is not commonly used for Markov-chains or stochastic processes in general and it is different from other definitions of stability. It is however consistent with the notion of stability of queueing systems and this is the reason we use it here.

A Markov-chain is said to be aperiodic if all its states are aperiodic.

### 2.4.6 Steady-State Probabilities

Consider an irreducible, aperiodic and stable Markov-chain. Then the following limit exists.

$$\mathbf{\Pi} = \lim_{n\to\infty} \Pi_n = \lim_{n\to\infty} \Pi_0 \mathbf{P}^n \tag{206}$$

and it satisfies

$$\mathbf{\Pi} = row\ of \lim_{n\to\infty} \mathbf{P}^n \tag{207}$$

where $row\ of\ \lim_{n\to\infty} \mathbf{P}^n$ is any row of the matrix $\mathbf{P}^n$ as $n$ approaches $\infty$. All the rows are the same in this matrix at the limit. The latter signifies the fact that the limit $\mathbf{\Pi}$ is independent of the initial distribution. In other words, after the Markov-chain runs for a long time, it forgets its initial distribution and converges to $\mathbf{\Pi}$.

We denote by $\pi_j$, $j = 0, 1, 2, \ldots$, the components of the vector $\mathbf{\Pi}$. That is, $\pi_j$ is the steady-state probability of the Markov-chain to be at state $j$. Namely,

$$\pi_j = \lim_{n\to\infty} \pi_n(j) \quad \text{for all } j. \tag{208}$$

By equation (200), we obtain

$$\pi_n(j) = \sum_{i=0}^{\infty} P_{ij}\pi_{n-1}(i), \tag{209}$$

then by the latter and (208), we obtain

$$\pi_j = \sum_{i=0}^{\infty} P_{ij}\pi_i. \tag{210}$$

Therefore, recalling that $\pi$ is a proper probability distribution, we can conclude that for an irreducible, aperiodic and stable Markov-chain, the steady-state probabilities can be obtained by solving the following steady-state equations:

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij} \quad \text{for all } j, \tag{211}$$

$$\sum_{j=0}^{\infty} \pi_j = 1 \tag{212}$$

and

$$\pi_j \geq 0 \quad \text{for all } j. \tag{213}$$

In this case:

$$\pi_j = \frac{1}{E[R_j]}. \tag{214}$$

To explain the latter, consider a large number of sequential transitions of the Markov-chain denoted $\bar{N}$, and let $R_j(i)$ be the $i$th return time to state $j$. We assume that $\bar{N}$ is large enough so we can neglect edge effects. Let $N(j)$ be the number of times the process visits state $j$ during the $\bar{N}$ sequential transitions of the Markov-chain. Then

$$\pi_j \approx \frac{N(j)}{\bar{N}} \approx \frac{N(j)}{\sum_{k=1}^{N(j)} R_j(k)} \approx \frac{N(j)}{E[R_j]N(j)} = \frac{1}{E[R_j]}.$$

When the state space $\Theta$ is finite, one of the equations in (211) is redundant and replaced by (212).

In matrix notation equation (211) is written as: $\mathbf{\Pi} = \mathbf{\Pi P}$.

Note that if we consider an irreducible, aperiodic and stable Markov-chain, then also a unique non-negative steady-state solution vector $\mathbf{\Pi}$ of the steady-state equation (211) exists. However, in this case, the $j$th component of $\mathbf{\Pi}$, namely $\pi_j$, is not a probability but it is the proportion of time in steady-state that the Markov-chain is in state $j$.

Note also that the steady-state vector $\mathbf{\Pi}$ is called the stationary distribution of the Markov-chain, because if we set $\mathbf{\Pi}_0 = \mathbf{\Pi}$, $\mathbf{\Pi}_1 = \mathbf{\Pi P} = \mathbf{\Pi}$, $\mathbf{\Pi}_2 = \mathbf{\Pi P} = \mathbf{\Pi}$, ..., i.e., $\mathbf{\Pi}_n = \mathbf{\Pi}$ for all $n$.

We know that for an irreducible, aperiodic and stable Markov-chain,

$$\sum_{i=0}^{\infty} P_{ji} = 1.$$

This is because we must go from $j$ to one of the states in one step. Then

$$p_j \sum_{i=0}^{\infty} P_{ji} = p_j.$$

Then by (211), we obtain the following steady-state equations:

$$\pi_j \sum_{i=0}^{\infty} P_{ji} = \sum_{i=0}^{\infty} \pi_i P_{ij} \quad \text{for } j = 0, 1, 2 \ldots\ . \tag{215}$$

These equations are called *global balance equations*. Equations of this type are often used in queueing theory. Intuitively, they can be explained as requiring that the long-term frequency of transitions out of state $j$ should be equal to the long term frequency of transitions into state $j$.

**Homework 2.17**

1. Show that a discrete-time Markov-chain (MC) with two states where the rows of the transition probability matrix are identical is a Bernoulli process.
2. Prove that in any finite MC, at least one state must be recurrent.
3. Provide examples of MCs defined by their transition probability matrices that their states (or some of the states) are periodic, aperiodic, transient, null recurrent and positive recurrent. Provide examples of irreducible and reducible (not irreducible) and of stable and unstable MCs. You may use as many MCs as you wish to demonstrate the different concepts.
4. For different $n$ values, choose an $n \times n$ transition probability matrix $\mathbf{P}$ and an initial vector $\mathbf{\Pi}_0$. Write a program to compute $\mathbf{\Pi}_1$, $\mathbf{\Pi}_2$, $\mathbf{\Pi}_3$, ... and demonstrate convergence to a limit in some cases and demonstrate that the limit does not exist in other cases.
5. Prove equation (214).
6. Consider a binary communication channel between a transmitter and a receiver where $B_n$ is the value of the $n$th bit at the receiver. This value can be either equal to 0, or equal to 1. Assume that the event [a bit to be erroneous] is independent of the value received and only depends on whether or not the previous bit is erroneous or correct. Assume the following:
   $P(B_{n+1}$ is erroneous $\mid B_n$ is correct$) = 0.0001$
   $P(B_{n+1}$ is erroneous $\mid B_n$ is erroneous$) = 0.01$
   $P(B_{n+1}$ is correct $\mid B_n$ is correct$) = 0.9999$
   $P(B_{n+1}$ is correct $\mid B_n$ is erroneous$) = 0.99$
   Compute the steady-state error probability. □

### 2.4.7 Birth and Death Process

In many real life applications, the state of the system sometimes increases by one, and at other times decreases by one, and no other transitions are possible. Such a discrete-time Markov-chain $\{X_n\}$ is called a *birth-and-death process.* In this case, $P_{ij} = 0$ if $|i-j| > 1$ and $P_{ij} > 0$ if $|i-j| = 1$.

Then by the first equation of (211), we obtain,

$$p_0 P_{01} = p_1 P_{10}.$$

Then substituting the latter in the second equation of (211), we obtain

$$p_1 P_{12} = p_2 P_{21}.$$

Continuing in the same way, we obtain

$$p_i P_{i,i+1} = p_{i+1} P_{i+1,i}, \quad i = 0, 1, 2, \dots \; . \tag{216}$$

These equations are called *local balance equations.* They together with the normalizing equation

$$\sum_{i=1}^{\infty} p_i$$

constitute a set of steady-state equations for the steady-state probabilities. They are far simpler than (211).

**Homework 2.18**

Solve the local balance equations together with the normalizing equations for the $p_i$, $i = 0, 1, 2, \ldots$ .

**Guide**

Recursively, write all $p_i$, $i = 0, 1, 2, 3, \ldots$ in terms of $p_0$. Then use the normalizing equation and isolate $p_0$. □

### 2.4.8 Reversibility

Consider an irreducible, aperiodic and stable Markov-chain $\{X_n\}$. Assume that this Markov-chain has been running for a long time to achieve stationarity with transition probability matrix $\mathbf{P} = [P_{ij}]$, and consider the process $X_n,\ X_{n-1},\ X_{n-2},\ \ldots$, going back in time. This reversed process is also a Markov-chain because $X_n$ has dependence relationship only with $X_{n-1}$ and $X_{n+1}$ and conditional on $X_{n+1}$, it is independent of $X_{n+2},\ X_{n+3},\ X_{n+4},\ \ldots$ . Therefore,

$$P(X_{n-1} = j \mid X_n = i) = P(X_{n-1} = j \mid X_n = i, X_{n+1} = i_{n+1}, X_{n+2} = i_{n+2},\ \ldots).$$

In the following we derive the transition probability matrix, denoted $\mathbf{Q} = [Q_{ij}]$ of the process $\{X_n\}$ in reverse. Accordingly Define

$$Q_{ij} = P(X_n = j \mid X_{n+1} = i). \tag{217}$$

By the definition of conditional probability, we obtain,

$$Q_{ij} = \frac{P(X_n = j \cap X_{n+1} = i)}{P(X_{n+1} = i)} \tag{218}$$

or

$$Q_{ij} = \frac{P(X_n = j)P(X_{n+1} = i \mid X_n = j)}{P(X_{n+1} = i)} \tag{219}$$

and if $\pi_j$ denotes the steady-state probability of the Markov-chain $\{X_n\}$ to be in state $j$, and let $n \to \infty$, we obtain

$$Q_{ij} = \frac{\pi_j P_{ji}}{\pi_i}. \tag{220}$$

A Markov-chain is said to be *time reversible* if $Q_{ij} = P_{ij}$ for all $i$ and $j$. Substituting $Q_{ij} = P_{ij}$ in (220), we obtain,

$$\pi_i P_{ij} = \pi_j P_{ji} \text{ for all } i, j. \tag{221}$$

The set of equations (221) is also a necessary and sufficient condition for time reversibility. This set of equations is called the *detailed balance conditions*. In other words, a necessary and sufficient condition for reversibility is that there exists a solution that sums up to unity for the detailed balance conditions. Furthermore, if such solution exists, it is the stationary probability of the Markov chain, namely, it also solves also the global balance equations.

Intuitively, a Markov-chain $X_n$ is time-reversible if for a large $k$ (to ensure stationarity) the Markov-chain $X_k,\ X_{k+1}, X_{k+2}\ \ldots$ is statistically the same as the process $X_k,\ X_{k-1}, X_{k-2}\ \ldots$. In other words, by considering the statistical characteristics of the two processes, you cannot tell which one is going forward and which is going backward.

**Homework 2.19**

Provide an example of a Markov-chain that is time reversible and another one that is not time reversible. □

### 2.4.9 Multi-Dimensional Markov-chains

So far, we discussed single dimensional Markov-chains. If the state space is made of finite vectors instead of scalars, we can easily convert them to scalars and proceed with the above described approach. For example, if the state-space is (0,0) (0,1) (1,0) (1,1) we can simply change the names of the states to 0,1,2,3 by assigning the values 0, 1, 2 and 3 to the states (0,0), (0,1), (1,0) and (1,1), respectively. In fact we do not even have to do it explicitly. All we need to do is to consider a $4 \times 4$ transition probability matrix as if we have a single dimension Markov-chain. Let us now consider an example of a multi-dimensional Markov-chain.

Consider a bit-stream transmitted through a channel. Let $Y_n = 1$ if the $n$th bit is received correctly, and let $Y_n = 0$ if the $n$th bit is received incorrectly. Assume the following

$$P(Y_n = i_n \mid Y_{n-1} = i_{n-1}, Y_{n-2} = i_{n-2}) \\ = P(Y_n = i_n \mid Y_{n-1} = i_{n-1}, Y_{n-2} = i_{n-2}, Y_{n-3} = i_{n-3}, Y_{n-4} = i_{n-4},\ \ldots).$$

$$P(Y_n = 0 \mid Y_{n-1} = 0, Y_{n-2} = 0) = 0.9$$
$$P(Y_n = 0 \mid Y_{n-1} = 0, Y_{n-2} = 1) = 0.7$$
$$P(Y_n = 0 \mid Y_{n-1} = 1, Y_{n-2} = 0) = 0.6$$
$$P(Y_n = 0 \mid Y_{n-1} = 1, Y_{n-2} = 1) = 0.001.$$

By the context of the problem, we have

$$P(Y_n = 1) = 1 - P(Y_n = 0)$$

so,

$$P(Y_n = 1 \mid Y_{n-1} = 0, Y_{n-2} = 0) = 0.1$$
$$P(Y_n = 1 \mid Y_{n-1} = 0, Y_{n-2} = 1) = 0.3$$
$$P(Y_n = 1 \mid Y_{n-1} = 1, Y_{n-2} = 0) = 0.4$$
$$P(Y_n = 1 \mid Y_{n-1} = 1, Y_{n-2} = 1) = 0.999.$$

**Homework 2.20**

Explain why the process $\{Y_n\}$ is not a Markov-chain. □

Now define the $\{X_n\}$ process as follows:
$X_n = 0$ if $Y_n = 0$ and $Y_{n-1} = 0$.

$X_n = 1$ if $Y_n = 0$ and $Y_{n-1} = 1$.
$X_n = 2$ if $Y_n = 1$ and $Y_{n-1} = 0$.
$X_n = 3$ if $Y_n = 1$ and $Y_{n-1} = 1$.

**Homework 2.21**

Explain why the process $\{X_n\}$ is a Markov-chain, produce its transition probability matrix, and compute its steady-state probabilities. □

## 2.5 Continuous Time Markov-chains

### 2.5.1 Definitions and Preliminaries

A continuous-time Markov-chain is a continuous-time stochastic process $\{X_t\}$. At any point in time $t$, $\{X_t\}$ describes the state of the process which is discrete. We will consider only continuous-time Markov-chain where $X_t$ takes values that are nonnegative integer. The time between changes in the state of the process is exponentially distributed. In other words, the process stays constant for an exponential time duration before changing to another state.

In general, a continuous-time Markov-chain $\{X_t\}$ is defined by the property that for all real numbers $s \geq 0$, $t \geq 0$ and $0 \leq v < s$, and integers $i \geq 0$, $j \geq 0$ and $k \geq 0$,

$$P(X_{t+s} = j \mid X_t = i, X_v = k_v, v \leq t) = P(X_{t+s} = j \mid X_t = i). \tag{222}$$

That is, the probability distribution of the future values of the process $X_t$, represented by $X_{t+s}$, given the present value of $X_t$ and the past values of $X_t$ denoted $X_v$, is independent of the past and depends only on the present.

A general continuous-time Markov-chain can also be defined as a continuous-time discrete space stochastic process with the following properties.

1. Each time the process enters state $i$, it stays at that state for an amount of time which is exponentially distributed with parameter $\delta_i$ before making a transition into a different state.
2. When the process leaves state $i$, it enters state $j$ with probability denoted $P_{ij}$. The set of $P_{ij}$s must satisfy the following:

$$\begin{array}{ll} (1) & P_{ii} = 0 \;\text{ for all } i \\ (2) & \sum_j P_{ij} = 1. \end{array}$$

An example of a continuous-time Markov-chain is a Poisson process with rate $\lambda$. The state at time $t$, $\{X_t\}$ can be the number of occurrences by time $t$ which is the counting process $N(t)$. In this example of the Poisson counting process $\{X_t\} = N(t)$ increases by one after every exponential time duration with parameter $\lambda$.

Another example is the so-called *pure birth process* $\{X_t\}$. It is a generalization of the counting Poisson process. Again $\{X_t\}$ increases by one every exponential amount of time but here, instead of having a fixed parameter $\lambda$ for each of these exponential intervals, this parameter depends of the state of the process and it is denoted $\delta_i$. In other words, when $\{X_t\} = i$, the time until the next occurrence in which $\{X_t\}$ increases from $i$ to $i + 1$ is exponentially distributed with parameter $\delta_i$. If we set $\delta_i = \lambda$ for all $i$, we have the Poisson counting process.

### 2.5.2 Birth and Death Process

As in the case of the discrete-time Markov chain, in many real-life applications, such as various queueing systems, that lend themselves to continuous-time Markov-chain modelling, the state of the system in one point in time sometimes increases by one, and at other times decreases by one, but never increase or decrease by more than one at one time instance. Such a continuous-time Markov-chain $\{X_t\}$, as its discrete-time counterpart, is called a *birth-and-death process*. In such a process, the time between occurrences in state $i$ is exponentially distributed, with parameter $\delta_i$, and at any point of occurrence, the process increases by one (from its previous value $i$ to $i + 1$) with probability $\upsilon_i$ and decreases by one (from $i$ to $i - 1$) with probability $\vartheta_i = 1 - \upsilon_i$. The transitions from $i$ to $i + 1$ are called *births* and the transitions from $i$ to $i - 1$ are called *deaths*. Recall that the mean time between occurrences, when in state $i$, is $1/\delta_i$. Hence, the birth rate in state $i$, denoted $b_i$, is given by

$$b_i = \delta_i \upsilon_i$$

and the death rate ($d_i$) is given by

$$d_i = \delta_i \vartheta_i.$$

Summing up these two equations gives the intuitive result that the total rate at state $i$ is equal to the sum of the birth-and-death rates. Namely,

$$\delta_i = b_i + d_i$$

and therefore the mean time between occurrences is

$$\frac{1}{\delta_i} = \frac{1}{b_i + d_i}.$$

**Homework 2.22**

Show the following:

$$\vartheta_i = \frac{d_i}{b_i + d_i}$$

and

$$\upsilon_i = \frac{b_i}{b_i + d_i}. \quad \square$$

Birth-and-death processes apply to queueing systems where customers arrive one at a time and depart one at a time. Consider for example a birth-and-death process with the death rate higher than the birth rate. Such a process could model, for example, a stable single-server queueing system.

### 2.5.3 First Passage Time

An important problem that has applications in many fields, such as biology, finance and engineering, is how to derive the distribution or moments of the time it takes for the process to transit from state $i$ to state $j$. In other words, given that the process is in state $i$ find the distribution of a random variable representing the time it takes to enter state $j$ for the first time. This random variable is called the *first passage time from $i$ to $j$*. Let us derive the mean of the first passage time from $i$ to $j$ in a birth-and-death process for the case $i < j$. To solve this problem we start with a simpler one. Let $U_i$ be the mean passage time to go from $i$ to $i+1$. Then

$$U_0 = \frac{1}{b_0}. \tag{223}$$

and

$$U_i = \frac{1}{\delta_i} + \vartheta_i[U_{i-1} + U_i]. \tag{224}$$

**Homework 2.23**

Explain equations (223) and (224).

**Guide**

Notice that $U_{i-1}$ is the mean passage time to go from $i-1$ to $i$, so $U_{i-1}+U_i$ is the mean passage time to go from $i-1$ to $i+1$. Equation (224) essentially says that $U_i$ the mean passage time to go from $i$ to $i+1$ is equal to the mean time the process stays in state $i$ (namely $1/\delta_i$), plus the probability to move from $i$ to $i-1$, times the mean passage time to go from $i-1$ to $i+1$. Notice that the probability of moving from $i$ to $i+1$ is not considered because if the process moves from $i$ to $i+1$ when it completes its sojourn in state $i$ then the process reaches the target (state $i+1$), so no further time needs to be considered. □

Therefore,

$$U_i = \frac{1}{b_i + d_i} + \frac{d_i}{b_i + d_i}[U_{i-1} + U_i] \tag{225}$$

or

$$U_i = \frac{1}{b_i} + \frac{d_i}{b_i}U_{i-1}. \tag{226}$$

Now we have a recursion by which we can obtain $U_0, U_1, U_2, \ldots$, and the mean first passage time between $i$ and $j$ is given by the sum

$$\sum_{k=i}^{j} U_k.$$

**Homework 2.24**

Let $b_i = \lambda$ and $d_i = \mu$ for all $i$, derive a closed form expression for $U_i$. □

### 2.5.4 Transition Probability Function

Define the *transition probability function* $P_{ij}(t)$ as the probability that given that the process is in state $i$ at time $t_0$, then a time $t$ later, it will be in state $j$. That is,

$$P_{ij}(t) = P[X(t_0 + t) = j \mid X(t_0) = i]. \quad (227)$$

The continuous time version of the Chapman-Kolmogorov equations are

$$P_{ij}(t+\tau) = \sum_{n=0}^{\infty} P_{in}(t)P_{nj}(\tau) \quad \text{for all } t \geq 0, \tau \geq 0. \quad (228)$$

Using the latter to derive the limit

$$\lim_{\triangle t \to 0} \frac{P_{ij}(t+\triangle t) - P_{ij}(t)}{\triangle t}$$

we obtain the so called Kolmogorov's Backward Equations:

$$P'_{ij}(t) = \sum_{n \neq i} \delta_i P_{in} P_{nj}(t) - \delta_i P_{ij}(t) \text{ for all } i,j \text{ and } t \geq 0. \quad (229)$$

For a birth-and-death process the latter becomes

$$P'_{0j}(t) = b_0\{P_{1j}(t) - P_{0j}(t)\} \quad (230)$$

and

$$P'_{ij}(t) = b_i P_{i+1,j}(t) + d_i P_{i-1,j}(t) - (b_i + d_i)P_{ij}(t) \text{ for all } i > 0. \quad (231)$$

### 2.5.5 Steady-State Probabilities

As in the case of the discrete-time Markov-chain, define a continuous-time Markov-chain to be called *irreducible* if there is a positive probability for any state to reach every state, and we define a continuous-time Markov-chain to be called *positive recurrent* if for any state, if the process visits and then leaves that state, the random variable that represents the time it returns to that state has finite mean. As for discrete-time Markov-chains, a continuous-time Markov-chain is said to be *stable* if all its states are positive recurrent.

Henceforth we only consider continuous-time Markov-chains that are irreducible, aperiodic and stable. Then the limit of $P_{ij}(t)$ as $t$ approaches infinity exists, and we define

$$\pi_j = \lim_{t \to \infty} P_{ij}(t). \quad (232)$$

The $\pi_j$ values are called steady-state probabilities or stationary probabilities of the continuous-time Markov-chain. In particular, $\pi_j$ is the steady-state probability of the continuous-time Markov-chain to be at state $j$. We shall now describe how the steady-state probabilities $\pi_j$ can be obtained.

We now construct the matrix $\mathbf{Q}$ which is called the *infinitesimal generator* of the continuous-time Markov-chain. The matrix $\mathbf{Q}$ is a matrix of one step infinitesimal rates $Q_{ij}$ defined by

$$Q_{ij} = \delta_i P_{ij} \text{ for } i \neq j \quad (233)$$

and

$$Q_{ii} = -\sum_{j \neq i} Q_{ij}. \tag{234}$$

**Remarks:**

- The state-space can be finite or infinite and hence the matrices **P** and **Q** can also be finite or infinite.
- In Eq. (233), $Q_{ij}$ is the product of the rate to leave state $i$ and the probability of transition to state $j$ from state $i$ which is the rate of transitions from $i$ to $j$.

To obtain the steady-state probabilities $\pi_j$s, we solve the following set of steady-state equations:

$$0 = \sum_i \pi_i Q_{ij} \quad \text{for all } j \tag{235}$$

and

$$\sum_j \pi_j = 1. \tag{236}$$

Denoting $\mathbf{\Pi} = [\pi_0, \pi_1, \pi_2, \ \ldots]$, Eq. (235) can be written as

$$0 = \mathbf{\Pi Q}. \tag{237}$$

To explain Eqs. (235), notice that, by (233) and (234), for a particular $j$, the equation

$$0 = \sum_i \pi_i Q_{ij} \tag{238}$$

is equivalent to the equations

$$\pi_j \sum_{i \neq j} Q_{ji} = \sum_{i \neq j} \pi_i Q_{ij} \tag{239}$$

or

$$\pi_j \sum_{i \neq j} \delta_j P_{ji} = \sum_{i \neq j} \pi_i \delta_i P_{ij} \tag{240}$$

which give the following global balance equations if we consider all $j$.

$$\pi_j \sum_{i \neq j} \delta_j P_{ji} = \sum_{i \neq j} \pi_i \delta_i P_{ij} \quad \text{for all } j, \tag{241}$$

or using the $Q_{ij}$ notation,

$$\pi_j \sum_{i \neq j} Q_{ji} = \sum_{i \neq j} \pi_i Q_{ij} \quad \text{for all } j. \tag{242}$$

The quantity $\pi_i Q_{ij}$ which is the steady-state probability of being in state $i$ times the infinitesimal rate of a transition from state $i$ to state $j$ is called the *probability flux* from state $i$ to state $j$. Eq. (238) says that the total probability flux from all states into state $j$ is equal to the total probability flux out of state $j$ to all other states. To explain this equality, consider a long period of time $L$. Assuming the process returns to all states infinitely many times, during a long time period $L$, the number of times the process moves into state $j$ is equal (in the limit $L \to \infty$) to the number of times the process moves out of state $j$. This leads to Eq. (240) with the factor

$L$ in both sides. The concept of probability flux is equivalent to the concept of the long-term frequency of transitions discussed above in the context of discrete-time Markov chains.

Similar to the case of discrete-time Markov-chains, the set of equations (235) and (236) is dependent and one of the equations in (235) is redundant in the finite state space case.

For continuous-time birth-and-death processes, $Q_{ij} = 0$ for $|i - j| > 1$. As in the discrete-time case, under this special condition, the global balance equations (241) can be simplified to the local balance equations. We start with the first equation of (241) and using the condition $Q_{ij} = 0$ for $|i - j| > 1$, we obtain

$$\pi_0 Q_{01} = \pi_1 Q_{10} \tag{243}$$

The second equation is

$$\pi_1[Q_{10} + Q_{12}] = \pi_0 Q_{01} + \pi_2 Q_{21}. \tag{244}$$

Then Eq. (244) can be simplified using (243) and we obtain

$$\pi_1 Q_{12} = \pi_2 Q_{21}. \tag{245}$$

In a similar way, by repeating the process, we obtain the following local balance equations.

$$\pi_i Q_{i,i+1} = \pi_{i+1} Q_{i+1,i} \quad i = 0, 1, 2, \ \ldots \ . \tag{246}$$

### 2.5.6 Multi-Dimensional Continuous Time Markov-chains

The extension discussed earlier regarding multi-dimensional discrete-time Markov-chains applies also to the case of continuous-time Markov-chains. If the state-space is made of finite vectors instead of scalars, as discussed, there is a one-to-one correspondence between vectors and scalars, so a multi-dimensional continuous-time Markov-chain can be converted to a single-dimension continuous-time Markov-chain and we proceed with the above described approach that applies to the single dimension.

### 2.5.7 The Curse of Dimensionality

In many applications, the **Q** matrix is too large, so it may not be possible to solve the steady-state equations (235) in reasonable time. Actually, the case of a large state-space (or large **Q** matrix) is common in practice.

This is often occur when the application lead to a Markov-chain model that is of high dimensionality. Consider for example a 49 cell GSM mobile network, and assume that every cell has 23 voice channels. Assuming Poisson arrivals and exponential call time duration and cell sojourn times. Then this cellular mobile network can be modeled as a continuous time Markov-chain with each state representing the number of busy channels in each cell. In this case, the number of states is equal to $24^{49}$, so a numerical solution of the steady-state equations is computationally prohibitive.

### 2.5.8 Simulations

When a numerical solution is not possible, we often rely on simulations. Fortunately, due to the special structure of the continuous-time Markov-chain together with a certain property of the

Poisson process called PASTA (Poisson Arrivals See Time Averages), simulations of continuous time Markov-chain models can be simplified and expedited so they lead to accurate results. To explain the PASTA property, consider a stochastic process for which steady-state probabilities exist. If we are interested in obtaining certain steady-state statistical characteristics of the process (like the $\pi_i$ in a continuous-time Markov-chain), we could inspect the entire evolution of the process (in practice, for a long enough time period), or we could use an independent Poisson inspector. (We already discussed the property of the Poisson process to see time-averages.) The PASTA principle means that if the arrivals follow a Poisson process, we do not need a separate Poisson inspector, but we could inspect the process at occurrences of points in time just before points of arrivals.

Note that in practice, since we are limited to a finite number of inspections, we should choose a Poisson process that will have sufficient number of occurrences (inspections) during the simulation of the stochastic process we are interested in obtaining its steady-state statistics.

In many cases, when we are interested in steady-state statistics of a continuous time Markov-chain, we can conveniently find a Poisson process which is part of the continuous-time Markov-chain we are interested in and use it as a Poisson inspector. For example, if we consider a queueing system in which the arrival process follows a Poisson process, such process could be used for times of arrivals of the inspector if it, at any inspection, does not count (include) its own particular arrival. In other words, we consider a Poisson inspector that arrives just before its own arrival occurrences.

See Chapter 4 for more information and examples on simulations.

### 2.5.9 Reversibility

We have discussed the **time reversibility** concept in the context of discrete-time Markov-chains. In the case of a continuous-time Markov-chain the notion of time reversibility is similar. If you observe the process $X_t$ for a large $t$ (to ensure stationarity) and if you cannot tell from its statistical behavior if it is going forward or backward, it is time reversible.

Consider stationary continuous-time Markov-chain that has a unique steady-state solution. Its $[P_{ij}]$ matrix characterizes a discrete-time Markov-chain. This discrete-time Markov-chain, called the *embedded chain* of our continuous-time Markov-chain, has $[P_{ij}]$ as its transition probability matrix. This embedded chain is in fact the sequence of states that our original continuous-time chain visits where we ignore the time spent in each state during each visit to that state. We already know the condition for time reversibility of the embedded chain, so consider our continuous-time chain and assume that it has been running for a long while, and consider its reversed process going backwards in time. In the following we show that also the reversed process spends an exponentially distributed amount of time in each state. Moreover, we will show that the reverse process spends an exponentially distributed amount of time with parameter $\delta_i$ when in state $i$ which is equal to the time spent in state $i$ by the original process.

$$\begin{aligned} P\{X(t)=i,\ \text{for } t\in[u-v,u] \mid X(u)=i\} &= \frac{P\{X(t)=i,\ \text{for } t\in[u-v,u]\cap X(u)=i\}}{P[X(u)=i]} \\ &= \frac{P[X(u-v)=i]e^{-\delta_i v}}{P[X(u)=i]} = e^{-\delta_i v}. \end{aligned}$$

The last equality is explained by reminding the reader that the process is in steady-state so the probability that the process is in state $i$ at time $(u - v)$ is equal to the probability that the process is in state $i$ at time $u$.

Since the continuous-time Markov-chain is composed of two parts, its embedded chain and the time spent in each state, and since we have shown that the reversed process spends time in each state which is statistically the same as the original process, a condition for time reversibility of a continuous-time Markov-chain is that its embedded chain is time reversible.

As we have learned when we discussed reversibility of stationary discrete-time Markov-chains, the condition for reversibility is the existence of positive $\hat{\pi}_i$ for all states $i$ that sum up to unity that satisfy the detailed balance equations:

$$\hat{\pi}_i P_{ij} = \hat{\pi}_j P_{ji} \text{ for all adjacent } i, j. \tag{247}$$

Recall that this condition is necessary and sufficient for reversibility and that if such a solution exists, it is the stationary probability of the process. The equivalent condition in the case of a stationary continuous-time Markov-chain is the existence of positive $\pi_i$ for all states $i$ that sum up to unity that satisfy the detailed balance equations of a continuous-time Markov-chain, defined as:

$$\pi_i Q_{ij} = \pi_j Q_{ji} \text{ for all adjacent } i, j. \tag{248}$$

**Homework 2.25**

Derive (248) from (247). □

It is important to notice that for a birth-and-death process, its embedded chain is time-reversible. Consider a very long time $L$ during that time, the number of transitions from state $i$ to state $i+1$, denoted $T_{i,i+1}(L)$, is equal to the number of transitions, denoted $T_{i+1,i}(L)$, from state $i + 1$ to $i$ because every transition from $i$ to $i + 1$ must eventually follow by a transition from $i + 1$ to $i$. Actually, there may be a last transition from $i$ to $i + 1$ without the corresponding return from $i + 1$ to $i$, but since we assume that $L$ is arbitrarily large, the number of transitions is arbitrarily large and being off by one transition for an arbitrarily large number of transitions is negligible.

Therefore, for arbitrary large $L$,

$$\frac{T_{i,i+1}(L)}{L} = \frac{T_{i+1,i}(L)}{L}. \tag{249}$$

Since for a birth-and-death process $Q_{ij} = 0$ for $\mid i-j \mid > 1$ and for $i = j$, and since for arbitrarily large $L$, we have

$$\pi_i Q_{i,i+1} = \frac{T_{i,i+1}(L)}{L} = \frac{T_{i+1,i}(L)}{L} = \pi_{i+1} Q_{i+1,i}, \tag{250}$$

so our birth-and-death process is time reversible. This is an important result for the present context because many of the queueing models discussed in this book involve birth-and-death processes.

## 2.6 Renewal Process

An informal way to describe a *renewal process* is a generalization of the Poisson process where the inter-arrival times are not necessarily exponentially distributed. In the renewal process the inter-arrival are positive IID with finite mean, but they are not necessarily exponential. Because of this generalization a renewal process does not have to be memoryless. There are discrete-time and continuous-time renewal processes. In a discrete-time renewal process the inter-arrival times take only positive integer values while in a continuous-time renewal processes the inter-arrival times are positive real valued.

**Homework 2.26**

Provide examples of renewal processes that are not memoryless.

**Guide**

Since the exponential random variable is the only continuous random variable with the memoryless property, take any other positive continuous random variable for the inter-arrival times and you will have a renewal process which is not memoryless. □

**Note:** The definition of a renewal process is sometimes extended to cases that allow a zero value for the inter-arrival times to occur with positive probability. This implies that more than one arrival can occur at the same point in time.

# 3 General Queueing and Teletraffic Concepts

In general, queueing systems may be characterized by complex input process, service time distribution, number of servers (or channels), buffer size (or waiting room) and queue disciplines. In practice, such queueing processes and disciplines are often not amenable to analysis. Nevertheless, insight can be often gained using simpler queueing models. Modelling simplification is often made when the aim is to analyze a complex queueing system or network, such as the Internet, where packets on their ways to their destinations arrive at a router where they are stored and then forwarded according to addresses in their headers. One of the most fundamental elements in this process is the single-server queue (SSQ). One of the aims of telecommunications research is to explain traffic and management processes and their effect on queueing performance. In this section, we briefly cover basic queueing theory concepts. We shall bypass mathematically rigorous proofs and rely instead on simpler intuitive explanations.

## 3.1 Notation

A commonly used shorthand notation, called Kendall notation [51], for such single queue models describes the arrival process, service distribution, the number of servers and the buffer size (waiting room) as follows:

{arrival process}/{service distribution}/{number of servers}/{buffer size}-{queue discipline}

Commonly used characters for the first two positions in this shorthand notation are: D (Deterministic), M (Markovian - Poisson for the arrival process or Exponential for the service time distribution required by each customer), G (General), GI (General and independent), and Geom (Geometric). The fourth position is used for the number of buffer places including the buffer spaces available at the servers. This means that if there are $k$ servers and no additional waiting room is available then $k$ will be written in the fourth position. The fourth position is not used if the waiting room is unlimited. The fifth position is used for the queue discipline. Namely, the order in which the customers are served in the queue. For example: First In First Out (FIFO), Last In First Out (LIFO), Processor Sharing (PS) where all customers in the queue obtain service, and random order (random). The fifth position is not used for the case of the FIFO queue discipline. Notice that the dash notation "-" before the fifth position is used to designate the fifth position. This "-" designation avoids ambiguity in case the fourth position is missing.

For example, D/D/1 denotes an SSQ where both the inter-arrival times and the service times are deterministic. This means that the inter-arrival times are all equal to each other and the service times are all equal to each other. M/M/1 denotes an SSQ with Poisson arrival process and exponential service time with infinite buffer, and FIFO service order. GI/M/1 denotes a generalization of M/M/1 where the arrival process is a renewal process and not necessarily Poisson. Then, G/M/1 is a further generalization of GI/M/1 where the arrival process is not restricted to be a renewal process, i.e., the inter-arrival times do not have to be IID, but can be dependent and not necessarily identically distributed. G/G/1 is a further generalization of G/M/1 where the service times are not necessarily exponentially distributed and may even depend on each other. G/G/1 is the most general infinite buffer FIFO SSQ considered in queueing theory where both the arrival and service processes are general. M/G/$k$/$k$ denotes a k-server queue with Poisson arrivals and general service times without additional waiting room accept at the servers with the arrival process being Poisson. Then the $k$-server queue M/G/$k$/$N$ (with $N \geq k$) also with Poisson arrivals and generally distributed service times is a generalization of M/G/$k$/$k$ where a waiting room that can accommodate up to $N - k$ customers is added. M/G/1-PS denotes a single server processor sharing queue with Poisson arrivals and generally distributed customer service time requirement. Notice that in an M/G/1-PS queue although the service time of a customer/packet starts immediately upon arrival it may continue for a longer time than its service requirement, because the server capacity is always shared among all customers/packets in the system.

## 3.2 Utilization

An important measure for queueing systems performance is the utilization, denoted $\hat{U}$. It is the proportion of time that a server is busy on average. In many systems, the server is paid for its time regardless if it is busy or not. Normally, the time that transmission capacity is not used is time during which money is spent but no revenue is earned. It is therefore important to design systems that will maintain high utilization.

If you have two identical servers and one is busy 0.4 of the time and the other 0.6. Then the utilization is 0.5. We always have that $0 \leq \hat{U} \leq 1$. If we consider an M/M/$\infty$ queue (Poisson arrivals, exponentially distributed service times and infinite servers) and the arrival rate is finite, the utilization is zero because the mean number of busy servers is finite and the mean number of idle servers is infinite.

Consider a G/G/1 queue (that is, an SSQ with arbitrary arrival process and arbitrary service time distribution, with infinite buffer). Let $S$ be a random variable representing the service time and let $E[S] = 1/\mu$, i.e., $\mu$ denotes the service rate. Further, let $\lambda$ be the arrival rate. Assume that $\mu > \lambda$ so that the queue is *stable*, namely, that it will not keep growing forever, and that whenever it is busy, eventually it will reach the state where the system is empty. For a stable G/G/1 queue, we have that that $\hat{U} = \lambda/\mu$. To show the latter let $L$ be a very long period of time. The average number of customers (amount of work) arrived within time period $L$ is: $\lambda L$. The average number of customers (amount of work) that has been served during time period $L$ is equal to $\mu \hat{U} L$. Note that during $L$, customers are being served only during $\hat{U}$ proportion of $L$ when the system is not empty. Since $L$ can be made arbitrarily large and the queue is stable, these two values can be considered equal. Thus, $\mu \hat{U} L = \lambda L$. Hence, $\hat{U} = \lambda/\mu$.

Often, we are interested in the distribution of the number of customers (or jobs or packets) in the system. Consider a G/G/1 queue and let $p_n$ be the probability that there are $n$ customers in the system. Having the utilization, we can readily obtain $p_0$ the probability that the G/G/1 queue is empty. Specifically,

$$p_0 = 1 - \hat{U} = 1 - \lambda/\mu. \tag{251}$$

In the case of a multi-server queue, which in the most general case is denoted by G/G/$k$/$k+n$, the utilization will be defined as the overall average utilization of the individual servers. That is, each server will have its own utilization defined by the proportion of time it is busy, and the utilization of the entire multi-server system will be the average of the individual server utilizations.

## 3.3 Little's Formula

Another important and simple queueing theory result that applies to G/G/1 queue (and to other systems) is known as *Little's Formula* [59, 83, 84]. It has two forms. The first form is:

$$E[Q] = \lambda E[D] \tag{252}$$

where $E[Q]$ and $E[D]$ represent the stationary mean queue-size including the customer in service and the mean delay (system waiting time) of a customer from the moment it arrives until its service is completed, respectively. In the remainder of this book, when we use terms such as *mean queue-size* and *mean delay*, we refer to their values in steady-state, i.e., stationary mean queue-size and delay, respectively.

The second form is:

$$E[N_Q] = \lambda E[W_Q] \tag{253}$$

where $E[N_Q]$ and $E[W_Q]$ represent the mean number of customers in the queue in steady-state excluding the customer in service and the mean delay of a customer, in steady-state, from the moment it arrives until its service commences (waiting time in the queue), respectively.

An intuitive (non-rigorous) way to explain Eq. (252) is by considering a customer that just left the system (completed service). This customer sees behind his/her back on average $E[Q]$ customers. Who are these customers? They are the customers that had been arriving during the time that our customer was in the system. Their average number is $\lambda E[D]$.

Another explanation is based on the amusement part analogy of [12]. Consider an amusement park where customers arrive at a rate of $\lambda$ per time unit. Assume that the park is in stationary condition which implies that it is open 24 hours a day and that the arrival rate $\lambda$ does not change in time. After arriving at the park, a customer spends time at various sites, and then leaves. The mean time a customer spends in the park is represented by $E[D]$. Assume that the park charge every customer one dollar per unit time spent in the park. The mean queue size $E[Q]$ in this case represents the mean number customers in the park in steady state.

Under these assumptions, the rate at which the park earns its income in steady state is $E[Q]$ per unit time because each of the $E[Q]$ customers pays one dollar per unit time. Now let $L$ be an arbitrarily long period of time. The mean number of customers that arrive during $L$ is $\lambda L$. Because $L$ is arbitrarily long, the time a customer spends in the park $E[D]$ is negligible relative to $L$, so we can assume that all the customers that arrive during $L$ also left during $L$. Since a customer on average pays $E[D]$ dollars for its visit in the park, and there are on average $\lambda L$ customers visiting during $L$, the total income earned by the park on average during $L$ is $\lambda LE[Q]$. This gives that the rate at which the park earns its income in steady state per unit time is

$$\frac{\lambda LE[D]}{L} = \lambda E[D].$$

We also know that the latter is equal to $E[Q]$. Therefore, $\lambda E[D] = E[Q]$.

For a graphical proof of Little's Formula for the case of G/G/1 queue see [13]. The arguments there may be summarized as follows. Consider a stable G/G/1 queue that starts at time $t = 0$ with an empty queue. Let $A(t)$ be the number of arrivals up to time $t$, and let $D(t)$ be the number of departures up to time $t$. The queue-size (number in the system) at time $t$ is denoted $Q(t)$ and is given by $Q(t) = A(t) - D(t)$, $t \geq 0$. Let $L$ be an arbitrarily long period of time.

Then the mean queue-size $E[Q]$ is given by

$$E[Q] = \frac{1}{L}\int_0^L Q(t)dt. \tag{254}$$

Also notice that

$$\int_0^L Q(t)dt = \sum_{i=1}^{A(L)} W_i \tag{255}$$

where $W_i$ is the time spent in the system by the $i$th customer. (Notice that since $L$ is arbitrarily large, there have been arbitrarily large number of events during $[0, L]$ where our stable G/G/1 queue became empty, so $A(L) = D(L)$.) Therefore,

$$\frac{1}{L}\int_0^L Q(t)dt = \frac{1}{L}\sum_{i=1}^{A(L)} W_i \tag{256}$$

and realizing that

$$\lambda = A(L)/L, \tag{257}$$

and

$$E[D] = \frac{1}{A(L)}\sum_{i=1}^{A(L)} W_i, \tag{258}$$

we obtain

$$E[Q] = \frac{1}{L}\int_0^L Q(t)dt = \frac{A(L)}{L}\frac{1}{A(L)}\sum_{i=1}^{A(L)} W_i = \lambda E[D]. \tag{259}$$

Little's formula applies to many systems. Its applicability is not limited to single-server queues, or single queue systems, or systems with infinite buffer. However, the system must be in steady-state for Little's formula to apply.

Little's formula is applicable to almost any queueing system in steady state. The system may consist of more than one queue, more than one server, the order does not need to be FIFO, the arrivals do not need to follow Poisson process, and service time do not need to be exponential.

Interestingly, the result $\hat{U} = \lambda/\mu$ for a G/G/1 queue can also be obtained using Little's formula. Let us consider a system to be just the server (excluding the infinite buffer). The mean time a customer spends in this system is $1/\mu$ because this is the mean service time. The mean arrival rate into that system must be equal to $\lambda$ because all the customers that arrive at the queue eventually arrive at the server - nothing is lost. Let us now consider the number of customers at the server, denoted $N_s$. Clearly, $N_s$ can only take the values zero or one, because no more than one customer can be at the server at any point in time. We also know that the steady-state probability $P(N_s = 0)$ is equal to $\pi_0$. Therefore,

$$E[N_s] = 0\pi_0 + 1(1 - \pi_0) = 1 - \pi_0 = \hat{U}.$$

By Little's formula, we have

$$E[N_s] = \lambda(1/\mu),$$

so

$$\hat{U} = \frac{\lambda}{\mu}.$$

Conventional notations in queueing theory for a k-server queue are

$$A = \frac{\lambda}{\mu}$$

and

$$\rho = \frac{A}{k}.$$

Thus, for a G/G/1 queue

$$E[N_s] = \frac{\lambda}{\mu} = \hat{U} = \rho.$$

To obtain (253) from (252), notice that

$$E[Q] = E[N_Q] + E[N_s] = E[N_Q] + \frac{\lambda}{\mu} \tag{260}$$

and

$$E[D] = E[W_Q] + 1/\mu. \tag{261}$$

Substituting (260) and (261) in (252), (253) follows.

Another interesting application of Little's formula relates the blocking probability $P_b$ of a G/G/1/$k$ queue (a G/G/1 queue with a buffer of size $k$) with its server utilization [42, 72]. Again, consider the server as an independent system. Since the mean number of customers in this system is $\hat{U}$, and the arrival rate into this system is $(1 - P_b)\lambda$, we obtain by Little's formula:

$$\hat{U} = (1 - P_b)\lambda\mu^{-1}, \tag{262}$$

where $\mu^{-1}$ is the mean service time. Having $\rho = \lambda/\mu$, we obtain

$$P_b = 1 - \frac{\hat{U}}{\rho}. \tag{263}$$

## 3.4 Traffic

The concept of amount of traffic is very important for in the telecommunications industry. It is important for the network designer, so that networks are properly designed and dimensioned so that they have sufficient capacity to route the traffic to the destinations and meet required quality of service of customers. As traffic on its way from source to destination may traverse several domains (parts of the Internet) each of which controlled and managed by different operator, the operators must negotiate and agree on the level of service they provide and the amount of traffic that is carried between them. To measure the quantity of traffic, we consider the arrival rate of customer calls as well as the amount of network resources the calls require. Clearly, a short phone call will require less resources than a long one. In particular, traffic is measured in units called *erlangs* and it is defined as $\lambda/\mu$, namely, if a quantity of traffic $A$ [erlangs] are offered to a system then $A$ is given by

$$A = \frac{\lambda}{\mu}.$$

In other words, the quantity of traffic in erlangs is the product of the call arrival rate and the mean time that a single server will serve a call (i.e. the mean time that call occupies a server). note that this is the second term mentioned in this book named after the Danish mathematician Agner Krarup Erlang. The first was the Erlang distribution.

Intuitively, as the traffic is $\lambda$ times the mean service time, it represents the number of arrivals per service time. If it is equal to one, it will require on average one server forever. If it is equal to two, it will require two servers forever, etc. Accordingly, if the traffic is $A$, then $A$ is the mean number of servers occupied in steady state by this traffic. This is discussed illustrated by considering the M/M/$\infty$ (or M/G/$\infty$) (see also Chapter 7). Consider an M/M/$\infty$ queue with arrival rate $\lambda$ and service rate $\mu$. Then by Little's formula, the mean number of calls (or customers) in steady state in the systems is equal to $\lambda$ times the mean time a call spends in the system which is $1/\mu$, which is $\lambda/\mu = A$. The mean number of calls in the system is equal to the mean number of servers occupied in steady state.

The M/M/$k$/$k$ (or M/G/$k$/$k$) model is widely used in telecommunications network design. It is applicable to a telephony cable where many circuits (analogous to servers in the model), optical fibre cable comprises many wavelength channels, or a wireless systems that comprises many channels that serve many users simultaneously. In Chapter 8, extends the discussion to distinguish between offered, carried and lost traffic traffic. We also show there that the carried traffic (the proportion of traffic that is not blocked and enters the system to obtain service is equal to the mean number of busy servers in steady state.

In Telephony, the term *holding time*, is often used for the time spent in the system by a call. In such systems, calls that are not blocked, namely, they are admitted to the system, immediately start their service upon their arrivals. Let $h$ be the mean holding time. Because our holding time definition is associate with systems without waiting time, $h$ is also the mean service time, namely

$$h = \frac{1}{\mu}.$$

The total traffic offered to a system is normally generated by many users. If we have $N$ users and the $i$th user generates $A_i$ erlangs. The total traffic generated by the $N$ users is

$$A = \sum_{i=1}^{N} A_i \quad \text{[erlang]}.$$

**Homework 3.1**

Consider a wireless system that provides channels each of which can serve one phone call. There are 100 users making phone calls. Each user makes on average one phone call per hour and the average duration of a phone call is three minutes. What is the total traffic in erlangs that the 100 users generate.

**Guide**

In such questions, it is important first to choose a consistent time unit. Here is is convenient to choose minutes. Accordingly, the arrival rate of each user is $\lambda_i = 1/60$ calls per minute. The mean call duration (or holding time) is 3 minutes, so $A_i = 3/60 = 1/20$ [erlang], and the total traffic is $100/20 = 5$ [erlang]. □

The quantity of traffic measured in erlangs is also called *traffic intensity*. See Section 7.1 for more details. Then another related concept is *traffic volume*. Traffic volume is measured in units of erlang-hour (or erlang-minute, or call-hour or call-minute, etc.) and it is a measure of the traffic processed by a facility during a given period of time. Traffic volume is the product of the traffic intensity and the given time period, namely,

$$\textit{Traffic Volume} = \textit{Traffic Intensity} \times \textit{Time Period}.$$

Let $A, \lambda, \mu$, and $h$ be the traffic intensity, arrival rate, service rate and holding time, respectively. Then

$$A = \frac{\lambda}{\mu} = \lambda h.$$

Let $a_T$ be the number of arrivals during the given period of time $T$. Then the relevant arrival rate can be estimated by

$$\lambda = \frac{a_T}{T}.$$

Let $V_T$ be the traffic volume during $T$. Then

$$V_T = AT = \frac{a_T hT}{T} = a_T h.$$

This leads to a different definition of traffic volume, namely, the product of the number of calls and the mean holding time, and this explains the traffic volume units of call-hour or call-minute. The latter result for $V_T$ can be illustrated by the following example.

Consider a period of time of three hours and during this period of time, 120 calls have arrived and their average holding time is three minutes, then the traffic volume is $3 \times 120 = 360$ call-minute, or 360 erlang-minute, or $360/60 = 6$ erlang-hour. Then, the traffic intensity in erlangs is obtained by dividing the traffic volume in erlang-hour by the number of hours in the given period of time. In our case, the traffic intensity is $6/3 = 2$ [erlangs].

## 3.5 Work Conservation

Another important concept in queuing theory is the concept of *work conservation*. A queuing system is said to be work conservative if a server is never idle whenever there is still work to be done. For example, G/G/1 and G/G/1/$k$ are work conservative. However, a stable G/G/$k$ is not work conservative because a server can be idle while there are customers served by other servers.

## 3.6 PASTA

Many of the queueing models we consider in this book involve Poisson arrival processes. The PASTA property discussed in the previous chapter is important for analysis and simulations of such queueing models. Let us further explain and prove this important property.

The PASTA property implies that arriving customers in steady state will find the number of customers in the system obeying its steady-state distribution. In other words, the statistical characteristics (e.g., mean, variance, distribution) of the number of customers in the system observed by an arrival is the same as those observed by an independent Poisson inspector. This is not true in general. Consider the *lonely person* example of a person lives alone and never has another person comes to his/her house. The period of time s/he stay at home follows some positive distribution and the period of time s/he is outside follows some other positive distribution. When this person comes home s/he always finds that there are no people in the house upon its arrival, but if we use an independent Poisson inspector to evaluate the proportion of time that person is in the house, the inspector will find sometimes that there is one person in the house and in other times that there is no-one in the house. The arrival process of this person is not a Poisson process as the inter-arrival times are not exponentially distributed - each of them is a sum of a period of time the person stays in the house and a period of time s/he stays outside. Also notice that the probability of having arrivals during an arbitrarily small interval $dt$ following an arrival of the person to the home is clearly smaller than the probability of arrivals during an arbitrarily chosen time interval $dt$ because no arrivals occur when the person is at home.

In addition to the assumption of Poisson arrivals, for PASTA to be valid we also need the condition that arrivals after time $t$ are independent of the queue size at time $t$, $Q(t)$. For example, if we have a single-server queue with Poisson arrivals and the service times have the property that the service of a customer must always terminate before the next arrival, then the arrivals always see an empty queue, and, of course, an independent arrival does not. However, in all the queueing systems that we study, this condition holds because normally the server cannot predict the exact time of the next arrival because of the pure chance nature of the Poisson process.

To prove PASTA we consider the limit

$$A_k(t) = \lim_{\Delta t \to 0} P[Q(t) = k \mid \text{ an arrival occurs within } (t, t + \Delta t)].$$

Using Bayes' theorem and the condition that arrivals after time $t$ are independent of $Q(t)$, we obtain that

$$A_k(t) = P[Q(t) = k]. \tag{264}$$

Then, by taking the limit of both sides of (264), we complete the proof that the queue size seen by an arrival is statistically identical to the queue size seen by an independent observer. □

**Homework 3.2**

Prove Eq. (264). □

**Homework 3.3 [13]**

Consider a queueing system with $k$ servers and the total waiting room for customers in the system (including waiting room in the queue and at the server) is $N$, such that $N > k$. This system is always full. That is, it always has $N$ customers in the system. In the beginning there are $N$ customers in the system. When a customer leaves the system immediately another customer arrives. You are given that the mean service time is $1/\mu$.

Find the mean time from the moment a customer arrives until it leaves the system as a function of $N, k$ and $\mu$.

**Solution**

Let $\lambda$ be the arrival rate into the system (which is also equal to the departure rate). We use the term *System S* to denote the system composed of only the servers without the waiting room in the queue outside the servers, and we use the term *System E* to denote the system composed of both the servers and the waiting room in the queue outside the servers. Therefore, $\lambda$ is the arrival rate into *System S* and into *System E*, and it is also equal to the departure rate.

Then, by applying Little's formula to *System S*, we have:

$$\lambda\left(\frac{1}{\mu}\right) = k$$

so

$$\lambda = k\mu.$$

Let $E[D]$ be the mean time from the moment a customer arrives until it leaves the system.

Then, by applying Little's formula to *System E*, we have:

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

Substituting $\lambda$, we obtain

$$E[D] = \frac{N}{k\mu}.$$

□

**Homework 3.4**

Consider an M/G/1/2 queue with arrival rate of one packet per millisecond and the mean service time is one millisecond. It is also given that $E[N_Q]$ the mean number of customers in the queue (not including a customer in the service) is equal to 0.2. What is the blocking probability? A numerical answer is required.

**Solution**

Let $\pi_0$, $\pi_1$ and $\pi_2$ be the steady state probabilities that the total number of customers in this M/G/1/2 queueing system (including those in service and those waiting in the queue) is equal to 0, 1, and 2, respectively. Then $N_Q = 1$ when there are 2 customers in the system, and $N_Q = 0$ when there are either 0 or 1 customers in the system. Therefore:

$$E[N_Q] = \pi_2(1) + (\pi_0 + \pi_1)0 = \pi_2.$$

It is also given that $E[N_Q] = 0.2$. Therefore, $\pi_2 = 0.2$, and since the arrival process is Poisson, by PASTA it is also the blocking probability. □

## 3.7 Bit-rate Versus Service Rate

In many telecommunications design problems, we know the bit-rate of an output link and we can estimate the load on the system in terms of the arrival rate of items of interest that require service. Examples for such items include packets, messages, jobs, or calls. To apply queueing theory for evaluation of quality of service measures, it is necessary, as an intermediate step, to calculate the *service rate* of the system which is the number of such items that the system can serve per unit time. For example, if we know that the average message size is 20 MBytes [MB] and the service rate of the output link of our system is 8 Gigabits per second [Gb/s] and we are interested in the service rate of the link in terms of messages per second, we first calculate the message size in bits which is $20 \times 8 = 160$ Mega bits [Mb] or $160 \times 10^6$ bits. Then the service rate is

$$\frac{8 \times 10^9}{160 \times 10^6} = 50 \text{ messages per second.}$$

**Homework 3.5**

Consider a single server queue with a limited buffer. The arrival rate is one job per second and the mean job size is 0.1 Gigabyte. The server serves the jobs according to a FIFO discipline at a bit-rate rate of one Gb/s (Gigabit per second). Jobs that arrive and find the buffer full are blocked and cleared of the system. The proportion of jobs blocked is 2%. Find the input rate in [Gb/s], the service rate in jobs per second, the queue utilization, and throughput rates in [Gb/s] and in jobs per second. Note that the service rate is the capacity of the server to render service and the throughput is the actual output in either jobs per second or Gb per second.

**Solution**

Since 0.1 Gbyte = 0.8 Gb, then the input rate in [Gb/s] is given by $0.8/1 = 0.8$ [Gb/s].

The service rate is $1/0.8 = 1.25$ jobs per seconds.

The queue utilization is obtained by using Little's formula considering the server as the system and only the traffic that enters the buffer and reaches the server is the input to this system.

Therefore, the queue utilization is the mean queue size in this system (of only the server) and it is obtained by 1(1-0.02)/1.25 = 0.784.

The throughout is obtained by the product of the arrival rate and (1-0.02), so it is equal to 0.98 jobs per second or 0.8(1-0.02) = 0.784 [Gb/s]. □

## 3.8 Queueing Models

In this book, we discuss various queueing models that are amenable to analysis. The analysis is simplest for D/D/ type queues where the inter-arrival and service times are deterministic (fixed values). They will be discussed in the next section. Afterwards, we will consider the so-called Markovian queues. These queues are characterized by the Poisson arrival process, independent exponential service times and independence between the arrival process and the service times. They are denoted by M in the first two positions (i.e., M/M/ $\cdot$ /$\cdot$). Because of the memoryless property of Markovian queues, these queues are amenable to analysis. In fact, they are all continuous-time Markov-chains with the state being the *queue-size* defined as the number in the system $n$ and the time between state transitions is exponential. The reason that these time periods are exponential is that at any point in time, the remaining time until the next arrival, or the next service completion, is a competition between various exponential random variables.

# 4 Simulations

In many cases, analytical solutions are not available, so simulations are used to estimate performance measures. Simulations are also used to evaluate accuracy of analytical approximations.

## 4.1 Confidence Intervals

Regardless of how long we run a simulation involving random processes, we will never obtain the exact mathematical result of a steady-state measure we are interested in. To assess the error of our simulation, we begin by running a certain number, say $n$, of simulation experiments and obtain $n$ observed values, denoted $a_1, a_2, \ \ldots, \ a_n$, of the measure of interest.

Let $\bar{a}$ be the observed mean and $\sigma_a^2$ the observed variance of these $n$ observations. Their values are given by

$$\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{265}$$

and

$$\sigma_a^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})^2. \tag{266}$$

Then the confidence interval of $\bar{a}$, with confidence $\alpha$, $0 \leq \alpha \leq 1$, is given by $(\bar{a} - U_r, \bar{a} + U_r)$, where

$$U_r = \{t_{(1-\alpha)/2,(n-1)}\}\frac{\sigma_a}{\sqrt{n}} \tag{267}$$

where $t_{(1-\alpha)/2,(n-1)}$ is the appropriate percentage point for Student's t-distribution with $n-1$ degrees of freedom. The $t_{(1-\alpha)/2,(n-1)}$ values are available in standard tables. For example: $t_{0.025,5} = 2.57$ and $t_{0.025,10} = 2.23$. That is, if we are interested in 95% confidence and we have $n = 6$ observations, we will use $t_{0.025,5} = 2.57$ to obtain the confidence interval, and if we have $n = 11$ observations, we will use $t_{0.025,10} = 2.23$.

Microsoft (MS) Excel$^{TM}$ provides the function TINV whereby $TINV(1-\alpha, n-1)$ gives the appropriate constant based in t-distribution for confidence $\alpha$ and $n-1$ degrees of freedom. Then the confidence interval of $\bar{a}$, with confidence $\alpha$, $0 \leq \alpha \leq 1$, is given by

$$(\bar{a} - U_r, \bar{a} + U_r),$$

where

$$U_r = TINV(1-\alpha, n-1)\frac{\sigma_a}{\sqrt{n}}.$$

Let us now consider the above-mentioned two examples of $n = 6$ and $n = 11$. Using MS Excel$^{TM}$, $TINV(0.05, 5) = 2.57$ and $TINV(0.05, 10) = 2.23$. That is, if we are interested in 95% confidence and we have $n = 6$ observations, we will use $TINV(0.05, 5) = 2.57$ to obtain the confidence interval, and if we have $n = 11$ observations, we will use $TINV(0.05, 10) = 2.23$.

The larger the number of observations, the smaller is the 95% confidence interval. As certain simulations a very time consuming, a decision needs to be made on the tradeoff between time and accuracy. In many cases, when the simulations are not very time consuming, we can increase the number of observations until required accuracy (length of confidence interval) is achieved.

**Homework 4.1**

From an assignment in Section 1.11 (see also Equation (70)), we have learned that the mean return in every roulette game is negative. In other words, in the long run, one loses money by playing roulette. Recall, that we have made the independence assumption that the results of all roulette games are independent. Now consider the following strategy. Place a bet of $B = 1$ dollar on the red. If you lose, double the bet in the next game. As long as you keep losing, keep doubling and redoubling your bets until you win. The argument for this strategy is that the number of times the player needs to bet until a win is geometrical distributed, and since the latter has a finite mean, eventually, the player wins. Henceforth, we will use the following terminology: a *game* is one bet – one turn of the wheel, and a *sequence* represents the number of games required to achieve a win or to run out of money. The drawback is twofold. Firstly, you may run out of money, and secondly, you may reach the *table limit* (an upper bound on $B$) set by the Casino management. Here we assume for simplicity that there is no table limit. However, we do take into consideration the case of running out of money. In particular, assume that you have $10,000 to bet with at the beginning of a sequence. This means that a sequence cannot be longer than 13 games (why?) and if you run out of money, you lost $8,191 (why?).

Recall the discussion and assignments in the Section 1.10 on how to generate deviates from any probability distribution and use computer simulations to show consistency with the results of the previous assignment in Section 1.11. In particular, run at least six independent runs, each of which of several million sequences where in each sequence the player has fresh starts with $10,000 regardless of the result of the previous sequence. In each run (of million of sequences) calculate the total net winnings (total won minus total lost), and the total number of games played. The ratio between the two will give you an estimate for the mean win/loss per game. Repeating these runs multiple times, will provide a confidence interval for this mean value. Then compare these results with the mean value obtained by Equation (70) and demonstrate consistency between the results. □

We use here the Student's t-distribution (and not Gaussian) because it is the right distribution to use when we attempt to estimate the mean of a population which is normally distributed when we have a small sample size. In fact, the need to estimate such mean based on a small sample gave rise to the development of the Student's t-distribution. In the next section we will guide the reader on how to write queueing simulations for a G/G/1 queue.

## 4.2 Simulation of a G/G/1 Queue

We will now present an example of how to simulate a G/G/1 queue using an approach called *Discrete Event Simulation* [29]. Although the example presented here is for a G/G/1 queue, the principles can be easily extended to multi server and/or finite buffer queues. The first step is to generate a sequence of inter-arrival times and service times in accordance with the given distributions. (Note the discussion in Section 1.10.1 regarding the generation of random deviates.) In our example, starting at time 0, let us consider the following inter-arrival times: 1, 2, 1, 8, 4, 5, . . . , and the following sequence of service times: 4, 6, 4, 2, 5, 1, . . . .

In writing a computer simulation for G/G/1, we aim to fill in the following table for several 100,000s or millions arrivals (rows).

| arrival time | service duration | queue-size on arrival | service starts | service ends | delay |
|---|---|---|---|---|---|
| 1 | 4 | 0 | 1 | 5 | 4 |
| 3 | 6 | 1 | 5 | 11 | 8 |
| 4 | 4 | 2 | | | |
| 12 | 2 | | | | |
| 16 | 5 | | | | |
| 21 | 1 | | | | |

The following comments explain how to fill in the table.

- The arrival times and the service durations values are readily obtained from the inter-arrival and service time sequences.
- Assuming that the previous rows are already filled in, the "queue-size on arrival" is obtained by comparing the arrival time of the current arrivals and the values in the "service ends" column of the previous rows. In particular, the queue size on arrival is equal to the number of customers that arrive before the current customer (previous rows) that their "service ends" time values are greater than the arrival time value of the current arrival.
- The "service starts" value is the maximum of the "arrival time" value of the current arrival and the "service end" value of the previous arrival. Also notice that if the queue size on arrival of the current arrival is equal to zero, the service start value is equal to the "arrival time" value of the current arrival and if the queue size on arrival of the current arrival is greater than zero the service start value is equal to the "service end" value of the previous arrival.
- The "service ends" value is simply the sum of the "service starts" and the "service duration" values of the current arrival.
- The "delay" value is the difference between the "service ends" and the "arrival time" values.

Using the results obtained in the last column, we can estimate the delay distribution and moments in steady-state. However, the "queue-size on arrival" values for all the customers do not, in general, provide directly the steady-state queue-size distribution and moments. To estimate accurately the steady-state queue-size distribution, we will need to have inspections performed by an independent Poisson inspector. Fortunately, due to PASTA, for M/G/1 (including M/M/1 and M/D/1) the "queue-size on arrival" values can be used directly to obtained the steady-state queue-size distribution and moments and a separate Poisson inspector is not required. Observing the queue-size just before the arrivals provides the right inspections for steady-state queue-size statistics. However, if the arrival process does not follow a Poisson process, a separate independent Poisson inspector is required. In such a case, we generate a

Poisson process: $t_1$, $t_2$, $t_3$, $\ldots$, and for each $t_i$, $i = 1, 2, 3, \ldots$ we can invoke the queue-size at time $t_i$, denoted $Q_i$, in a similar way to the one we obtained the "queue-size on arrival" values. The $Q_i$ values are then used to evaluate the queue-size distribution and moments.

An alternative way to evaluate the queue size distribution of a G/G/1 queue is to record the total time spent in each state. If there was an event (arrival or departure) at time $t_j$ when the G/G/1 queue entered state $i$ and the next event (arrival or departure) at $t_k$ when the G/G/1 queue exited state $i$, then the period $t_j - t_k$ is added to a counter recording the total time spent in the state $i$.

**Homework 4.1**

Fill in the above table by hand. □

**Homework 4.2**

Write a computer simulation for a P/P/1 queue (a single-server queue with Pareto inter-arrival and service time distributions) to derive estimates for the mean and distribution of the delay and of the queue-size. Perform the simulations for a wide range of parameter values. Compute confidence interval as described in Section 4. □

**Homework 4.3**

Repeat the simulations, of the previous homework, for a wide range of parameter values, for a U/U/1 queue, defined as a single-server queue with Uniform inter-arrival and service time distributions, and for an M/M/1 queue. For the M/M/1 queue, verify that your simulation results are consistent with respective analytical results. For the U/U/1 queue, use the Poisson inspector approach and the "time recording" approach and verify that the results are consistent. □

**Homework 4.4**

Discuss the accuracy of your estimations in the different cases. □

**Homework 4.5**

Use the principles presented here for a G/G/1 queue simulation to write a computer simulation for a G/G/$k$/$k$ queue. In particular, focus on the cases of an M/M/$k$/$k$ queue and a U/U/$k$/$k$ queue, defined as a $k$-server system without additional waiting room where the inter-arrival and service times are uniformly distributed, and compute results for the blocking probability for these two cases. For a meaningful comparison use a wide range of parameter values. □

# 5 Deterministic Queues

We consider here the simple case where inter-arrival and service times are deterministic. To avoid ambiguity, we assume that if an arrival and a departure occur at the same time, the departure occurs first. Such an assumption is not required for Markovian queues where the queue size process follows a continuous-time Markov-chain because the probability of two events occurring at the same time is zero, but it is needed for deterministic queues. Unlike many of the Markovian queues that we study in this book, for deterministic queues steady-state queue size distribution does not exist because the queue size deterministically fluctuate according to a certain pattern. Therefore, for deterministic queues we will use the notation $P(Q = n)$, normally designating the steady-state probability of the queue-size to be equal to $n$ in cases where such steady-state probability exists, for the proportion of time that there are $n$ customers in the queue, or equivalently, $P(Q = n)$ is the probability of having $n$ in the queue at a randomly (uniformly) chosen point in time. Accordingly, the mean queue size $E[Q]$ will be defined by

$$E[Q] = \sum_{n=0}^{\infty} nP(Q = n).$$

We will use the term blocking probability $P_b$ to designate the proportion of packets that are blocked. To derive performance measures such as mean queue size, blocking probability and utilization, in such deterministic queues, we follow the queue-size process, for a certain transient period, until we discover a pattern (cycle) that repeats itself. Then we focus on a single cycle and obtain the desired measures of that cycle.

## 5.1 D/D/1

If we consider the case $\lambda > \mu$, the D/D/1 queue is unstable. In this case the queue size constantly grows and approaches infinity as $t \to \infty$, and since there are always packets in the queue waiting for service, the server is always busy, thus the utilization is equal to one.

Let us consider now a stable D/D/1 queue, assuming $\lambda < \mu$. Notice that for D/D/1, given our above assumption that if an arrival and a departure occur at the same time, the departure occurs first, the case $\lambda = \mu$ will also be stable. Assume that the first arrival occurs at time $t = 0$. The service time of this arrival will terminate at $t = 1/\mu$. Then another arrival will occur at time $t = 1/\lambda$ which will be completely served at time $t = 1/\lambda + 1/\mu$, etc. This gives rise to a deterministic cyclic process where the queue-size takes two values: 0 and 1 with transitions from 0 to 1 in points of time $n(1/\lambda)$, $n = 0, 1, 2, \ldots$, and transitions from 1 to 0 in points of time $n(1/\lambda) + 1/\mu$, $n = 0, 1, 2, \ldots$ . Each cycle is of time-period $1/\lambda$ during which there is a customer to be served for a time-period of $1/\mu$ and there is no customer for a time-period of $1/\lambda - 1/\mu$. Therefore, the utilization is given by $\hat{U} = (1/\mu)/(1/\lambda) = \lambda/\mu$ which is consistent with what we know about the utilization of G/G/1.

As all the customers that enter the system are served before the next one arrives, the mean queue-size of D/D/1 must be equal to the mean queue-size at the server, and therefore, it is also equal to the utilization. In other words, the queue-size alternates between the values 1 and 0, spending a time-period of $1/\mu$ at state 1, then a time-period of $1/\lambda - 1/\mu$ at state 0, then again $1/\mu$ time at state 1, etc. If we pick a random point in time, the probability that there is one in the queue is given by $P(Q = 1) = (1/\mu)/(1/\lambda)$, and the probability that there are no customers in the queue is given by $P(Q=0) = 1 - (1/\mu)/(1/\lambda)$. Therefore, the mean queue-size is given by $E[Q] = 0P(Q=0) + 1P(Q=1) = (1/\mu)/(1/\lambda) = \hat{U}$.

Moreover, we can show that out of all possible G/G/1 queues, with $\lambda$ being the arrival rate and $\mu$ the service rate, no-one will have lower mean queue-size than D/D/1. This can be shown using Little's formula $E[Q] = \lambda E[D]$. Notice that for each of the relevant G/G/1 queues $E[D] = 1/\mu + E[W_Q] \geq 1/\mu$, but for D/D/1 $E[W_Q] = 0$. Thus, $E[D]$ for any G/G/1 queue must be equal or greater than that of D/D/1, and consequently by Little's formula, $E[Q]$ for any G/G/1 queue must be equal or greater than that of D/D/1.

## 5.2 D/D/$k$

Here we consider deterministic queues with multiple servers. The inter-arrival times are again always equal to $1/\lambda$, and the service time of all messages is equal to $1/\mu$. Again if we consider the case $\lambda > k\mu$, the D/D/$k$ queue is unstable. In this case the queue size constantly increases and approaches infinity as $t \to \infty$, and since there are always more than $k$ packets in the queue waiting for service, all $k$ servers are constantly busy, thus the utilization is equal to one.

Now consider the stable case of $\lambda < k\mu$, so that the arrival rate is below the system capacity. Notice again that given our above assumption that if an arrival and a departure occur at the same time, the departure occurs first, the case $\lambda = k\mu$ will also be stable. Extending the D/D/1 example to a general number of servers, the behavior of the D/D/$k$ queue is analyzed as follows. As $\lambda$ and $\mu$ satisfy the stability condition $\lambda < k\mu$, there must exist an integer $\hat{n}$, $1 \leq \hat{n} \leq k$ such that

$$(\hat{n}-1)\mu < \lambda \leq \hat{n}\mu, \tag{268}$$

or equivalently

$$\frac{\hat{n}-1}{\lambda} < \frac{1}{\mu} \leq \frac{\hat{n}}{\lambda}. \tag{269}$$

**Homework 5.1**

Show that

$$\hat{n} = \left\lceil \frac{\lambda}{\mu} \right\rceil \tag{270}$$

satisfies $1 \leq \hat{n} \leq k$ and (269). Recall that $\lceil x \rceil$ designates the smallest integer greater or equal to x.

**Guide**

Notice that

$$\frac{\hat{n}}{\lambda} = \frac{\left\lceil \frac{\lambda}{\mu} \right\rceil}{\lambda} \geq \frac{\frac{\lambda}{\mu}}{\lambda} = \frac{1}{\mu}.$$

Also,

$$\frac{\hat{n}-1}{\lambda} = \frac{\left\lceil \frac{\lambda}{\mu} \right\rceil - 1}{\lambda} < \frac{\frac{\lambda}{\mu}}{\lambda} = \frac{1}{\mu}.$$

□

The inequality

$$\frac{\hat{n}-1}{\lambda} < \frac{1}{\mu},$$

means that if the first arrival arrives at $t = 0$, there will be additional $\hat{n} - 1$ arrivals before the first customer leaves the system. Therefore, the queue-size increases incrementally taking the value $j$ at time $t = (j - 1)/\lambda$, $j = 1, 2, 3, \ldots, \hat{n}$. When the queue reaches $\hat{n}$ for the first time, which happens at time $(\hat{n} - 1)/\lambda$, the cyclic behavior starts. Then, at time $t = 1/\mu$ the queue-size reduces to $\hat{n} - 1$ when the first customer completes its service. Next, at time $t = \hat{n}/\lambda$, the queue-size increases to $\hat{n}$ and decreases to $\hat{n} - 1$ at time $t = 1/\lambda + 1/\mu$ when the second customer completes its service. This cyclic behavior continuous forever whereby the queue-size increases from $\hat{n}-1$ to $\hat{n}$ at time points $t = (\hat{n}+i)/\lambda$, and decreases from $\hat{n}$ to $\hat{n}-1$ at time points $t = i/\lambda + 1/\mu$, for $i = 0, 1, 2, \ldots$ . The cycle length is $1/\lambda$ during which the queue-size process is at state $\hat{n}$, $1/\mu - (\hat{n} - 1)/\lambda$ of the cycle time, and it is at state $\hat{n} - 1$, $\hat{n}/\lambda - 1/\mu$ of the cycle time. Thus,

$$P(Q = \hat{n}) = \frac{\lambda}{\mu} - (\hat{n} - 1)$$

and

$$P(Q = \hat{n} - 1) = \hat{n} - \frac{\lambda}{\mu}.$$

The mean queue-size $E[Q]$, can be obtained by

$$E[Q] = (\hat{n} - 1)P(Q = \hat{n} - 1) + \hat{n}P(Q = \hat{n})$$

which after some algebra gives

$$E[Q] = \frac{\lambda}{\mu}. \tag{271}$$

**Homework 5.2**

Perform the algebraic operations that lead to (271). □.

This result is consistent with Little's formula. As customers are served as soon as they arrive, the time each of them spends in the system is the service time $1/\mu$ - multiplying it by $\lambda$, gives by Little's formula the mean queue size. Since $E[Q]$ in D/D/$k$ gives the number of busy servers, the utilization is given by

$$\hat{U} = \frac{\lambda}{k\mu}. \tag{272}$$

Notice that Equations (271) and (272) applies also to D/D/$\infty$ for finite $\lambda$ and $\mu$. Eq. (271) gives the mean queue-size of D/D/$\infty$ (by Little's formula, or by following the arguments that led to Eq. (271)) and for D/D/$\infty$, we have that $\hat{U} = 0$ by (272). Also notice that in D/D/$\infty$ there are infinite number of servers and the number of busy servers is finite, so the average utilization per server must be equal to zero.

## 5.3 D/D/$k$/$k$

In D/D/$k$/$k$ there is no waiting room beyond those available at the servers. Recall that to avoid ambiguity, we assume that if an arrival and a departure occur at the same time, the departure occurs first. Accordingly, if $\lambda \leq k\mu$, then we have the same queue behavior as in D/D/$k$ as no losses will occur. The interesting case is the one where $\lambda > k\mu$ and this is the case we focus on. Having $\lambda > k\mu$, or $1/\mu > k/\lambda$, implies that

$$\tilde{n} = \left\lceil \frac{\lambda}{\mu} \right\rceil - k$$

satisfies

$$\frac{k + \tilde{n} - 1}{\lambda} < \frac{1}{\mu} \leq \frac{k + \tilde{n}}{\lambda}.$$

**Homework 5.3**

Prove the last statement.

**Guide**

Notice that

$$\frac{k + \tilde{n}}{\lambda} = \frac{\left\lceil \frac{\lambda}{\mu} \right\rceil}{\lambda} \geq \frac{\frac{\lambda}{\mu}}{\lambda} = \frac{1}{\mu}.$$

Also,

$$\frac{k + \tilde{n} - 1}{\lambda} = \frac{\left\lceil \frac{\lambda}{\mu} \right\rceil - 1}{\lambda} < \frac{\frac{\lambda}{\mu}}{\lambda} = \frac{1}{\mu}.$$

□

### 5.3.1 The D/D/$k$/$k$ process and its cycles

Again, consider an empty system with the first arrival occurring at time $t = 0$. There will be additional $k - 1$ arrivals before all the servers are busy. Notice that because $1/\mu > k/\lambda$, no service completion occurs before the system is completely full. Then $\tilde{n}$ additional arrivals will be blocked before the first customer completes its service at time $t = 1/\mu$ at which time the queue-size decreases from $k$ to $k - 1$. Next, at time $t = (k + \tilde{n})/\lambda$, the queue-size increases to $k$ and reduces to $k - 1$ at time $t = 1/\lambda + 1/\mu$ when the second customer completes its service. This behavior of the queue-size alternating between the states $k$ and $k - 1$ continues until all the first $k$ customers complete their service which happens at time $t = (k - 1)/\lambda + 1/\mu$ when the $k$th customer completes its service, reducing the queue-size from $k$ to $k-1$. Next, an arrival at time $t = (2k + \tilde{n} - 1)/\lambda$ increased the queue-size from $k - 1$ to $k$. Notice that the point in time $t = (2k + \tilde{n} - 1)/\lambda$ is an end-point of a cycle that started at $t = (k - 1)/\lambda$. This cycles comprises two parts: the first is a period of time where the queue-size stays constant at $k$ and all the arrivals are blocked, and the second is a period of time during which no losses occur and the queue-size alternates between $k$ and $k - 1$. Then a new cycle of duration $(k + \tilde{n})/\lambda$ starts and this new cycle ends at $t = (3k + 2\tilde{n} - 1)/\lambda$. In general, for each $j = 1,\ 2,\ 3,\ \ldots$, a cycle of duration $(k+\tilde{n})/\lambda$ starts at $t = (jk+(j-1)\tilde{n}-1)/\lambda$ and ends at $t = ((j+1)k+j\tilde{n}-1)/\lambda$.

### 5.3.2 Blocking probability, mean queue-size and utilization

In every cycle, there are $k + \tilde{n}$ arrivals out of which $\tilde{n}$ are blocked. The blocking probability is therefore

$$P_b = \frac{\tilde{n}}{k+\tilde{n}}.$$

Since

$$k + \tilde{n} = \left\lceil \frac{\lambda}{\mu} \right\rceil,$$

the blocking probability is given by

$$P_b = \frac{\left\lceil \frac{\lambda}{\mu} \right\rceil - k}{\left\lceil \frac{\lambda}{\mu} \right\rceil}. \tag{273}$$

Let $A = \lambda/\mu$, the mean-queue size is obtained using Little's formula to be given by

$$E[Q] = \frac{\lambda}{\mu}(1 - P_b) = \frac{kA}{\lceil A \rceil}. \tag{274}$$

As in D/D/$k$, since every customer that enters a D/D/$k$/$k$ system does not wait in a queue, but immediately enters service, the utilization is given by

$$\hat{U} = \frac{E[Q]}{k} = \frac{A}{\lceil A \rceil}. \tag{275}$$

### 5.3.3 Proportion of time spent in each state

Let us now consider a single cycle and derive the proportion of time spent in the states $k - 1$ and $k$, denoted $P(Q = k - 1)$ and $P(Q = k)$, respectively. In particular, we consider the first cycle of duration

$$\frac{k+\tilde{n}}{\lambda} = \frac{\lceil A \rceil}{\lambda}$$

that starts at time

$$t_s = \frac{k-1}{\lambda}$$

and ends at time

$$t_e = \frac{2k+\tilde{n}-1}{\lambda}.$$

We define the first part of this cycle (the part during which arrivals are blocked) to begin at $t_s$ and to end at the point in time when the $\tilde{n}$th arrival of this cycle is blocked which is

$$t_{\tilde{n}} = t_s + \frac{\tilde{n}}{\lambda} = \frac{k-1+\tilde{n}}{\lambda} = \frac{\lceil A \rceil - 1}{\lambda}.$$

The second part of the cycle starts at $t_{\tilde{n}}$ and ends at $t_e$. The queue-size is equal to $k$ for the entire duration of the first part of the cycle. However, during the second part of the cycle, the queue-size alternates between the values $k$ and $k-1$ creating a series of $k$ mini-cycles each of duration $1/\lambda$. Each of these mini-cycles is again composed of two parts. During the first part of each mini-cycle, $Q = k$, and during the second part of each mini-cycle, $Q = k - 1$. The first mini-cycle starts at time $t_{\tilde{n}}$ and ends at

$$t_{1e} = t_{\tilde{n}} + \frac{1}{\lambda} = \frac{\lceil A \rceil}{\lambda}.$$

The first part of the first mini-cycle starts at time $t_{\tilde{n}}$ and ends at time $1/\mu$, and the second part starts at $1/\mu$ and ends at time $t_{1e}$. Thus, the time spent in each mini-cycle at state $Q = k - 1$ is equal to

$$t_{1e} - \frac{1}{\mu} = \frac{\lceil A \rceil}{\lambda} - \frac{1}{\mu} = \frac{\lceil A \rceil}{\lambda} - \frac{\frac{\lambda}{\mu}}{\lambda} = \frac{\lceil A \rceil - A}{\lambda}.$$

Because there are $k$ mini-cycles in a cycle, we have that the total time spent in state $Q = k-1$ during a cycle is

$$\frac{k(\lceil A \rceil - A)}{\lambda}.$$

Because $P(Q = k - 1)$ is the ratio of the latter to the total cycle duration, we obtain,

$$P(Q = k - 1) = \frac{\frac{k(\lceil A \rceil - A)}{\lambda}}{\frac{\lceil A \rceil}{\lambda}}. \qquad (276)$$

The time spent in state $Q = k$ during each cycle is the total cycle duration minus the time spent in state $Q = k - 1$. Therefore, we obtain

$$P(Q = k) = \frac{\frac{\lceil A \rceil}{\lambda} - \frac{k(\lceil A \rceil - A)}{\lambda}}{\frac{\lceil A \rceil}{\lambda}}. \qquad (277)$$

**Homework 5.4**

1. Show that the results for the queue-size probabilities $P(Q = k - 1)$ and $P(Q = k)$ in (276) and (277) are consistent with the result for the mean queue-size in (274). In other words, show that

$$(k - 1)P(Q = k - 1) + kP(Q = k) = E[Q]$$

or equivalently

$$(k-1)\left\{\frac{\frac{k(\lceil A \rceil - A)}{\lambda}}{\frac{\lceil A \rceil}{\lambda}}\right\} + k\left\{\frac{\frac{\lceil A \rceil}{\lambda} - \frac{k(\lceil A \rceil - A)}{\lambda}}{\frac{\lceil A \rceil}{\lambda}}\right\} = \frac{kA}{\lceil A \rceil}.$$

2. Consider a D/D/3/3 queue with $1/\mu = 5.9$ and $1/\lambda = 1.1$. Start with the first arrival at $t = 0$ and produce a two-column table showing the time of every arrival and departure until $t = 20$, and the corresponding queue-size values immediately following each one of these events.

3. Write a general simulation program for a D/D/$k$/$k$ queue and use it to validate (274) and the results for $P(Q = k - 1)$ and $P(Q = k)$ in (276) and (277). Use it also to confirm the results you obtained for the D/D/3/3 queue.
4. Consider a D/D/1/$n$ queue for $n > 1$. Describe the evolution of its queue-size process and derive formulae for its mean queue-size, mean delay, utilization, and blocking probability. Confirm your results by simulation $\square$.

## 5.4 Summary of Results

The following table summarizes the results on D/D/1, D/D/$k$ and D/D/$k$/$k$. Note that we do not consider the case $\lambda = k\mu$ for which the results for the case $\lambda < k\mu$ are applicable assuming that if a departure and an arrival occur at the same time, the departure occurs before the arrival.

| Model | Condition | $E[Q]$ | $\hat{U}$ |
|---|---|---|---|
| D/D/1 | $\lambda < \mu$ | $\lambda/\mu$ | $\lambda/\mu$ |
| D/D/1 | $\lambda > \mu$ | $\infty$ | 1 |
| D/D/$k$ | $\lambda < k\mu$ | $A = \lambda/\mu$ | $A/k$ |
| D/D/$k$ | $\lambda > k\mu$ | $\infty$ | 1 |
| D/D/$k$/$k$ | $\lambda < k\mu$ | $A$ | $A/k$ |
| D/D/$k$/$k$ | $\lambda > k\mu$ | $kA/\lceil A \rceil$ | $A/\lceil A \rceil$ |

**Homework 5.5**

Justify the following statements.

1. D/D/1 is work conservative.
2. D/D/$k$ is work conservative (following a certain finite initial period) if $\lambda > k\mu$.
3. D/D/$k$ is not work conservative if $\lambda < k\mu$.
4. D/D/$k$/$k$ is not work conservative for all possible values of the parameters $\lambda$ and $\mu$ if we assume that if arrival and departure occurs at the same time, then the arrival occurs before the departure.

**Guide**

Notice that D/D/$k$ is work conservative if there are more than $k$ customers in the system. Notice that for D/D/$k$/$k$ (under the above assumption) there are always periods of time during which less than $k$ servers are busy. $\square$.

# 6 M/M/1

Having considered the straightforward cases of deterministic queues, we will now discuss queues where the inter-arrival and service times are non-deterministic. We will begin with cases where the inter-arrival and service times are independent and exponentially distributed (memoryless). Here we consider the M/M/1 queue where the arrival process follows a Poisson process with parameter $\lambda$ and service times are assumed to be IID and exponentially distributed with parameter $\mu$, and are independent of the arrival process. As M/M/1 is a special case of G/G/1, all the results that are applicable to G/G/1 are also applicable to M/M/1. For example, $\hat{U} = \lambda/\mu$, $p_0 = 1 - \lambda/\mu$ and Little's formula. It is the simplest Markovian queue; it has only a single server and an infinite buffer. It is equivalent to a continuous-time Markov-chain on the states: 0, 1, 2, 3, .... Assuming that the M/M/1 queue-size process starts at state 0, it will stay in state 0 for a period of time that is exponentially distributed with parameter $\lambda$ then it moves to state 1. The time the process stays in state $n$, for $n \geq 1$, is also exponentially distributed, but this time, it is a competition between two exponential random variable, one of which is the time until the next arrival - exponentially distributed with parameter $\lambda$, and the other is the time until the next departure - exponentially distributed with parameter $\mu$. As discussed in Section 1.10.2, the minimum of the two is therefore also exponential with parameter $\lambda + \mu$, and this minimum is the time the process stays in state $n$, for $n \geq 1$. We also know from the discussion in Section 1.10.2 that after spending an exponential amount of time with parameter $\lambda + \mu$, the process will move to state $n + 1$ with probability $\lambda/(\lambda + \mu)$ and to state $n - 1$ with probability $\mu/(\lambda + \mu)$.

## 6.1 Steady-State Queue Size Probabilities

As the M/M/1 queue-size process increases by only one, decreases by only one and stays an exponential amount of time at each state, it is equivalent to a birth-and-death process. Therefore, by Eqs. (233) and (234), the infinitesimal generator for the M/M/1 queue-size process is given by

$Q_{i,i+1} = \lambda$ for $i$=0, 1, 2, 3, ...
$Q_{i,i-1} = \mu$ for $i$= 1, 2, 3, 4, ...
$Q_{0,0} = -\lambda$
$Q_{i,i} = -\lambda - \mu$ for $i$=1, 2, 3, ... .

Substituting this infinitesimal generator in Eq. (235) we readily obtain the following global balance steady-state equations for the M/M/1 queue. $\pi_0\lambda = \pi_1\mu$
$\pi_1(\lambda + \mu) = \pi_2\mu + \pi_0\lambda$
and in general for $i \geq 1$:

$$\pi_i(\lambda + \mu) = \pi_{i+1}\mu + \pi_{i-1}\lambda \quad (278)$$

To explain (278) intuitively, Let $L$ be a very long time. During $L$, the total time that the process stays in state $i$ is equal to $\pi_i L$. For the case $i \geq 1$, since the arrival process is a Poisson process, the mean number of transitions out of state $i$ is equal to $(\lambda + \mu)\pi_i L$. This can be explained as follows. For the case $i \geq 1$, the mean number of events that occur during $L$ in state $i$ is $(\lambda + \mu)\pi_i L$ because as soon as the process enters state $i$ it stays there on average an amount of time equal to $1/(\mu + \lambda)$ and then it moves out of state $i$ to either state $i + 1$, or to state $i - 1$. Since during time $\pi_i L$ there are, on average, $(\lambda + \mu)\pi_i L$ interval times of size

$1/(\mu + \lambda)$, then $(\lambda + \mu)\pi_i L$ is also the mean number of events (arrivals and departures) that occur in state $i$ during $L$. In a similar way we can explain that for the case $i = 0$, the mean number of transition out of state $i = 0$ is equal to $(\lambda)\pi_i L$ for $i = 0$. It is the mean number of events that occur during $L$ (because there are no departures at state 0).

Recalling the notion of probability flux, introduced in Section 2.5.5, we notice that the global balance equations (278) equate for each state the total probability flux out of the state and the total probability flux into that state.

A solution of the global balance equations (278) together with the following normalizing equation that will guarantee that the sum of the steady-state probabilities must be equal to one:

$$\sum_{j=0}^{\infty} \pi_j = 1 \tag{279}$$

will give the steady-state probabilities of M/M/1.

However, the global balance equations (278) can be simplified recursively as follows. We first write the first equation: $\pi_0 \lambda = \pi_1 \mu$
Then we write the second equation $\pi_1(\lambda + \mu) = \pi_2 \mu + \pi_0 \lambda$
Then we observe that these two equation yield $\pi_1 \lambda = \pi_2 \mu$
Then recursively using all the equations (278), we obtain:

$$\pi_i \lambda = \pi_{i+1} \mu, \text{ for } i = 0,\ 1,\ 2,\ \ldots\ . \tag{280}$$

Notice that the steady-state equations (280) are the detailed balance equations of the continuous-time Markov chain that describes the stationary behaviour of the queue-size process of M/M/1. What we have notice here is that the global balance equations, in the case of the M/M/1 queue, are equivalent to the detailed balance equations. In this case, a solution of the detailed balance equations and (279) that sum up to unity will give the steady-state probability distribution of the queue-size. Recalling the discussion we had in Section 2.5.9, this implies that the M/M/1 queue is reversible, which in turn implies that the output process of M/M/1 is also a Poisson process. This is an important result that will be discussed later in Section 6.8.

Another way to realize that the queue size process of M/M/1 is reversible is to recall that this process is a birth-and-death process. And we already know from Section 2.5.9 that birth-and-death processes are reversible.

Let $\rho = \lambda/\mu$, by (280)we obtain,

$\pi_1 = \rho \pi_0$
$\pi_2 = \rho \pi_1 = \rho^2 \pi_0$
$\pi_3 = \rho \pi_2 = \rho^3 \pi_0$
and in general:

$$\pi_i = \rho^i \pi_0 \text{ for } i = 0,\ 1,\ 2,\ \ldots\ . \tag{281}$$

As M/M/1 is a special case of G/G/1, we can use Eq. (251) to obtain $\pi_0 = 1 - \rho$, so

$$\pi_i = \rho^i (1 - \rho) \text{ for } i = 0,\ 1,\ 2,\ \ldots\ . \tag{282}$$

Let $Q$ be a random number representing the queue-size in steady-state. Its mean is obtained by $E[Q] = \sum_{i=0}^{\infty} i\pi_i$. This leads to:

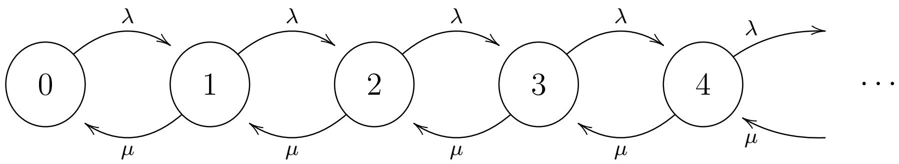
$$E[Q] = \frac{\rho}{1-\rho}. \quad (283)$$

**Homework 6.1**

Perform the algebraic operations that lead to (283). □

## 6.2 State Transition Diagram of M/M/1

In general, state transition diagrams are used to represent a system as a collection of states and activities associated with various relationships among the states. Such diagrams show how the system moves from one state to another, and the rates of movements between states. State transition diagrams have many applications related to design and analysis of real-time and object-oriented systems. Queueing systems that are modeled by continuous time Markov chains are often described by their state transition diagram that provides the complete information of their detailed balance equations. In particular, the state transition diagram of M/M/1 is[1]:




The states are the numbers in the circles: 0, 1, 2, 3, . . ., and the rates downwards and upwards are $\mu$ and $\lambda$, respectively. We observe that the rates of transitions between the states in the state transition diagram of M/M/1 are consistent with the rates in the detailed balance equations of M/M/1 (280).

## 6.3 Delay Statistics

By (282), and by the PASTA principle, an arriving customer will have to pass a geometric number of IID phases, each of which is exponentially distributed with parameter $\mu$, until it leaves the system. We have already shown that a geometrically distributed sum of an IID exponentially distributed random variables is exponentially distributed (see Eq. (155) in Section 1.15.2). Therefore the total delay of any arriving customer in an M/M/1 system must be exponentially distributed.

Therefore, to derive the density of the delay, all that is left to do is to obtain its mean which can be derived by (283) invoking Little's formula. Another way to obtain the mean delay is by noticing from (282) that the number of phases is geometrically distributed with mean $1/(1-\rho)$. Observe that this mean must equal $E[Q]+1$ which is the mean queue-size observed

[1]The author would like to thank Yin Chi Chan for his help in producing the various state transition diagrams in this book.

by an arriving customer plus one more phase which is the service time of the arriving customer. Thus, the mean number of phases is

$$E[Q]+1=\frac{\rho}{1-\rho}+1=\frac{1-\rho+\rho}{1-\rho}=\frac{1}{1-\rho}.$$

**Homework 6.2**

Prove that the number of phases is geometrically distributed with mean $1/(1-\rho)$.

**Guide**

Let $P_h$ be the number of phases. We know that in steady-state an arriving customer will find $Q$ customers in the system, where

$$P(Q=i)=\pi_i=\rho^i(1-\rho).$$

Since $P_h=Q+1$, we have

$$P(P_h=n)=P(Q+1=n)=P(Q=n-1)=\rho^{n-1}(1-\rho).$$

□

The mean delay equals the mean number of phases times the mean service time $1/\mu$. Thus,

$$E[D]=\frac{1}{(1-\rho)\mu}=\frac{1}{\mu-\lambda}. \tag{284}$$

**Homework 6.3**

Verify that (283) and (284) are consistent with Little's formula. □

Substituting $1/E[D]=\mu-\lambda$ as the parameter of exponential density, the density of the delay distribution is obtained to be given by

$$\delta_D(x)=\begin{cases}(\mu-\lambda)e^{(\lambda-\mu)x} & \text{if } x\geq 0\\ 0 & \text{otherwise.}\end{cases} \tag{285}$$

Having derived the distribution of the total delay (in the queue and in service), let us now derive the distribution of the queueing delay (excluding the service time). That is, we are interested in deriving $P(W_Q>t), \quad t\geq 0$. By the Law of Total Probability, we obtain:

$$\begin{aligned}P(W_Q>t) &= \rho P(W_Q>t|\text{server busy})+(1-\rho)P(W_Q>t|\text{server not busy})\\ &= \rho P(W_Q>t|\text{server busy}) \quad t\geq 0.\end{aligned} \tag{286}$$

To find $P(W_Q>t|\text{server busy})$, let us find $P(N_q=n|\text{server busy})$.

$$\begin{aligned} P(N_q = n|\text{server busy}) &= \frac{P(N_q = n \cap \text{server busy})}{P(\text{server busy})} \\ &= \frac{P(Q = n+1 \cap n+1 \geq 1)}{\rho} \\ &= \frac{\rho^{n+1}(1-\rho)}{\rho} \quad n = 0,1,2,3\ldots \\ &= \rho^n(1-\rho) \quad n = 0,1,2,3\ldots. \end{aligned}$$

Note that this is the same geometric distribution as that of $P(Q = n)$. Therefore, the random variable $\{W_Q > t|\text{server busy}\}$ is a geometric sum of exponential random variables and therefore has exponential distribution. As a result,

$$P(W_Q > t|\text{server busy}) = e^{-(\mu-\lambda)t}$$

and by (286) we obtain

$$P(W_Q > t) = \rho e^{-(\mu-\lambda)t} \quad t \geq 0.$$

## 6.4 Mean Delay of Delayed Customers

So far we were interested in delay statistics of all customers. Now suppose that we are interested in the mean delay of only those customers that found the server busy upon their arrivals and had to wait in the queue before they commence service. We assume that the arrival rate $\lambda$ and the service rate $\mu$ are given, then the mean number of customers in the queue $E[N_Q]$ is given by

$$E[N_Q] = E[Q] - E[N_s] = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}.$$

Denote:
$\hat{D}$ = The delay of a delayed customer including the service time
$\hat{W}_Q$ = The delay of a delayed customer in the queue excluding the service time.

To obtain $E[\hat{W}_Q]$, we use Little's formula where we consider the queue (without the server) as the system and the arrival rate of the delayed customers which is $\lambda\rho$. Thus

$$E[\hat{W}_Q] = \frac{E[N_Q]}{\lambda\rho} = \frac{1}{\mu-\lambda},$$

and

$$E[\hat{D}] = E[\hat{W}_Q] + \frac{1}{\mu} = \frac{1}{\mu-\lambda} + \frac{1}{\mu}.$$

Now, let us check the latter using the Law of Iterated Expectation as follows:

$$\begin{aligned} E[D] &= (1-\rho)[\text{Mean total delay of a non-delayed customer}] \\ &+ \quad \rho[\text{Mean total delay of a delayed customer}] \\ &= (1-\rho)\frac{1}{\mu} + \rho\left(\frac{1}{\mu-\lambda} + \frac{1}{\mu}\right) = \frac{1}{\mu-\lambda}. \end{aligned}$$

and we observe that consistency is achieved. Notice that this consistency check is an alternative way to obtain $E[\hat{D}]$.

**Homework 6.4**

Derive $E[\hat{W}_Q]$ using the Law of Iterated Expectation. □

**Homework 6.5**

Packets destined to a given destination arrive at a router according to a Poisson process with a rate of 2000 packets per millisecond. The router has a very large buffer and serves these packets by transmitting them through a single 10.5 Gb/s output link. The service policy is First Come First Served. The packet sizes are exponentially distributed with a mean of 625 bytes. Answer the following assuming the system is in steady state.

Compute the mean queue size and the mean total delay (including queueing time and service time). What do you observe from the answer?

**Solution**

$\lambda = 2000$ [packet/millisecond] $= 2 \times 10^6$ [packet/s]

$\mu = 10.5$ [Gb/s]/(625 × 8) [bits] $= 2.1 \times 10^6$ [packet/s].

Consider an M/M/1 queue for the case: $\lambda = 2 \times 10^6$ and $\mu = 2.1 \times 10^6$. We obtain

$$\rho = \frac{\lambda}{\mu} = 0.952 \text{ approx.}$$

$$E[Q] = \frac{\rho}{[1-\rho]} = 20 \text{ [packets]}$$

$$E[D] = \frac{E[Q]}{\lambda} = 10^{-5} \text{ [seconds].}$$

The delay is very small even if the utilization is high because of the high bit-rate (service rate).

Notice that the mean delay in M/M/1 is given by

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

so for a fixed $\rho$ and arbitrarily large $\mu$, the delay is arbitrarily small.

$$E[\hat{D}] = \frac{1}{\mu - \lambda} + \frac{1}{\mu} = 1.05 \times 10^{-5} \text{ [seconds] approx.}$$

□

## 6.5 Using Z-Transform

The Z-transform defined in Section 1.15, also known as Probability Generating Function, is a powerful tool to derive statistics of queueing behavior.

As an example, we will now demonstrate how the Z-transform is used to derive the mean queue-size of M/M/1.

Let us multiply the $n$th equation of (280) by $z^n$. Summing up both sides will give

$$\frac{\Psi(z) - \pi_0}{z} = \rho\Psi(z) \tag{287}$$

where $\Psi(z) = \sum_{i=0}^{\infty} \pi_i z^i$. Letting $z$ approach 1 (from below) gives

$$\pi_0 = 1 - \rho \tag{288}$$

which is consistent with what we know already. Substituting it back in (287) gives after simple algebraic manipulation:

$$\Psi(z) = \frac{1 - \rho}{1 - \rho z}. \tag{289}$$

Taking derivative and substituting $z = 1$, after some algebra we obtain

$$E[Q] = \Psi^{(1)}(1) = \frac{\rho}{1 - \rho} \tag{290}$$

which is again consistent with what we know about M/M/1 queue.

**Homework 6.6**

1. Derive equations (287) – (290).
2. Derive the variance of the M/M/1 queue-size using Z-transform. □

## 6.6 Multiplexing

In telecommunications, the concept of multiplexing refers to a variety of schemes or techniques that enable multiple traffic streams from possibly different sources to share a common transmission resource. In certain situations such sharing of a resource can lead to a significant improvement in efficiency. In this section, we use the M/M/1 queueing model to gain insight into efficiency gain of multiplexing.

An important and interesting observation we can make by considering the M/M/1 queueing performance results (282)–(285) is that while the queue-size statistics are dependent only on $\rho$ (the ratio of the arrival rate to the service rate), the delay statistics (mean and distribution) are a function of what we call the *spare capacity* (or *mean net input*) which is the difference between the service rate and the arrival rate. To be more specific, it is a linear function of the reciprocal of that difference.

Assume that our traffic model obeys the M/M/1 assumptions. Then if the arrival rate increases from $\lambda$ to $N\lambda$ and we increase the service rate from $\mu$ to $N\mu$ (maintaining the same $\rho$), the mean queue-size and its distribution will remain the same. However, in this scenario the mean delay does not remain the same. It reduces by $N$ times to $1/[N(\mu - \lambda)]$.

This is applicable to a situation where we have $N$ individual M/M/1 queues each of which with arrival rate $\lambda$ and service rate $\mu$. Then we superpose (multiplex) all the arrival processes together which results in a Poisson process of rate $N\lambda$. An interesting question is the following. If we replace all the individual servers (each of which has service rate $\mu$) with one fast server that serves the superposed Poisson stream of rate $N\lambda$, what service rate this fast server should operate at.

If our QoS measure of interest is the mean delay, or the probability that the delay exceeds a certain value, and if for a given arrival rate $\lambda$ there is a service rate $\mu$ such that our delay-related QoS measure is just met, then if the arrival rate increases from $\lambda$ to $N\lambda$, and we aim to find the service rate $\mu^*$ such that the delay-related QoS measure is just met, we will need to make sure that the spare capacity is maintained, that is

$$\mu - \lambda = \mu^* - N\lambda \tag{291}$$

or

$$\mu^* = \mu + (N - 1)\lambda \tag{292}$$

so by the latter and the stability condition of $\mu > \lambda$, we must have that $\mu^* < N\mu$. We can therefore define a measure for multiplexing gain to be given by

$$M_{mg} = \frac{N\mu - \mu^*}{N\mu} \tag{293}$$

so by (292), we obtain

$$M_{mg} = \frac{N - 1}{N}(1 - \rho). \tag{294}$$

Recalling the stability condition $\rho < 1$ and the fact that $\pi_0 = 1 - \rho$ is the proportion of time that the server is idle at an individual queue, Eq. (294) implies that $(N - 1)/N$ is the proportion of this idle time gained by multiplexing. For example, consider the case $N = 2$, that is, we consider multiplexing of two M/M/1 queues each with parameters $\lambda$ and $\mu$. In this case, half of the server idle time (or efficiency wastage) in an individual queue can be gained back by multiplexing the two streams to be served by a server that serves at the rate of $\mu^* = \mu + (N - 1)\lambda = \mu + \lambda$. The following four messages follow from Eq. (294).

1. The multiplexing gain is positive for all $N > 1$.
2. The multiplexing gain increases with $N$.
3. The multiplexing gain is bounded above by $1 - \rho$.
4. In the limiting condition as $N \to \infty$, the multiplexing gain approaches its bound $1 - \rho$.

The $1 - \rho$ bound means also that if $\rho$ is very close to 1, then the multiplexing gain diminishes because in this case the individual M/M/1 queues are already very efficient in terms of server utilization so there is little room for improvement. On the other hand, if we have a case where the QoS requirements are strict (requiring very low mean queueing delay) such that the utilization $\rho$ is low, the potential for multiplexing gain is high.

Let us now apply our general discussion on multiplexing to obtain insight into performance comparison between two commonly used multiple access techniques used in telecommunications.

One such technique is called *Time Division Multiple Access* (TDMA) whereby each user is assigned one or more channels (in a form of time-slots) to access the network. Another approach, which we call *full multiplexing* (FMUX), is to let all users to separately send the data that they wish to transmit to a switch which then forwards the data to the destination. That is, all the data is stored in one buffer (in the switch) which is served by the entire available link capacity.

To compare between the two approaches, let us consider $N$ users each transmitting packets at an average rate of $R_u$ [bits/second]. The average packet size denoted $S_u$ [bits] is assumed equal for the different users. Let $\hat{\lambda}$ [packets/second] be the packet rate generated by each of the users. Thus, $\hat{\lambda} = R_u/S_u$. Under TDMA, each of the users obtains a service rate of $B_u$ [bits/sec]. Packet sizes are assumed to be exponentially distributed with mean $S_u$ [bits], so the service rate in packets/second denoted $\hat{\mu}$ is given by $\hat{\mu} = B_u/S_u$. The packet service time is therefore exponentially distributed with parameter $\hat{\mu}$. Letting $\hat{\rho} = \hat{\lambda}/\hat{\mu}$, the mean queue size under TDMA, is given by

$$E[Q_{TDMA}] = \frac{\hat{\rho}}{1-\hat{\rho}}, \tag{295}$$

and the mean delay is

$$E[D_{TDMA}] = \frac{1}{\hat{\mu}-\hat{\lambda}}. \tag{296}$$

In the FMUX case the total arrival rate is $N\hat{\lambda}$ and the service rate is $N\hat{\mu}$, so in this case, the ratio between the arrival and service rate remains the same, so the mean queue size that only depends on this ratio remains the same

$$E[Q_{FMUX}] = \frac{\hat{\rho}}{1-\hat{\rho}} = E[Q_{TDMA}]. \tag{297}$$

However, we can observe an $N$-fold reduction in the mean delay:

$$E[D_{FMUX}] = \frac{1}{N\hat{\mu}-N\hat{\lambda}} = \frac{E[D_{TDMA}]}{N}. \tag{298}$$

Consider a telecommunication provider that wishes to meet packet delay requirement of its $N$ customers, assuming that the delay that the customers experienced under TDMA was satisfactory, and assuming that the M/M/1 assumptions hold, such provider does not need a total capacity of $N\hat{\mu}$ for the FMUX alternative. It is sufficient to allocate $\hat{\mu} + (N-1)\hat{\lambda}$.

**Homework 6.7**

Consider a telecommunication provider that aims to serve a network of 100 users each transmits data at overall average rate of 1 Mb/s. The mean packet size is 1 kbit. Assume that packets lengths are exponentially distributed and that the process of packets generated by each user follows a Poisson process. Further assume that the mean packet delay requirement is 50 millisecond. How much total capacity (bit-rate) is required to serve the 100 users under TDMA and under FMUX.

**Guide**

The arrival rate of each user is 1 Mb/s / 1 kbit = 1000 packets/s. For TDMA, use Eq. (296) and substitute $E[D_{TDMA}] = 0.05$ and $\hat{\lambda} = 1000$, to compute $\hat{\mu}$. This gives $\hat{\mu} = 1020$ [packets/s] or bit-rate of 1.02 Mb/s per each user. For 100 users the required rate is 102,000 packets/s or bit-rate of 102 Mb/s. For FMUX the required rate is $\hat{\mu}+(N-1)\hat{\lambda}$ which is 100,020 packets/s or 100.02 Mb/s (calculate and verify it). The savings in using FMUX versus TDMA is therefore 1.98 Mb/s. □

## 6.7 Dimensioning Based on Delay Distribution

In the previous section we have considered dimensioning based on average delay. That is, the aim was to meet or to maintain QoS measured by average delay. It may be, however, more practical to aim for a percentile of the delay distribution; e.g., to require that no more than 1% of the packets will experience over 100 millisecond delay.

In the context of the M/M/1 model, we define two dimensioning problems. The first problem is: for given $\lambda$, $t \geq 0$ and $\alpha$, find Minimal $\mu^*$ such that

$$P(D > t) = e^{-(\mu^*-\lambda)t} \leq \alpha.$$

The solution is:

$$\mu^* = \lambda - \frac{\ln(\alpha)}{t}.$$

The second problem is: for given $\mu$, $t \geq 0$ and $\alpha$, find Maximal $\lambda^*$ such that

$$P(D > t) = e^{-(\mu-\lambda^*)t} \leq \alpha.$$

To solve this problem we solve for $\lambda^*$ the equation

$$P(D > t) = e^{-(\mu-\lambda^*)t} = \alpha.$$

However, for a certain range, the solution is not feasible because the delay includes the service time and can never be less than the service time. That is, for certain parameter values, even if the arrival rate is very low, the delay requirements cannot be met, simply because the service time requirements exceeds the total delay requirements. To find the feasible range set $\lambda^* = 0$, and obtain

$$\mu > \frac{-\ln(\alpha)}{t}.$$

This requires that the probability of the service time requirement exceeds $t$ is less than $\alpha$.

In other words, if this condition does not hold there is no feasible solution to the optimal dimensioning problem.

If a solution is feasible, the $\lambda^*$ is obtained by

$$\lambda^* = \frac{\ln(\alpha)}{t} + \mu.$$

## 6.8 The Departure Process

We have already mentioned in Section 6.1 the fact that the output process of an M/M/1 is Poisson. This is one of the results of the so-called Burke's theorem [17]. In steady-state, the departure process of a stable M/M/1, where $\rho < 1$, is a Poisson process with parameter $\lambda$ and is independent of the number in the queue after the departures occur. If we already know that the output process is Poisson, given that the arrival rate is $\lambda$ and given that there are no losses, all the traffic that enters must depart. Therefore the rate of the output process must also be equal to $\lambda$.

We have shown in Section 6.1 the reversibility of M/M/1 queue-size process showing that the detailed balance equations and the normalizing equation yield the steady-state distribution of the queue-size process.

By reversibility, in steady-state, the arrival process of the reversed process must also follow a Poisson process with parameter $\lambda$ and this process is the departure process of the forward process. Therefore the departures follow a Poisson process and the inter-departure times are independent of the number in the queue after the departures occur in the same way that inter-arrival times are independent of a queue size before the arrivals.

Now that we know that in steady-state the departure process of a sable M/M/1 queue is Poisson with parameter $\lambda$, we also know that, in steady-state, the inter-departure times are also exponentially distributed with parameter $\lambda$. We will now show this fact without using the fact that the departure process is Poisson directly. Instead, we will use it indirectly to induce PASTA for the reversed arrival process to obtain that, following a departure, in steady-state, the queue is empty with probability $1 - \rho$ and non-empty with probability $\rho$. If the queue is non-empty, the time until the next departure is exponentially distributed with parameter $\mu$ – this is the service-time of the next customer. If the queue is empty, we have to wait until the next customer arrival which is exponentially distributed with parameter $\lambda$ and then we will have to wait until the next departure which will take additional time which is exponentially distributed. All together, if the queue is empty, the time until the next departure is a sum of two exponential random variables, one with parameter $\lambda$ and the other with parameter $\mu$. Let $U_1$ and $U_2$ be two independent exponential random variables with parameters $\lambda$ and $\mu$, respectively. Define $U = U_1 + U_2$, notice that $U$ is the convolution of $U_1$ and $U_2$, and note that $U$ has hypoexponential distribution. Having the density $f_U(u)$, the density $f_D(t)$ of a random variable $D$ representing the inter-departure time will be given by

$$f_D(t) = \rho\mu e^{-\mu t} + (1 - \rho) f_U(t). \tag{299}$$

Knowing that $f_U(u)$ is a convolution of two exponentials, we obtain

$$\begin{aligned} f_U(t) &= \int_{u=0}^{t} \lambda e^{-\lambda u} \mu e^{-\mu(t-u)} du \\ &= \frac{\lambda\mu}{\mu - \lambda} \left( e^{-\lambda t} - e^{-\mu t} \right). \end{aligned}$$

Then by the latter and (299), we obtain

$$f_D(t) = \rho\mu e^{-\mu t} + (1-\rho)\frac{\lambda\mu}{\mu-\lambda}\left(e^{-\lambda t} - e^{-\mu t}\right) \tag{300}$$

which after some algebra gives

$$f_D(t) = \lambda e^{-\lambda t}. \tag{301}$$

This result is consistent with Burke's theorem.

**Homework 6.8**

Complete all the algebraic details in the derivation of equations (299) – (301). □

Another way to show consistency with Burke's theorem is the following. Consider a stable ($\rho < 1$) M/M/1 queue. Let $d_\epsilon$ be the unconditional number of departures, in steady state, that leave the M/M/1 queue during a small interval of time of size $\epsilon$, and let $d_\epsilon(i)$ be the number of departures that leave the M/M/1 queue during a small interval of time of size $\epsilon$ if there are $i$ packets in our M/M/1 queue at the beginning of the interval. Then, $P(d_\epsilon(i) > 0) = o(\epsilon)$ if $i = 0$, and $P(d_\epsilon(i) = 1) = \epsilon\mu + o(\epsilon)$ if $i > 0$. Therefore, in steady-state,

$$P(d_\epsilon = 1) = (1-\rho)0 + (\rho)\mu\epsilon + o(\epsilon) = \epsilon\lambda + o(\epsilon),$$

which is a property consistent with the assertion of Poisson output process with parameter $\lambda$ in steady-state.

**Homework 6.9**

So far we have discussed the behaviour of the M/M/1 departure process in steady-state. You are now asked to demonstrate that the M/M/1 departure process may not be Poisson with parameter $\lambda$ if we do not assume steady-state condition. Consider an M/M/1 system with arrival rate $\lambda$ and service rate $\mu$, assume that $\rho = \lambda/\mu < 1$ and that there are no customers in the system at time 0. Derive the distribution of the number of customers that leave the system during the time interval $(0, t)$. Argue that this distribution is, in most cases, not Poisson with parameter $\lambda t$ and find a special case when it is.

**Guide**

Let $D(t)$ be a random variable representing the number of customers that leave the system during the time interval $(0, t)$. Let $X_p(\lambda t)$ be a Poisson random variable with parameter $\lambda t$ and consider two cases: (a) the system is empty at time $t$, and (b) the system is not empty at time $t$. In case (a), $D(t) = X_p(\lambda t)$ (why?) and in case (b) $D(t) = X_p(\lambda t) - Q(t)$ (why?) and use the notation used in Section 2.5 $P_{00}(t)$ to denote the probability that in time $t$ the system is empty, so the probability that the system is not empty at time $t$ is $1 - P_{00}(t)$. Derive $P_{00}(t)$ using Eqs. (230) and (231). Then notice that

$$D(t) = P_{00}(t)X_p(\lambda t) + [1 - P_{00}(t)][X_p(\lambda t) - Q(t)]. \quad \square$$

Consider the limit

$$D_k(t) = \lim_{\Delta t \to 0} P[Q(t) = k \mid \text{a departure occurs within}(t - \Delta t, t)].$$

Considering the fact that the reversed process is Poisson and independence between departures before time $t$ and Q(t), we obtain that

$$D_k(t) = P[Q(t) = k]. \tag{302}$$

Then, by taking the limit of both sides of (302), we show that the queue size seen by a leaving customer is statistically identical to the queue size seen by an independent observer. □

**Homework 6.10**

Write a simulation of the M/M/1 queue by measuring queue size values in two ways: (1) just before arrivals and (2) just after departures. Verify that the results obtained for the mean queue size in steady-state are consistent. Use confidence intervals. Verify that the results are also consistent with analytical results. Repeat your simulations and computation for a wide range of parameters values (different $\rho$ values). Plot all the results in a graph including the confidence intervals (bars). □

## 6.9 Mean Busy Period and First Passage Time

The *busy period* of a single-server queueing system is defined as the time between the point in time the server starts being busy and the point in time the server stops being busy. In other words, it is the time elapsed from the moment a customer arrives at an empty system until the first time the system is empty again. Recalling the first passage time concept defined in Section 2.5, and that the M/M/1 system is in fact a continuous-time Markov-chain, the busy period is also the first passage time from state 1 to state 0. The end of a busy period is the beginning of the so called *idle period* - a period during which the system is empty. We know the mean of the idle period in an M/M/1 queue. It is equal to $1/\lambda$ because it is the mean time until a new customer arrives which is exponentially distributed with parameter $\lambda$. A more interesting question is what is the mean busy period. Let $T_B$ and $T_I$ be the busy and the idle periods, respectively. Noticing that $E[T_B]/(E[T_B] + E[T_I])$ is the proportion of time that the server is busy, thus it is equal to $\rho$. Considering also that $E[T_I] = 1/\lambda$, we obtain

$$\frac{E[T_B]}{E[T_B] + \frac{1}{\lambda}} = \rho. \tag{303}$$

Therefore,

$$E[T_B] = \frac{1}{\mu - \lambda}. \tag{304}$$

Interestingly, for the M/M/1 queue the mean busy period is equal to the mean delay of a single customer! This may seem counter intuitive. However, we can realize that there are many busy periods each of which is made of a single customer service time. It is likely that for the majority of these busy periods (service times), their length is shorter than the mean delay of a customer.

Furthermore, the fact that for the M/M/1 queue the mean busy period is equal to the mean delay of a single customer can be proven by considering an M/M/1 queue with a service policy of Last In First Out (LIFO). So far we have considered only queues that their service policy is First In First Out (FIFO). Let us consider an M/M/1 with LIFO with preemptive priority. In such a queue the arrival and service rates $\lambda$ and $\mu$, respectively, are the same as those of the FIFO M/M/1, but in the LIFO queue, the customer just arrived has priority over all other customers that arrived before it and in fact interrupts the customers currently in service. (More information on LIFO queues is available in Section 17.1.)

The two queues we consider, the FIFO and the LIFO, are both birth-and-death processes with the same parameters so their respective queue size processes are statistically the same. Then by Little's formula their respective mean delays are also the same. Also the delay of a customer in an M/M/1 LIFO queue we consider is equal to the busy period in M/M/1 FIFO queue (why?) so the mean delay must be equal to the busy period in M/M/1 with FIFO service policy.

**Homework 6.11**

Derive an expression for the mean first passage time for M/M/1 from state $n$ to state 0 and from state 0 to state $n$, for $n \geq 3$. □

**Homework 6.12**

For a wide range of parameter values, simulate an M/M/1 system with FIFO service policy and an M/M/1 system with LIFO service policy with preemptive priority and compare their respective results for the mean delay, the variance of the delay, the mean queue size and the mean busy period. □

## 6.10 A Markov-chain Simulation of M/M/1

A simulation of am M/M/1 queue can be made as a special case of G/G/1 as described before, or it can be simplified by taking advantage of the M/M/1 Markov-chain structure if we are not interested in performance measures that are associated with times (such as delay distribution). If our aim is to evaluate queue size statistics or blocking probability, we can avoid tracking the time. All we need to do is to collect the relevant information about the process at PASTA time-points without even knowing what is the running time at these points. Generally speaking, using the random walk simulation approach, also called the *Random Walk simulation* approach, we simulate the evolution of the states of the process based on the transition probability matrix and collect information on the values of interest at selective PASTA points without being concerned about the time. We will now explain how these ideas may be applied to few relevant examples.

If we wish to evaluate the mean queue size of an M/M/1 queue, we can write the following simulation.

Variables and input parameters: $Q$ = queue size; $\hat{E}(Q)$ = estimation for the mean queue size; $N$ = number of $Q$-measurements taken so far which is also equal to the number of arrivals so far; $MAXN$ = maximal number of $Q$-measurements taken; $\mu$ = service rate; $\lambda$ = arrival rate.

Define function: $I(Q) = 1$ if $Q > 0$; $I(Q) = 0$ if $Q = 0$.
Define function: $R(01)$ = a uniform $U(0,1)$ random deviate. A new value for R(01) is generated every time it is called.

Initialization: $Q = 0$; $\hat{E}[Q] = 0$; $N = 0$.

1. If $R(01) \leq \lambda/(\lambda + I(Q)\mu)$, then $N = N + 1$, $\hat{E}(Q) = [(N-1)\hat{E}(Q) + Q]/N$, and $Q = Q + 1$;

else, $Q = Q - 1$.

2. If $N < MAXN$ go to 1; else, print $\hat{E}(Q)$.

This signifies the simplicity of the simulation. It has only two If statements: one to check if the next event is an arrival or a departure according to Eq. (63), and the second is merely a stopping criterion.

**Comments:**

1. The operation $Q = Q + 1$ is performed after the $Q$ measurement is taken. This is done because we are interested in $Q$ values seen by arrivals just before they arrive. If we include the arrivals after they arrive we violate the PASTA principle. Notice that if we do that, we never observe a $Q = 0$ value which of course will not lead to an accurate estimation of $E[Q]$.
2. If the condition $R(01) \leq \lambda/(\lambda + I(Q)\mu)$ holds we have an arrival. Otherwise, we have a departure. This condition is true with probability $\lambda/(\lambda + I(Q)\mu)$. If $Q = 0$ then $I(Q) = 0$ in which case the next event is an arrival with probability 1. This is clearly intuitive. If the system is empty no departure can occur, so the next event must be an arrival. If $Q > 0$, the next event is an arrival with probability $\lambda/(\lambda + \mu)$ and a departure with probability $\mu/(\lambda + \mu)$. We have here a competition between two exponential random variables: one (arrival) with parameter $\lambda$ and the other (departure) with parameter $\mu$. According to the discussion in Section 1.10.2 and as mentioned in the introduction to this section, the probability that the arrival "wins" is $\lambda/(\lambda + \mu)$, and the probability that the departure "wins" is $\mu/(\lambda + \mu)$.
3. In a case of a departure, all we do is decrementing the queue size; namely, $Q = Q - 1$. We do not record the queue size at these points because according to PASTA arrivals see time-averages. (Notice that due to reversibility, if we measure the queue size immediately **after** departure points we will also see time-averages.)

**Homework 6.13**

Simulate an M/M/1 queue using a Markov-chain simulation to evaluate the mean queue-size for the cases of Section 4.2. Compare the results with the results obtain analytically and with those obtained using the G/G/1 simulation principles. In your comparison consider accuracy (closeness to the analytical results) the length of the confidence intervals and running times. □

# 7 M/M/$\infty$

The next queueing system we consider is the M/M/$\infty$ queueing system where the number of servers is infinite. Because the number of servers is infinite, the buffer capacity is unlimited and arrivals are never blocked. We assume that the arrival process is Poisson with parameter $\lambda$ and each server renders service which is exponentially distributed with parameters $\mu$. As in the case of M/M/1, we assume that the service times are independent and are independent of the arrival process.

## 7.1 Offered and Carried Traffic

The concept of *offered traffic* is one of the most fundamentals in the field of teletraffic. It is often used in practice in telecommunications network design and in negotiations between telecommunications carriers. The offered traffic is defined as the mean number of arrivals per mean service time. Namely, it is equal to the ratio $\lambda/\mu$. It is common to use the notation $A$ for the offered traffic, so we denote $A = \lambda/\mu$ in the context of the M/M/$\infty$ queue. Notice that we use the notation $A$ here for the ratio $\lambda/\mu$ while we used the notation $\rho$ for this ratio in the M/M/1 case as $\rho$ is defined by the ratio $A/k$ and in an SSQ $k = 1$ so $\rho = A$. Clearly, both $A$ and $\rho$ represent the offered traffic by definition, but $\rho$ is used for SSQs and $A$ for multi server queues. In addition, $\rho$ and $A$ represent the mean number of busy servers in the M/M/1 and M/M/$\infty$ cases, respectively. We have already shown that this is true for a G/G/1 queue (and therefore also for M/M/1). We will now show that it is true for an M/M/$\infty$ queue. According to Little's formula, the mean number of customers in the system is equal to the arrival rate ($\lambda$) times the mean time a customer spends in the system which is equal to $1/\mu$ in the case of M/M/$\infty$. Because there are no losses in M/M/$\infty$, all the arriving traffic enters the service-system, so we obtain

$$E[Q] = \lambda(1/\mu) = A. \tag{305}$$

In M/M/1, we must have that $\rho$ cannot exceed unity for stability. In M/M/1 $\rho$ also represents the server utilization which cannot exceeds unity. However, in M/M/$\infty$, $A$ can take any non-negative value and we often have $A > 1$. M/M/$\infty$ is stable for any $A \geq 0$. Notice that in M/M/$\infty$ the service rate increases with the number of busy servers and when we reach a situation where the number of busy servers $j$ is higher that $A$ (namely $j > A = \lambda/\mu$), we will have that the system service rate is higher than the arrival rate (namely $j\mu > \lambda$).

As discussed in Chapter 3, offered traffic is measured in *erlangs*. One erlang represents traffic load of one arrival, on average, per mean service time. This means that traffic load of one erlang, if admitted to the system, will require a resource of one server on average. For example this could be provided by service from one server continuously busy, or from two servers each of which is busy 50% of the time,

Another important teletraffic concept is the *carried traffic*. It is defined as the mean number of customers, calls or packets leaving the system after completing service during a time period equal to the mean service time. Carried traffic is also measured in erlangs and it is equal to the mean number of busy servers which is equal to the mean queue size. It is intuitively clear that if, on average, there are $n$ busy servers each completing service for one customer per one mean service time, we will have, on average, $n$ service completions per service time. In the case of M/M/$\infty$, the carried traffic is equal to $A$ which is also the *offered traffic*, namely the mean number of arrivals during a mean service time. The equality:

$$\text{offered traffic} = \text{carried traffic}$$

is due to the fact that all traffic is admitted as there are no losses in M/M/$\infty$.

In practice, the number of servers (channels or circuits) is limited, and the offered traffic is higher than the carried traffic because some of the calls are blocked due to call congestion when all circuits are busy. A queueing model which describes this more realistic case is the M/M/$k$/$k$ queueing model discussed in the next chapter.

Another term to describe traffic which is often used is the *traffic intensity* [45]. Then *offered traffic intensity* and *carried traffic intensity* are synonymous to offered traffic and carried traffic, respectively. In cases of infinite buffer/capacity systems, such as M/M/1 and M/M/$\infty$, the term traffic intensity is used to describe both offered traffic and carried traffic. Accordingly, $\rho$ is called traffic intensity in the M/M/1 context and $A$ is the traffic intensity in an M/M/$\infty$ system. Others use the term traffic intensity in multiservice system for the offered load per server [89]. To avoid confusion, we will only use the term traffic intensity in the context of a single server queue with infinite buffer, in which case, traffic intensity is always equal to $\rho = \lambda/\mu$.

## 7.2 Steady-State Equations

As for M/M/1, the queue-size process of an M/M/$\infty$ system can also be viewed as a continuous-time Markov-chain with the state being the queue-size (the number of customers in the system). As for M/M/1, since in M/M/$\infty$ queue-size process is stationary and its transitions can only occur upwards by one or downwards by one, the queue-size process is a birth-and-death process and therefore it is reversible. As As in M/M/1, the arrival rate is independent of changes in the queue-size. However, unlike M/M/1, in M/M/$\infty$, the service rate does change with the queue-size. When there are $n$ customers in the system, and at the same time, $n$ servers are busy, the service rate is $n\mu$, and the time until the next event is exponentially distributed with parameter $\lambda + n\mu$, because it is a competition between $n + 1$ exponential random variables: $n$ with parameter $\mu$ and one with parameter $\lambda$.

Considering a birth-and-death process that represents the queue evolution of an M/M/$\infty$ queueing system, and its reversibility property, the steady-state probabilities $\pi_i$ (for $i = 0, 1, 2, \ldots$) of having $i$ customers in the system satisfy the following detailed balance (steady-state) equations:
$\pi_0\lambda = \pi_1\mu$
$\pi_1\lambda = \pi_2 2\mu$
...
and in general:

$$\pi_n\lambda = \pi_{n+1}(n + 1)\mu, \text{ for } n = 0, \ 1, \ 2, \ \ldots \ . \tag{306}$$

The sum of the steady-state probabilities must be equal to one, so we again have the additional normalizing equation

$$\sum_{j=0}^{\infty} \pi_j = 1. \tag{307}$$

We note that the infinitesimal generator of M/M/$\infty$ is given by

$Q_{i,i+1} = \lambda$ for $i$=0, 1, 2, 3, ...
$Q_{i,i-1} = i\mu$ for $i$= 1, 2, 3, 4, ...
$Q_{0,0} = -\lambda$
$Q_{i,i} = -\lambda - i\mu$ for $i$=1, 2, 3, ... .

## 7.3 Solving the Steady-State Equations

Using the $A$ notation we obtain

$\pi_1 = A\pi_0$
$\pi_2 = A\pi_1/2 = A^2\pi_0/2$
$\pi_3 = A\pi_2/3 = A^3\pi_0/(3!)$

and in general:

$$\pi_n = \frac{A^n \pi_0}{n!} \text{ for } n = 0,\ 1,\ 2,\ \ldots\ . \tag{308}$$

To obtain $\pi_0$, we sum up both sides of Eq. (308), and because the sum of the $\pi_n$s equals one, we obtain

$$1 = \sum_{n=0}^{\infty} \frac{A^n \pi_0}{n!}. \tag{309}$$

By the definition of Poisson random variable, see Eq. (29), we obtain

$$1 = \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!}. \tag{310}$$

Thus,

$$e^{\lambda} = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$$

which is also the well-known Maclaurin series expansion of $e^{\lambda}$. Therefore, Eq. (309) reduces to

$$1 = \pi_0 e^A, \tag{311}$$

or

$$\pi_0 = e^{-A}. \tag{312}$$

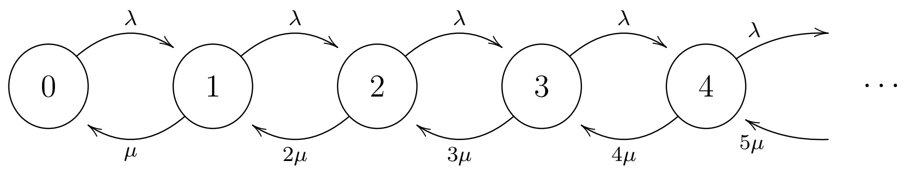
Substituting the latter in Eq. (308), we obtain

$$\pi_n = \frac{e^{-A} A^n}{n!} \text{ for } n = 0,\ 1,\ 2,\ \ldots\ . \tag{313}$$

By Eq. (313) we observe that the distribution of the number of busy channels (simultaneous calls or customers) in an M/M/$\infty$ system is Poisson with parameter $A$.

## 7.4 State Transition Diagram of M/M/∞

The state transition diagram of M/M/∞ is similar to that of M/M/1 except that the rate downwards from state $n$ ($n = 1, 2, 3, \ldots$) is $n\mu$ rather than $\mu$ reflecting the fact at state $n$ there are $n$ servers serving the $n$ customers. The state transition diagram of M/M/∞ is:




We observe that the rates of transitions between the states in the state transition diagram of M/M/∞ are consistent with the rates in the detailed balance equations of M/M/∞ (306).

## 7.5 Insensitivity

The above results for $\pi_i$, $i = 0, 1, 2\ \ldots$ and for the mean number of busy servers are insensitive to the shape of the service time (holding time) distribution [10, 66, 70, 82]; all we need to know is the mean service time and the results are insensitive to higher moments. In other words, the above results apply to an M/G/∞ system. This is important because it makes the model far more robust which allows us to use its analytical results for many applications where the service time is not exponential. This insensitivity property is valid also for the M/G/$k$/$k$ system [37, 81, 82, 87].

To explain the insensitivity property of M/G/∞ with respect to the mean occupancy, consider an arbitrarily long period of time $L$ and also consider the queue size process function, that represents the number of busy servers at any point in time between 0 and $L$. The average number of busy servers is obtained by the area under the queue size process function divided by $L$. This area is closely approximated by the number of arrivals during $L$ which is $\lambda L$ times the mean holding (service) time of each arrival ($1/\mu$). Therefore the mean number of busy servers, which is also equal to the mean number of customers in the system (queue size), is equal to $A = \lambda/\mu$ (notice that the $L$ is canceled out here). Since all the traffic load enters the system ($A$) is also the carried traffic load.

The words "closely approximated" are used here because there are some customers that arrive before $L$ and receive service after $L$ and there are other customers that arrive before time 0 and are still in the system after time 0. However because we can choose $L$ to be arbitrarily long, their effect is negligible.

Since in the above discussion, we do not use moments higher than the mean of the holding time, this mean number of busy servers (or mean queue size) is insensitive to the shape of the holding-time distribution and it is only sensitive to its mean.

Moreover, the distribution of the number of busy servers in M/G/∞ is also insensitive to the holding time distribution. This can be explained as follows. We know that the arrivals follow a Poisson process. Poisson process normally occurs in nature by having a very large number of independent sources each of which generates occasional events (arrivals) [66] - for example, a large population of customers making phone calls. These customers are independent of each other. In M/G/$\infty$, each one of the arrivals generated by these customers is able to find a server and its arrival time, service time and departure time is independent of all other arrivals (calls). Therefore, the event that a customer occupies a server at an arbitrary point in time in steady-state is also independent of the event that any other customer occupies a server at that point in time. Therefore, the server occupancy events are also due to many sources generating occasional events. This explains the Poisson distribution of the server occupancy. From the above discussion, we know that the mean number of servers is equal to $A$, so we always have, in M/G/$\infty$, in steady-state, a Poisson distributed number of servers with parameter $A$ which is independent of the shape of the service-time distribution.

## 7.6 Applications

### 7.6.1 A multi-access model

An interesting application of the M/M/$\infty$ system is the following multi-access problem (see Problem 3.8 in [13]). Consider a stream of packets that their arrival times follow a Poisson process with parameter $\lambda$. If the inter-arrival times of any pair of packets (not necessarily a consecutive pair) is less than the transmission time of the packet that arrived earlier out of the two, these two packets are said to collide. Assume that packets have independent exponentially distributed transmission times with parameter $\mu$. What is the probability of no collision?

Notice that a packet can collide with any one or more of the packets that arrived before it. In other words, it is possible that it may not collide with its immediate predecessor, but it may collide with a packet that arrived earlier. However, if it does not collide with its immediate successor, it will not collide with any of the packets that arrive after the immediate successor.

Therefore, the probability that an arriving packet will not collide on arrival can be obtained to be the probability of an M/M/$\infty$ system to be empty, that is, $e^{-A}$. While the probability that its immediate successor will not arrive during its transmission time is $\mu/(\lambda+\mu)$. The product of the two, namely $e^{-A}\mu/(\lambda+\mu)$, is the probability of no collision.

### 7.6.2 Birth rate evaluation

Another application of the M/M/$\infty$ system (or M/G/$\infty$ system) is to the following problem. Consider a city with population 3,000,000, and assume that (1) there is no immigration in and out of the city, (2) the birth rate $\lambda$ in constant (time independent), and (3) life-time expectancy $\mu^{-1}$ in the city is constant. It is also given that average life-time of people in this city is 78 years. How to compute the birth rate?

Using the M/M/$\infty$ model (or actually the M/G/$\infty$ as human lifetime is not exponentially distributed) with $E[Q] = 3,000,000$ and $\mu^{-1} = 78$, realizing that $E[Q] = A = \lambda/\mu$, we obtain, $\lambda = \mu E[Q] = 3,000,000/78 = 38461$ new births per year or 105 new births per day.

**Homework 7.1**

Consider an M/M/$\infty$ queue, with $\lambda = 120$ [call/s], and $\mu = 3$ [call/s]. Find the steady state probability that there are 120 calls in the system. This should be done by a computer. Use ideas presented in Section 1.8.5 to compute the required probabilities.

# 8 M/M/$k$/$k$ and Extensions

We begin this chapter with the M/M/$k$/$k$ queueing system where the number of servers is $k$ assuming that the arrival process is Poisson with parameter $\lambda$ and that each server renders service which is exponentially distributed with parameters $\mu$. Later in the chapter we extend the model to cases where the arrival process is non-Poisson.

In the M/M/$k$/$k$ model, as in the other M/M/... cases, we assume that the service times are mutually independent and are independent of the arrival process. We will now discuss Erlang's derivation of the loss probability of an M/M/$k$/$k$ system that leads to the well known Erlang's Loss Formula, also known as Erlang B Formula.

## 8.1 M/M/$k$/$k$: Offered, Carried and Overflow Traffic

The offered traffic under M/M/$k$/$k$ is the same as under M/M/$\infty$ it is equal to

$$A = \lambda/\mu.$$

However, because some of the traffic is blocked the offered traffic is not equal to the carried traffic. To obtain the carried traffic given a certain blocking probability $P_b$, we recall that the carried traffic is equal to the mean number of busy servers. To derive the latter we again invoke Little's formula. We notice that the arrival rate into the service system is equal to $(1 - P_b)\lambda$ and that the mean time each customer (or call) spends in the system is $1/\mu$. The mean queue size (which is also the mean number of busy servers in the case of the M/M/$k$/$k$ queue) is obtained to be given by

$$E[Q] = \frac{(1 - P_b)\lambda}{\mu} = (1 - P_b)A. \tag{314}$$

Therefore the carried traffic is equal to $(1 - P_b)A$. Notice that since $P_b > 0$ in M/M/$k$/$k$, the carried traffic here is lower than the corresponding carried traffic for M/M/$\infty$ which is equal to $A$.

The *overflow traffic* (in the context of M/M/$k$/$k$ it is also called: *lost traffic*) is defined as the difference between the two. Namely,

$$\text{overflow traffic} = \text{offered traffic} - \text{carried traffic}.$$

Therefore, for M/M/$k$/$k$, the overflow traffic is

$$A - (1 - P_b)A = P_b A.$$

## 8.2 The Steady-State Equations and Their Solution

The steady-state equations for M/M/$k$/$k$ are the same as the first $k$ steady-state equations for M/M/$\infty$.

As for M/M/$\infty$, the queue-size process of an M/M/$k$/$k$ is a birth-and-death process where queue-size transitions can only occur upwards by one or downwards by one. Therefore, the

M/M/$k$/$k$ queue-size process is also reversible which means that solving its detailed balance equations and the normalizing equation yields the steady-state probabilities of the queue size.

The difference between the two systems is that the queue size of M/M/$k$/$k$ can never exceed $k$ while for M/M/$\infty$ it is unlimited. In other words, the M/M/$k$/$k$ system is a truncated version of the M/M/$\infty$ with losses occur at state $k$, but the ratio $\pi_j/\pi_0$ for any $0 \leq j \leq k$ is the same in both M/M/$k$/$k$ and M/M/$\infty$ systems.

Another difference between the two is associated with the physical interpretation of reversibility. Although both systems are reversible, while in M/M/$\infty$ the physical interpretation is that in the reversed process the input point process is the point process of call completion times in reverse, in M/M/$k$/$k$ the reversed process is the superposition of call completion times in reverse and call blocking times in reverse.

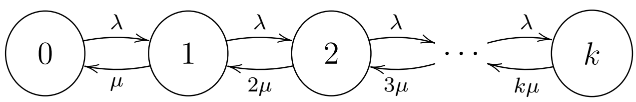
The reversibility property of M/M/$k$/$k$ and the truncation at $k$ imply that the detailed balance equations for M/M/$k$/$k$ are the same as the first $k$ detailed balance (steady-state) equations for M/M/$\infty$. Namely, these balance equations are:

$$\pi_n \lambda = \pi_{n+1}(n+1)\mu, \text{ for } n = 0,\ 1,\ 2,\ \ldots\ k-1. \tag{315}$$

The infinitesimal generator of M/M/$k$/$k$ is given by

$Q_{i,i+1} = \lambda$ for $i = 0, 1, 2, 3, \ \ldots, \ k-1$
$Q_{i,i-1} = i\mu$ for $i = 1, 2, 3, 4, \ \ldots, \ k$
$Q_{0,0} = -\lambda$
$Q_{i,i} = -\lambda - i\mu$ for $i = 1, 2, 3, \ \ldots, \ k-1$
$Q_{k,k} = -k\mu$.

The balance equations can also be described by the following state transition diagram of M/M/$k$/$k$:




The sum of the steady-state probabilities must be equal to one, so we again have the additional normalizing equation

$$\sum_{j=0}^{k} \pi_j = 1. \tag{316}$$

Accordingly, we obtain for M/M/$k$/$k$:

$$\pi_n = \frac{A^n \pi_0}{n!} \text{ for } n = 0,\ 1,\ 2,\ \ldots, k. \tag{317}$$

To obtain $\pi_0$, we again sum up both sides of the latter. This leads to

$$\pi_0 = \frac{1}{\sum_{n=0}^{k} \frac{A^n}{n!}}. \tag{318}$$

Substituting Eq. (318) in Eq. (317), we obtain

$$\pi_n = \frac{\frac{A^n}{n!}}{\sum_{n=0}^{k} \frac{A^n}{n!}} \text{ for } n = 0,\ 1,\ 2,\ \ldots, k. \tag{319}$$

The relationship between (319) and (308) is now clearer. Observing (319) that gives the distribution of the number of customers in an M/M/$k$/$k$ model, it is apparent that it is a truncated version of (308). Since (308) is merely the Poisson distribution, (319) is the truncated Poisson distribution. Accordingly, to obtain (319), we can simply consider (308), and firstly set $\pi_j = 0$ for all $\pi_j$ with $j > k$. Then for $0 \leq j \leq k$ we set the $\pi_j$ for the M/M/$k$/$k$ values by dividing the $\pi_j$ values of (308) by the sum $\sum_{j=0}^{k} \pi_j$ of the $\pi_j$ values in the M/M/$\infty$ model. This is equivalent to considering the M/M/$\infty$ model and deriving the conditional probability of the process being in state $j$ for $j = 0, 1, 2, \ldots, k$, conditioning on the process being within the states $j = 0, 1, 2, \ldots, k$. This conditional probability is exactly the steady-state probabilities $\pi_j$ of the M/M/$k$/$k$ model.

The most important quantity out of the values obtained by Eq. (319) is $\pi_k$. It is the probability that all $k$ circuits are busy, so it is the proportion of time that no new calls can enter the system, namely, they are blocked. It is therefore called *time congestion*. The quantity $\pi_k$ for an M/M/$k$/$k$ system loaded by offered traffic $A$ is usually denoted by $E_k(A)$ and is given by:

$$E_k(A) = \frac{\frac{A^k}{k!}}{\sum_{n=0}^{k} \frac{A^n}{n!}}. \tag{320}$$

Eq. (320) is known as Erlang's loss Formula, or Erlang B Formula, published first by A. K. Erlang in 1917 [24].

Due to the special properties of the Poisson process, in addition of being the proportion of time during which the calls are blocked, $E_k(A)$ also gives the proportion of calls blocked due to congestion; namely, it is the *call congestion* or *blocking probability*. A simple way to explain that for an M/M/$k$/$k$ system the call congestion (blocking probability) is equal to the time congestion is the following. Let $L$ be an arbitrarily long period of time. The proportion of time during $L$ when all servers are busy and every arrival is blocked is $\pi_k = E_k(A)$, so the time during $L$ when new arrivals are blocked is $\pi_k L$. The mean number of blocked arrivals during $L$ is therefore equal to $\lambda \pi_k L$. The mean total number of arrivals during $L$ is $\lambda L$. The blocking probability (call congestion) $P_b$ is the ratio between the two. Therefore:

$$P_b = \frac{\lambda \pi_k L}{\lambda L} = \pi_k = E_k(A).$$

Eq. (320) has many applications to telecommunications network design. Given its importance, it is necessary to be able to compute Eq. (320) quickly and exactly for large values of $k$. This will enable us to answer a dimensioning question, for example: "how many circuits are required so that the blocking probability is no more than 1% given offered traffic of $A = 1000$?".

## 8.3 Recursion and Jagerman Formula

Observing Eq. (320), we notice the factorial terms which may hinder such computation for a large $k$. We shall now present an analysis which leads to a recursive relation between $E_m(A)$

and $E_{m-1}(A)$ that gives rise to a simple and scalable algorithm for the blocking probability. By Eq. (320), we obtain

$$\frac{E_m(A)}{E_{m-1}(A)} = \frac{\frac{\frac{A^m}{m!}}{\sum_{j=0}^{m}\frac{A^j}{j!}}}{\frac{\frac{A^{m-1}}{(m-1)!}}{\sum_{j=0}^{m-1}\frac{A^j}{j!}}} = \frac{\frac{\frac{A^m}{m!}}{\sum_{j=0}^{m}\frac{A^j}{j!}}}{\frac{\frac{A^{m-1}}{(m-1)!}}{\sum_{j=0}^{m}\frac{A^j}{j!}-\frac{A^m}{m!}}} = \frac{A}{m}(1-E_m(A)). \tag{321}$$

Isolating $E_m(A)$, this leads to

$$E_m(A) = \frac{AE_{m-1}(A)}{m + AE_{m-1}(A)} \text{ for } m = \ 1,\ 2,\ \ldots, k. \tag{322}$$

**Homework 8.1**

Complete all the details in the derivation of Eq. (322). □

When $m = 0$, there are no servers (circuits) available, and therefore all customers (calls) are blocked, namely,

$$E_0(A) = 1. \tag{323}$$

The above two equations give rise to a simple recursive algorithm by which the blocking probability can be calculated for a large $k$. An even more computationally stable way to compute $E_m(A)$ for large values of $A$ and $m$ is to use the inverse [47]

$$I_m(A) = \frac{1}{E_m(A)} \tag{324}$$

and the recursion

$$I_m(A) = 1 + \frac{m}{A}I_{m-1}(A) \text{ for } m = \ 1,\ 2,\ \ldots, k. \tag{325}$$

with the initial condition $I_0(A) = 1$.

A useful formula for $I_m(A)$ due to Jagerman [48] is:

$$I_m(A) = A\int_0^\infty e^{-Ay}(1+y)^m dy. \tag{326}$$

**Homework 8.2**

Long long ago in a far-away land, John, an employee of a telecommunication provider company, was asked to derive the blocking probability of a switching system loaded by a Poisson arrival process of calls where the offered load is given by $A = 180$. These calls are served by $k = 200$ circuits.

The objective was to meet a requirement of no more than 1% blocking probability. The company has been operating with $k = 200$ circuits for some time and there was a concern that the blocking probability exceeds the 1% limit.

John was working for a while on the calculation of this blocking probability, but when he was very close to a solution, he won the lottery, resigned and went to the Bahamas. Mary, another employee of the same company, was given the task of solving the problem. She found some of John's papers where it was revealed that for an M/M/196/196 model and $A = 180$, the blocking probability is approximately 0.016. Mary completed the solution in a few minutes. Assuming that John's calculations were correct, what is the solution for the blocking probability of M/M/200/200 with $A = 180$? If the blocking probability in the case of $k = 200$ is more that 1%, what is the smallest number of circuits that should be added to meet the 1% requirement?

**Solution**

Using the Erlang B recursion (322) and knowing that $E_{196}(180) = 0.016$, we obtain

$E_{197}(180) \approx 0.0145$
$E_{198}(180) \approx 0.013$
$E_{199}(180) \approx 0.0116$
$E_{200}(180) \approx 0.0103$.

One more circuit should be added to achieve:

$E_{201}(180) \approx 0.0092$ □

## 8.4 The Special Case: M/M/1/1

**Homework 8.3**

Derive a formula for the blocking probability of M/M/1/1 in four ways: (1) by Erlang B Formula (320), (2) by the recursion (322), (3) by the recursion (325), and (4) by Jagerman Formula (326). □

The reader may observe a fifth direct way to obtain a formula for the blocking probability of M/M/1/1 using Little's formula. The M/M/1/1 system can have at most one customer in it. Therefore, its mean queue size is given by $E[Q] = 0\pi_0 + 1\pi_1 = \pi_1$ which is also its blocking probability. Noticing also that the arrival rate into the system (made only of successful arrivals) is equal to $\lambda(1 - E[Q])$, the mean time a customer stays in the system is $1/\mu$, and revoking Little's formula, we obtain

$$\frac{\lambda(1 - E[Q])}{\mu} = E[Q]. \tag{327}$$

Isolating $E[Q]$, the blocking probability is given by

$$\pi_1 = E[Q] = \frac{A}{1 + A}. \tag{328}$$

## 8.5 Lower bound of $E_k(A)$

By (314) and since $E[Q]$ satisfies $E[Q] < k$ [38], we obtain

$$A(1 - E_k(A)) < k,$$

or following simple algebraic operations

$$E_k(A) > 1 - \frac{k}{A}. \tag{329}$$

Clearly, the latter is relevant only when $k < A$; otherwise, a tighter bound is obtained by $E_k(A) \geq 0$. Thus, for the full range of $A \geq 0$ and $k = 1, 2, 3\ldots$ values, $E_k(A)$ satisfies [89]

$$E_k(A) \geq \max\left(0, 1 - \frac{k}{A}\right). \tag{330}$$

## 8.6 Monotonicity of $E_k(A)$

Intuitively, the blocking probability $E_k(A)$ is expected to increase as the offered traffic $A$ increases, and to decrease as the number of server $k$ increase.

To show that $E_k(A)$ is monotonic in $A$, we use induction on $k$ and the Erlang B recursion (322) [89].

**Homework 8.4**

Complete all the details in the proof that $E_k(A)$ is monotonic in $A$. □

Then, to show that $E_k(A)$ is monotonic in $k$, recall that by (321), for $k = 1, 2, 3, \ldots$ we have

$$\frac{E_k(A)}{E_{k-1}(A)} = \frac{A}{k}[1 - E_k(A)].$$

Then by (329) we have

$$1 - E_k(A) < \frac{k}{A}.$$

Thus,

$$\frac{E_k(A)}{E_{k-1}(A)} < \frac{A}{k}\frac{k}{A} = 1,$$

so

$$E_k(A) < E_{k-1}(A).$$

## 8.7 The Limit $k \to \infty$ with $A/k$ Fixed

As traffic and capacity (number of servers) increase, there is an interest in understanding the blocking behavior in the limit $k$ and $A$ both approach infinity with $A/k$ Fixed.

**Homework 8.5**

Prove that the blocking probability approaches zero for the case $A/k \leq 1$ and that it approaches $1 - k/A$ in the case $A/k > 1$.

**Guide (by Guo Jun based on [85])**

By (326) we have.

$$\frac{1}{\mathbf{E}(A,k)} = \int_0^\infty e^{-t}\left(1+\frac{t}{A}\right)^k dt.$$

Consider the case where $k$ increases in such a way that $A/k$ is constant. Then,

$$\begin{aligned}\lim_{k\to\infty}\frac{1}{\mathbf{E}(A,k)} &= \lim_{k\to\infty}\int_0^\infty e^{-t}\left(1+\tfrac{t}{A}\right)^k dt\\ &= \int_0^\infty e^{-t}\cdot e^{\frac{t}{A/k}}\, dt\\ &= \int_0^\infty e^{-(1-\frac{1}{A/k})t}\, dt.\end{aligned}$$

Then observe that

$$\lim_{k\to\infty}\frac{1}{\mathbf{E}(A,k)} = \begin{cases}\infty & \text{if } A/k \leq 1\\ \frac{1}{1-\frac{1}{A/k}} & \text{if } A/k > 1.\end{cases} \tag{331}$$

And the desired result follows. □

**Homework 8.6 (Jiongze Chen)**

Provide an alternative proof to the results of the previous homework using the Erlang B recursion.

**Guide**

Set $a = A/k$. Notice that in the limit $E_k(ak) \cong E_{k+1}(a(k+1)) \cong E_k(A)$ and provide a proof for the cases $k < A$ and $k = A$. Then, for the case $k > A$, first show that the blocking probability decreases as $k$ increases for a fixed $A$ (using the Erlang B recursion), and then argue that if the blocking probability already approaches zero for $A = k$, it must also approach zero for $k > A$. □

**Homework 8.7**

Provide intuitive explanation to the results of the previous homework.

**Guide**

Due to the insensitivity property, M/M/$k$/$k$ and M/D/$k$/$k$ experience the same blocking probability if the offered traffic in both system is the same. Observe that since the arrivals follow a Poisson process the variance is equal to the mean. Also notice that as the arrival rate $\lambda$ increases, the Poisson process approaches a Gaussian process. Having the variance equal to the mean, the standard deviation becomes negligible relative to the mean for a very large $\lambda$. With negligible variability, M/D/$k$/$k$ behaves like D/D/$k$/$k$ and the results follow. □

## 8.8 M/M/$k$/$k$: Dimensioning and Utilization

Taking advantage of the monotonicity of Erlang formula, we can also solve the dimensioning problem. We simply keep incrementing the number of circuits and calculate in each case the blocking probability. When the desired blocking probability (e.g., 1%) is reached, we have our answer.

**Homework 8.8**

Prove that if $A > A'$ then $E_n(A) > E_n(A')$ for $n = 1, 2, 3 \ldots$.
**Hint:** Consider the Erlang B recursion and use induction on $n$. □

We have already derived the mean number of busy circuits in an M/M/$k$/$k$ system fed by $A$ erlangs in (314) using Little's formula. Substituting $\pi_k$ for $P_b$ in (314), we obtain

$$E[Q] = (1 - \pi_k)A.$$

Note that it can also be computed by the weighted sum

$$E[Q] = \sum_{i=0}^{k} i\pi_i.$$

Accordingly, the utilization of an M/M/$k$/$k$ system is given by

$$\hat{U} = \frac{(1 - \pi_k)A}{k}. \tag{332}$$

**Homework 8.9**

Prove that $\sum_{i=0}^{k} i\pi_i = (1 - \pi_k)A$. □

In the following Table, we present the minimal values of $k$ obtained for various values of $A$ such that the blocking probability is no more than 1%, and the utilization obtained in each case. It is clearly observed that the utilization increases with the traffic.

Note that normally it is impossible to find $k$ that gives exactly 1% blocking probability, so we conservatively choose $k$ such that $E_k(A) < 1\%$, but we must make sure that $E_{k-1}(A) > 1\%$. Namely, if we further reduce $k$, the blocking probability requirement to be no more than 1% is violated.

| $A$ | $k$ | $E_k(A)$ | Utilization |
|---|---|---|---|
| 20 | 30 | 0.0085 | 66.10% |
| 100 | 117 | 0.0098 | 84.63% |
| 500 | 527 | 0.0095 | 93.97% |
| 1000 | 1029 | 0.0099 | 96.22% |
| 5000 | 5010 | 0.0100 | 98.81% |
| 10000 | 9970 | 0.0099 | 99.30% |

**Homework 8.10**

Reproduce the above Table. □

We also notice that for the case of $A = 10,000$ erlangs, to maintain no more than 1% blocking, $k$ value less than $A$ is required. Notice however that the carried traffic is not $A$ but $A(1-E_k(A))$. This means that for $A \geq 10,000$, dimensioning simply by $k = A$ will mean no more than 1% blocking and no less than 99% utilization - not bad for such a simple rule of thumb! This also implies that if the system capacity is much larger than individual service requirement, very high efficiency (utilization) can be achieved without a significant compromise on quality of service. Let us now further examine the case $k = A$.

## 8.9 M/M/$k$/$k$ under Critical Loading

A system where the offered traffic load is equal to the system capacity is called *critically loaded* [43]. Accordingly, in a critically loaded Erlang B System we have $k = A$. From the table below, it is clear that if we maintain $k = A$ and we increase them both, the blocking probability decreases, the utilization increases, and interestingly, the product $E_k(A)\sqrt{A}$ approaches a constant, which we denote $\tilde{C}$, that does not depend on $A$ or $k$. This implies that in the limit, the blocking probability decays at the rate of $1/\sqrt{k}$. That is, for a critically loaded Erlang B system, we obtain

$$\lim_{k\to\infty} E_k(A) = \frac{\tilde{C}}{\sqrt{k}}. \tag{333}$$

| $A$ | $k$ | $E_k(A)$ | Utilization | $E_k(A)\sqrt{A}$ |
|---|---|---|---|---|
| 10 | 10 | 0.215 | 78.5 % | 0.679 |
| 50 | 50 | 0.105 | 89.5% | 0.741 |
| 100 | 100 | 0.076 | 92.4% | 0.757 |
| 500 | 500 | 0.035 | 96.5% | 0.779 |
| 1000 | 1000 | 0.025 | 97.5% | 0.785 |
| 5000 | 5000 | 0.011 | 98.9% | 0.792 |
| 10000 | 10000 | 0.008 | 99.2% | 0.79365 |
| 20000 | 20000 | 0.00562 | 99.438% | 0.79489 |
| 30000 | 30000 | 0.00459 | 99.541% | 0.79544 |
| 40000 | 40000 | 0.00398 | 99.602% | 0.79576 |
| 50000 | 50000 | 0.00356 | 99.644% | 0.79599 |

To explain the low blocking probability in critically loaded large system, we refer back to our homework problem related to an Erlang B system with large capacity where the ratio $A/k$ is maintained constant. In such a case the standard deviation of the traffic is very small relative to the mean, so the traffic behaves close to deterministic. If 100 liters per second of water are offered, at a constant rate, to a pipe that has capacity of 100 liters per second, then the pipe can handle the offered load with very small losses.

## 8.10 Insensitivity and Many Classes of Customers

We have discussed in Section 7.5, the distribution and the mean of the number of busy servers is insensitive to the shape of the service time distribution (although it is still sensitive to the mean of the service time) in the cases of M/G/$\infty$ and M/G/$k$/$k$. For M/G/$k$/$k$, also the blocking probability is insensitive to the shape of the service time distribution [37, 81, 82].

However, we must make it very clear that the insensitivity property does not extend to the arrival process. We still require a Poisson process for the Erlang B formula to apply. If we have a more burtsy arrival process (e.g. arrivals arrive in batches) we will have more losses than predicted by Erlang B formula, and if we have a smoother arrival process than Poisson, we will have less losses than predicted by Erlang B formula. To demonstrate it, let us compare an M/M/1/1 system with a D/D/1/1 system. Suppose that each of these two systems is fed by $A$ erlangs, and that $A < 1$.

Arrivals into the D/D/1/1 system with $A < 1$ will never experience losses because the inter-arrivals are longer that the service times, so the service of a customer is always completed before the arrival of the next customer. Accordingly, by Little's formula: $E[Q] = A$, and since $E[Q] = 0 \times \pi_0 + 1 \times \pi_1$, we have that $\pi_1 = A$ and $\pi_0 = 1 - A$. In this case, the blocking probability $P_b$ is equal to zero and not to $\pi_1$. As there are no losses, the utilization will be given by $\hat{U} = \pi_1 = A$.

By contrast, for the M/M/1/1 system, $P_b = E_1(A) = E[Q] = \pi_1 = A/(1+A)$, so $\pi_0 = 1 - \pi_1 = 1/(1+A)$. To obtain the utilization we can either realized that it is the proportion of time our single server is busy, namely it is equal to $\pi_1 = A/(1+A)$, or we can use the above formula for $\hat{U}$ in M/M/$k$/$k$ system and obtain

$$\hat{U} = (1-\pi_k)A = [1 - A/(1+A)]A = A/(1+A). \tag{334}$$

This comparison is summarized in the following table:

| | M/M/1/1 | D/D/1/1 |
|---|---|---|
| $\pi_0$ | $1/(1+A)$ | $1-A$ |
| $\pi_1$ | $A/(1+A)$ | $A$ |
| $\hat{U}$ | $A/(1+A)$ | $A$ |
| $P_b$ | $A/(1+A)$ | 0 |
| $E[Q]$ | $A/(1+A)$ | $A$ |

Clearly, the steady-state equations (317) will not apply to a D/D/1/1 system.

We have already mentioned that for M/G/$k$/$k$ the distribution of the number of busy servers and therefore also the blocking probability is insensitive to the shape of the service time distribution (moments higher than the first). All we need is to know that the arrival process is Poisson, and the ratio of the arrival rate to the service rate of a single server and we can obtain the blocking probability using the Erlang B formula. Let us now consider the following problem.

Consider two classes of customers (packets). Class $i$ customers, $i = 1, 2$ arrive following an independent Poisson process at rate of $\lambda_i$ each of which requires independent and exponentially distributed service time with parameter $\mu_i$, for $i = 1, 2$. There are $k$ independent servers without waiting room (without additional buffer). The aim is to derive the blocking probability $E_k(A)$.

The combined arrival process of all the customers is a Poisson process with parameter $\lambda = \lambda_1 + \lambda_2$. Because the probability of an arbitrary customer to belong to the first class is

$$p = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{\lambda_1}{\lambda},$$

the service time of an arbitrary customer has hyperexponential distribution because with probability $p$ it is exponentially distributed with parameter $\mu_1$, and with probability $1-p$, it is exponentially distributed with parameter $\mu_2$.

Therefore, by the Law of iterated expectation, the mean service time (holding time) is given by

$$E[S] = \frac{p}{\mu_1} + \frac{1-p}{\mu_2}$$

so $A = \lambda E[S]$, and Erlang B can then be used to obtain the blocking probability $E_k(A)$.

Furthermore, let

$$A_i = \frac{\lambda_i}{\mu_i} \quad i = 1, 2$$

and observe that

$$E[S] = \left(\frac{\lambda_1}{\lambda}\right)\left(\frac{1}{\mu_1}\right) + \left(\frac{\lambda_2}{\lambda}\right)\left(\frac{1}{\mu_2}\right) = \frac{A_1 + A_2}{\lambda}.$$

Then

$$A = \lambda E[S] = A_1 + A_2.$$

**Homework 8.11 [13]**

Consider $n$ classes of calls that are offered to a loss system with $k$ servers and without waiting room. The arrival processes of the different classes follow independent Poisson processes. The arrival rate of calls of class $i$ is $\lambda_i$ and their mean holding times is $1/\mu_i$, $i = 1, 2, \ldots, n$. Find the blocking probability.

**Guide**

$$\lambda = \sum_{i=1}^{n} \lambda_i$$

$$E[S] = \sum_{i=1}^{n} \left(\frac{\lambda_i}{\lambda}\right)\left(\frac{1}{\mu_i}\right)$$

$$A = \lambda E[S] = \sum_{i=1}^{n} A_i$$

where

$$A_i = \frac{\lambda_i}{\mu_i} \quad i = 1, 2, \ldots n.$$

Then the blocking probability $E_k(A)$ is obtained by Erlang B formula. □

**Homework 8.12**

1. Consider an M/M/$\infty$ queueing system with the following twist. The servers are numbered 1, 2, ... and an arriving customer always chooses the server numbered lowest among all the free servers it finds. Find the proportion of time that each of the servers is busy [13].

   **Guide**

   Notice that the input (arrival) rate into the system comprises servers $n+1, n+2, n+3\ldots$ is equal to $\lambda E_n(A)$. Then using Little's formula notice that the mean number of busy servers among $n+1, n+2, n+2\ldots$ is equal to $AE_n(A)$. Repeat the procedure for the system comprises servers $n+2, n+3, n+4\ldots$, you will observe that the mean number of busy servers in this system is $AE_{n+1}(A)$. Then considering the difference between these two mean values, you will obtain that the mean number of busy servers in a system comprises only of the $n+1$ server is

   $$A[E_n(A) - E_{n+1}(A)].$$

   Recall that the mean queue size (mean number of busy server) of a system that comprises only the single server is (probability of server is busy) times 1 + (probability of server is idle) times 0, which is equal to the probability that the server is busy, we obtain that $A[E_n(A) - E_{n+1}(A)]$ is the probability that the server is busy.

An alternative way to look at this problem is the following. Consider the system made only of the $n+1$ server. The offered traffic into this system is $AE_n(A)$, the rejected traffic of this system is $AE_{n+1}(A)$. Therefore, the carried traffic of this system is $A[E_n(A) - E_{n+1}(A)]$. This means that the arrival rate of customers that actually enters this single server system is

$$\lambda_{\text{enters}(n+1)} = \lambda[E_n(A) - E_{n+1}(A)]$$

and since the mean time spent in this system is $1/\mu$, we have that the mean queue size in this single server system is

$$\lambda_{\text{enters}(n+1)}\frac{1}{\mu} = A[E_n(A) - E_{n+1}(A)]$$

which is the carried load. Based on the arguments above it is equal to the proportion of time the $n+1$ server is busy.

2. Show that if the number of servers is finite $k$, the proportion of time that server $n+1$ is busy is

$$A\left(1 - \frac{E_k(A)}{E_n(A)}\right)E_n(A) - A\left(1 - \frac{E_k(A)}{E_{n+1}(A)}\right)E_{n+1}(A) = A[E_n(A) - E_{n+1}(A)]$$

and provide intuitive arguments to why the result is the same as in the infinite server case.

3. Verify the results by discrete-event and Markov-chain simulations.

□

**Homework 8.13**

Consider an M/M/$k$/$k$ queue with a given arrival rate $\lambda$ and mean holding time $1/\mu$. Let $A = \lambda/\mu$. Let $E_k(A)$ be the blocking probability. An independent Poisson inspector inspects the M/M/$k$/$k$ queue at times $t_1, t_2, t_3, \ldots$ . What is the probability that the first arrival after an inspection is blocked?

**Answer:**

$$\frac{E_k(A)\lambda}{k\mu + \lambda}.$$

□

**Homework 8.14**

Bursts of data of exponential lengths with mean $1/\mu$ that arrive according to a Poisson process are transmitted through a bufferless optical switch. All arriving bursts compete for $k$ wavelength channels at a particular output trunk of the switch. If a burst arrives and all $k$ wavelength channels are busy, the burst is dumped at the wavelength bit-rate. While it is being dumped, if one of the wavelength channels becomes free, the remaining portion of the burst is successfully transmitted through the wavelength channel.

1. Show that the mean loss rate of data $E[Loss]$ is given by

$$E[Loss] = 1P(X = k+1) + 2P(X = k+2) + 3P(X = k+3) + \dots$$

   where $X$ is a Poisson random variable with parameter $A = \lambda/\mu$.

2. Prove that

$$E[Loss] = \frac{A\gamma(k, A)}{\Gamma(k)} - \frac{k\gamma(k+1, A)}{\Gamma(k+1)}$$

   where $\Gamma(k)$ is the Gamma function and $\gamma(k, A)$ is the lower incomplete Gamma function.

**Background information and guide**

The Gamma function is defined by

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt. \tag{335}$$

The lower incomplete Gamma function is defined by

$$\gamma(a, x) = \int_0^x t^{a-1}e^{-t}dt. \tag{336}$$

The upper incomplete Gamma function is defined by

$$\Gamma(a, x) = \int_x^\infty t^{a-1}e^{-t}dt. \tag{337}$$

Accordingly,

$$\gamma(a, x) + \Gamma(a, x) = \Gamma(a).$$

For an integer $k$, we have

$$\Gamma(k) = (k-1)! \tag{338}$$

$$\Gamma(k, x) = (k-1)!e^{-x}\sum_{m=0}^{k-1}\frac{x^m}{m!}. \tag{339}$$

Therefore,

$$e^{-A}\sum_{m=0}^{k-1}\frac{A^i}{i!} = \frac{\Gamma(k, A)}{\Gamma(k)} \tag{340}$$

so

$$1 - e^{-A}\sum_{m=0}^{k-1}\frac{A^i}{i!} = 1 - \frac{\Gamma(k, A)}{\Gamma(k)} = \frac{\Gamma(k) - \Gamma(k, A)}{\Gamma(k)} = \frac{\gamma(k, A)}{\Gamma(k)}. \tag{341}$$

Now notice that

$$\begin{aligned}
E[Loss] &= 1 \times P(X=k+1) + 2 \times P(X=k+2) + 3 \times P(X=k+3) + \ldots \\
&= \sum_{i=k+1}^{\infty} (i-k) A^i \frac{e^{-A}}{i!} \\
&= A \sum_{i=k+1}^{\infty} A^{i-1} \frac{e^{-A}}{(i-1)!} - k \sum_{i=k+1}^{\infty} A^i \frac{e^{-A}}{i!} \\
&= A \sum_{i=k}^{\infty} A^i \frac{e^{-A}}{i!} - k \sum_{i=k+1}^{\infty} A^i \frac{e^{-A}}{i!} \\
&= A \left[ 1 - e^{-A} \sum_{i=0}^{k-1} \frac{A^i}{i!} \right] - k \left[ 1 - e^{-A} \sum_{i=0}^{k} \frac{A^i}{i!} \right] \\
&= \frac{A\gamma(k,A)}{\Gamma(k)} - \frac{k\gamma(k+1,A)}{\Gamma(k+1)}. \qquad \square
\end{aligned}$$

## 8.11 First Passage Time and Time Between Events in M/M/$k$/$k$

As discussed the term first passage time between states $i$ and $j$ is the time it takes for a process in state $i$ to enter state $j$ for the first time. To enhance the understanding of Markov chains in general and M/M/$k$/$k$ in particular, we consider here several examples associated with the first passage time as well as problems associated with times between events in the context of M/M/$k$/$k$ These example will enhance understanding of Markov chains and provide preparation to the section on Markov chain simulation of M/M/$k$/$k$. It is also important to understand that derivations of first passage times and times between events are applicable to more general class of Markov chains and Markovian queueing models and they are clearly not limited M/M/$k$/$k$.

**Homework 8.15**

Consider an M/M/4/4 system with arrival rate $\lambda = 1$ per minute and holding time 180 seconds. Assume that at a given point in time $t$, there are 3 servers busy. In each of the questions below, write the formula first and then substitute correctly the numerical values and compute the final answer in numerical value.

1. Find the mean time from $t$ until the next arrival.

   **Solution**

   The mean time from $t$ until the next arrival is

   $$\frac{1}{\lambda} = \frac{1}{1} = 1.$$

2. Find the mean time from $t$ until the next event (either arrival or departure). (3 marks)

   **Solution**

$$\mu = \frac{1}{3}[\text{min.}^{-1}].$$

The mean time from $t$ until the next event is

$$\frac{1}{\lambda + 3\mu} = \frac{1}{1 + 3 \times (1/3)} = 0.5[\text{min.}].$$

3. Find the probability that the next event will be an arrival.

   **Solution**

$$\frac{\lambda}{\lambda + 3\mu} = \frac{1}{1 + 3 \times (1/3)} = 0.5.$$

4. Find the probability that the next event will be a departure.

   **Solution**

$$\frac{3\mu}{\lambda + 3\mu} = \frac{3 \times (1/3)}{1 + 3 \times (1/3)} = 0.5.$$

5. Find the mean time from $t$ until the next departure.

   **Solution**

   Let $X$ be the time from $t$ until the next departure. Then $X$ can be divided into two parts: $X_1$ and $X_2$, where $X_1$ is the time until the next event (which can be an arrival or a departure) and $X_2$ is the time from the next event until the first departure.

   Then $E[X_1]$ was already obtained in Part 2 of this question, and we have:

$$E[X_1] = \frac{1}{\lambda + 3\mu} = \frac{1}{1 + 3 \times (1/3)} = 0.5[\text{min.}].$$

   and $E[X_2]$ is obtained by the Law of Iterated Expectations as follows:

$$E[X_2] = \frac{\lambda}{\lambda + 3\mu} \times \frac{1}{4\mu} + \frac{3\mu}{\lambda + 3\mu} \times 0 = 0.5 \times 0.75 = 0.375[\text{min.}].$$

   Then we obtain

$$E[X] = E[X_1] + E[X_2] = 0.5 + 0.375 = 0.875[\text{min.}].$$

**Homework 8.15**

Consider an M/M/2/2 queueing system with $\lambda$ being the arrival rate and $1/\mu$ is the mean holding time. Assume the case $\lambda = 9$ and $\mu = 5$. At time $t$, there are two customers in the system. What is the mean time from time $t$ until the first time that the system is empty? (In other words, what is the mean first passage time from state 2 to state 0.)

**Solution**

Let $m_{2,0}$ be the mean first passage time from state 2 to state 0.

Let $m_{1,0}$ be the mean first passage time from state 1 to state 0.

Then,

$$m_{2,0} = \frac{1}{2\mu} + m_{1,0}$$

and

$$m_{1,0} = \frac{1}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} m_{2,0}$$

Substituting the above values for $\lambda$ and $\mu$, we obtain

$$m_{2,0} = \frac{1}{10} + m_{1,0}$$

and

$$m_{1,0} = \frac{1}{14} + \frac{9}{14} m_{2,0}$$

or

$$m_{2,0} = \frac{1}{10} + \frac{1}{14} + \frac{9}{14} m_{2,0}$$

from which we obtain

$$m_{2,0} = \frac{14}{50} + \frac{1}{5} = \frac{24}{50} = 0.48.$$

## 8.12 A Markov-chain Simulation of M/M/$k$/$k$

We have described a Markov-chain simulation in the context of the M/M/1 queue. In a similar way, we can use a Markov-chain simulation to evaluate the blocking probability of an M/M/$k$/$k$ queue, as follows.

Variables and input parameters:
$k$ = number of servers;
$Q$ = number of customers in the system (queue size);
$B_p$ = estimation for the blocking probability;
$N_a$ = number of customer arrivals counted so far;
$N_b$ = number of blocked customers counted so far;
$MAXN_a$ = maximal number of customer arrivals (it is used for the stopping condition);
$\mu$ = service rate;
$\lambda$ = arrival rate.

Define function: $R(01)$ = a uniform $U(0,1)$ random deviate. A new value for $R(01)$ is generated every time it is called.

Initialization: $Q = 0$; $N_a = 0$, $N_b = 0$.

1. If $R(01) \leq \lambda/(\lambda + Q\mu)$, then $N_a = N_a + 1$; if $Q = k$ then $N_b = N_b + 1$, else $Q = Q + 1$;

else, Q = Q - 1.

2. If $N_a < MAXN_a$ go to 1; else, print $B_p = N_b/N_a$.

Again, it is a very simple program of two If statements: one to check if the next event is an arrival or a departure, and the other a stopping criterion.

**Homework 8.15**

Simulate an M/M/$k$/$k$ queue based on the Markov-chain simulation principles to evaluate the blocking probability for a wide range of parameter values. Compare the results you obtained with equivalent results obtain analytically using the Erlang B Formula and with equivalent M/M/$k$/$k$ queue blocking probability results obtained based on discrete-event simulations as discussed in Section 4.2. In your comparison consider accuracy (closeness to the analytical results), the length of the confidence intervals and running times. □

**Homework 4.16**

Use the principles presented in Section 4.2 for discrete-event simulation of a G/G/1 queue to write three computer simulations for an M/G/$k$/$k$ queue where G is the service time modeled by

1. exponential distribution
2. uniform distribution (continuous)
3. deterministic (constant) value – it always takes the same value
4. Pareto distribution
5. non-parametric distribution,

where the mean service times in all cases is the same. Choose your own parameters for the different distributions (or values for the non-parametric distribution), but you must make sure to keep the mean service time the same in all cases.

Demonstrate the insensitivity property of M/G/$k$/$k$ by showing that the blocking probability is the same in all cases. Obtain the blocking probability also by Erlang B formula and by a Markov chain simulation for the case of exponentially distributed service times. Repeat your simulation experiments for different cases using a wide range of parameter values. □

**Homework 8.17**

Simulate equivalent U/U/$k$/$k$, M/U/$k$/$k$ ($U$ denotes here a uniform random variable) and M/M/$k$/$k$ models. (You may use programs you have written in previous assignments. Run these simulations for a wide range of parameter values and compare them numerically. Compare them also with equivalent results obtain analytically using the Erlang B Formula. Again, in your comparison consider accuracy (closeness to the analytical results), the length of the confidence intervals and running times. While in the previous assignment, you learn the effect of the method used on accuracy and running time, this time try also to learn how the different models affect accuracy and running times. □

**Homework 8.18**

Use the M/M/$k$/$k$ model to compare the utilization of an optical switch with full wavelength conversion and without wavelength conversion.

**Background information and guide**

Consider a switch with $T_I$ input trunks and $T_O$ output trunks. Each trunk comprises $F$ optical fibers each of which comprises $W$ wavelengths. Consider a particular output trunk and assume that the traffic directed to it follows a Poisson process with parameter $\lambda$ and that any packet is of exponential length with parameter $\mu$. Let $A = \lambda/\mu$. In the case of full wavelength conversion, every packet from any wavelength can be converted to any other wavelength, so the Poisson traffic with parameter $\lambda$ can all be directed to the output trunk and can use any of the $k = FW$ links. In the case of no wavelength conversion, if a packet arrives on a given wavelength at an input port, it must continue on the same wavelength at the output port, so now consider $W$ separate systems each has only $F$ links per trunk. Compare the efficiency that can be achieved for both alternatives, if the blocking probability is set limited to 0.001. In other words, in the wavelength conversion case, you have an M/M/$k$/$k$ system with $k = FW$, and in non-wavelength conversion case, you have $k = F$. Compute the traffic $A$ the gives blocking probability of 0.001 in each case and compare efficiency. Realistic ranges are $40 \leq W \leq 100$ and $10 \leq F \leq 100$. □

## 8.13 M/M/$k$/$k$ with Preemptive Priorities

So far we have considered a single class traffic without any priority given to some calls over others. Let us now consider a scenario that is common in many applications, where some calls have preemptive priorities over other lower priority calls. In this case, when a higher priority arrives and non of the $k$ servers is available, the higher priority call preempts one of the calls in service and enters service instead of the preempted lower priority call. We consider arriving calls to be of $m$ priority types. Where priority 1 represents the highest priority and priority $m$ represents the lowest priority. In general, if $i < j$ then priority $i$ is higher than priority $j$, so a priority $i$ arrival may preempt a priority $j$ customer upon its arrival.

The arrival process of priority $i$ customers follows a Poisson process with rate $\lambda_i$, for $i = 1, 2, 3, \ldots m$. The service time of all the customers is exponentially distributed with parameter $\mu$. The offered traffic of priority $i$ customers is given by

$$A_i = \frac{\lambda_i}{\mu}, \quad i = 1, 2, 3, \ldots m.$$

Let $P_b(i)$ be the blocking probability of priority $i$ customers. Because the priority 1 traffic access the system regardless of low priority loading, for the case $i = 1$, we have

$$P_b(1) = E_k(A_1).$$

To obtain $P_b(i)$ for $m \geq i > 1$, we first observe that the blocking probability of all the traffic of priority $i$ and higher priorities, namely, the traffic generated by priorities $1, 2, \ldots, i$, is given by

$$E_k(A_1 + A_2 + \ldots, A_i).$$

Next we observe that the lost traffic of priority $i$, $i = 1, 2, 3, \ldots m$, is given by the lost traffic of priorities $1, 2, 3, \ldots i$ minus the lost traffic of priorities $1, 2, 3, \ldots i - 1$, namely,

$$(A_1 + A_2 + \ldots, A_i)E_k(A_1 + A_2 + \ldots, A_i) - (A_1 + A_2 + \ldots, A_{i-1})E_k(A_1 + A_2 + \ldots, A_{i-1}).$$

Therefore, the value of $P_b(i)$ for $i > 1$, can be obtained as the ratio of the lost traffic of priority $i$ to the offered traffic of priority $i$, that is,

$$P_b(i) = \frac{\left(\sum_{j=1}^{i} A_j\right) E_k\left(\sum_{j=1}^{i} A_j\right) - \left(\sum_{j=1}^{i-1} A_j\right) E_k\left(\sum_{j=1}^{i-1} A_j\right)}{A_i}.$$

**Homework 8.19**

Assume that the traffic offered to a 10 circuit system is composed of two priority traffic: high and low. The arrival rate of the high priority traffic is 5 call/minute and that of the low priority traffic is 4 call/minute. The calls holding time of both traffic classes is exponentially distributed with a mean of 3 minutes. Find the blocking probability of each of the priority classes. □

## 8.14 Overflow Traffic of M/M/$k$/$k$

In many practical situations traffic that cannot be admitted to a $k$ server group overflows to another server group. In such a case the overflow traffic is not Poisson, but it is more bursty than a Poisson process. In other words, the variance of the number of arrivals in an interval is higher than the mean number of arrivals in that interval.

It is therefore important to characterize such overflow traffic by its variance and its mean. In particular, consider an M/M/$k$/$k$ queueing system with input offered traffic $A$ and let $M$ [Erlangs] be the traffic overflowed from this $k$-server system. As discussed in Section 8.1,

$$M = AE_k(A). \tag{342}$$

Let $V$ be the variance of the overflow traffic. Namely, $V$ is the variance of the number of busy servers in an infinite server systems to which the traffic overflowed from our M/M/$k$/$k$ is offered. The $V$ can be obtained by the so-called Riordan Formula as follows:

$$V = M\left(1 - M + \frac{A}{k + 1 + M - A}\right). \tag{343}$$

Note that $M$ and $V$ of the overflow traffic are completely determined by $k$ and $A$.

The variance to mean ratio of a traffic stream is called *Peakedness*. In our case, the peakedness of the overflow traffic is denoted $Z$ and is given by

$$Z = \frac{V}{M},$$

and it normally satisfies $Z > 1$.

## 8.15 Multi-server Loss Systems with Non-Poisson Input

Consider a generalization of an M/M/$k$/$k$ system to the case where the arrival process is not Poisson. As mentioned in the previous section, one way non-Poisson traffic occurs is when we consider a secondary server group to which traffic overflows from a primary server group. If we know the offered traffic to the primary server group (say $A$) and the number of servers in the primary and secondary groups are given by $k_1$ and $k_2$, respectively, then the blocking probability of the secondary server groups is obtained by

$$P_b(\text{secondary}) = \frac{E_{k_1+k_2}(A)}{E_{k_1}(A)}. \tag{344}$$

In this case we also know the mean and variance of the traffic offered to the secondary server group which is readily obtained by Equations (342) and (343).

However, the more challenging problem is the following: given a multi-server loss system with $k_2$ servers loaded by non-Poisson offered traffic with mean $M$ and variance $V$, find the blocking probability. This offered traffic could have come from or overflowed from various sources, and unlike the previous problem, here we do not know anything about the original offered traffic streams or the characteristics of any primary systems. All we know are the values of $M$ and $V$. This problem does not have an exact solution, but reasonable approximations are available. We will now present two approximations:

1. Equivalent Random Method (ERM)
2. Hayward Approximation.

### Equivalent Random Method (ERM)

We wish to estimate the blocking probability for a system with $k_2$ servers loaded by non-pure chance offered traffic with mean $M$ and variance $V$. We know that if this traffic is offered to a system with infinite number of servers (instead of $k_2$), the mean and variance of the number of busy servers will be $M$ and $V$, respectively.

Under the ERM, due to [91], we model the system as if the traffic was the overflow traffic from a primary system with $N_{eq}$ circuits and offered traffic $A_{eq}$ that follows Poisson process. If we find such $A_{eq}$ and $N_{eq}$ then by Eq. (344), the blocking probability in our $k_2$-server system, denoted $PB_{k_2}(N_{eq}, A_{eq})$, will be estimated by:

$$PB_{k_2}(N_{eq}, A_{eq}) = \frac{E_{N_{eq}+k_2}(A_{eq})}{E_{N_{eq}}(A_{eq})}.$$

To approximate $A_{eq}$ and $N_{eq}$, we use the following:

$$A_{eq} = V + 3Z(Z-1); \tag{345}$$

$$N_{eq} = \frac{A_{eq}(M+Z)}{M+Z-1} - M - 1. \tag{346}$$

Note that Equation (345) is an approximation, but Equation (346) is exact and it results in an approximation only because Equation (345) is an approximation.

**Hayward Approximation**

The Hayward approximation [31] is based on the following result. A multi-server system with $k$ servers fed by traffic with mean $M$ and variance $V$ has a similar blocking probability to that of an M/M/$k$/$k$ system with offered load $\frac{M}{Z}$ and $\frac{k}{Z}$ servers, hence Erlang B formula that gives $E_{\frac{k}{Z}}\left(\frac{M}{Z}\right)$ can be used if $\frac{k}{Z}$ is rounded to an integer.

**Homework 8.20**

The XYZ Corporation has measured their offered traffic to a set of 24 circuits and found that it has a mean of $M = 21$ and a variance of $V = 31.5$. Arriving calls cannot be queued and delayed before they are served. This means that if a call arrives and all the circuits (servers) are busy, the call is blocked and cleared from the system.

Use the Hayward Approximation as well as the Equivalent Random Method to estimate the blocking probability.

**Solution**

$M = 21; \quad V = 31.5;$

Peakedness: $Z = \frac{31.5}{21} = 1.5$.

**Hayward Approximation**

The mean offered traffic in the equivalent system = 21 / 1.5 = 14.

The number of servers in the equivalent system = 24/1.5 = 16.

Blocking probability = $E_{16}(14) = 0.1145$.

**Equivalent Random Method**

$$A_{eq} = V + 3Z(Z-1) = 31.5 + 3x1.5x0.5 = 33.75.$$

$$N_{eq} = \frac{A_{eq}(M+Z)}{M+Z-1} - M - 1 = \frac{33.75(21+1.5)}{21+1.5-1} - 21 - 1 = 13.32.$$

We will use $N_{eq} = 13$ conservatively.

From Erlang B:

$E_{13}(33.75) = 0.631$.

$E_{13+24}(33.75) = 0.075$.

Then, the blocking probability is obtained by

$$\frac{E_{13+24}(33.75)}{E_{13}(33.75)} = \frac{0.07536}{0.631} = 0.1194.$$

□

**Homework 8.21**

Assume that non-Poisson traffic with mean = 65 Erlangs and variance = 78 is offered to a loss system. Use both Hayward Approximation and the Equivalent Random Method to estimate the minimal number of circuits required to guarantee that the blocking probability is not more than 1%.

**Solution**

Let us use the notation $N^*$ to represent the minimal number of circuits required to guarantee that the blocking probability is not more than 1%. Previously, we use $k_2$ to represent the *given* number of servers in the secondary system. Now we use the notation $N^*$ to represent the desired number of servers in the system.

Given, $M = 65$ and $V = 78$, the peakedness is given by

$$Z = \frac{78}{65} = 1.2.$$

**Equivalent Random Method**

By (345): $A_{eq} = 78 + 3 \times 1.2 \times 0.2 = 78.72$.

By (346): $N_{eq} = \frac{78.72(65+1.2)}{65+1.2-1} - 65 - 1 = 13.92736 = 14$ approx.

A conservative rounding would be to round it down to $N_{eq} = 13$. This will result in a more conservative dimensioning because lower value for $N_{eq}$ will imply higher loss in the primary system, which leads to higher overflow traffic from the primary to the secondary system. This in turn will require more servers in the secondary system to meet a required blocking probability level.

In the present case, because the result is 13.92736 (so close to 14), we round it up to $N_{eq} = 14$. In any case, we need to be aware of the implication of our choice. We will repeat the calculation using the more conservative choice of $N_{eq} = 13$.

The blocking probability in the primary equivalent system is given by,

$$E_{14}(78.72) = 0.825.$$

Next, find the minimal value for $N^* + 14$ such that

$$\frac{E_{N^*+14}(78.72)}{0.825} \leq 0.01,$$

or,

$$E_{N^*+14}(78.72) \leq 0.00825.$$

By Erlang B formula: $N^* + 14 = 96$, so the number of required servers is estimated by $N^* = 82$.

Notice that the choice of $N^* + 14 = 96$ is made because

$$E_{96}(78.72) = 0.0071$$

satisfies the requirement and

$$E_{95}(78.72) = 0.0087,$$

does not satisfy the requirement.

Now we consider $N_{eq} = 13$.

The blocking probability in the primary equivalent system is given by,

$$E_{13}(78.72) = 0.8373.$$

Next, find minimal value for $N^* + 13$ such that

$$\frac{E_{N^*+13}(78.72)}{0.8373} \leq 0.01,$$

or,

$$E_{N^*+13}(78.72) \leq 0.008373.$$

By Erlang B formula: we choose $N^* + 13 = 96$ because as above $E_{96}(78.72) = 0.0071$ satisfies the requirement and $E_{95}(78.72) = 0.0087$ does not.

Because of our conservative choice of $N_{eq} = 13$, the more conservative choice for the number of required servers is $N^* = 83$.

**Hayward Approximation**

Mean offered traffic in the equivalent system = $65/1.2 = 54.16667$.

By Erlang B formula, the number of required servers in the equivalent system for 1% blocking is 68.

Then, $68 \times 1.2 = 81.6$. Rounding up conservatively, we obtain that 82 servers are required.

Now the designer will need to decide between 82 servers based on the Hayward approximation and based on a reasonable but less conservative approach according to Equivalent Random Method, or 83 according to the very conservative approach based on the Equivalent Random Method.

□

**Homework 8.22**

Consider again the two loss systems the primary and the secondary and use them to compare numerically between:

1. the exact solution;
2. the Hayward approximation;
3. the Equivalent Random Method approximation;
4. a 3rd approximation that is based on the assumption that the arrival process into the secondary system follows a Poisson process. For this approximation assume that the traffic lost in the primary system is offered to the secondary system following a Poisson process.

**Guide**

For the comparison, at first assume that you know $A$, $k_1$ and $k_2$ and compute $M$ and $V$, i.e., the mean and variance of the offered load to the secondary system as well as the blocking probability of traffic in the secondary system using (344).

Next, assume that $A$ and $k_1$ are not known but $k_2$ is known; also known are $M$ and $V$, i.e., the mean and variance of the offered load to the secondary system that you computed previously. And evaluate the blocking probability using both Hayward, the Equivalent Random Method and the Poisson approximations.

Compare the results for a wide range of parameters.

□

# 9 M/M/$k$

The M/M/$k$ queue is a generalization of the M/M/1 queue to the case of $k$ servers. As in M/M/1, for an M/M/$k$ queue, the buffer is infinite and the arrival process is Poisson with rate $\lambda$. The service time of each of the $k$ servers is exponentially distributed with parameter $\mu$. As in the case of M/M/1 we assume that the service times are independent and are independent of the arrival process.

## 9.1 Steady-State Equations and Their Solution

Letting $A = \lambda/\mu$, and assuming the stability condition $\lambda < k\mu$, or $A < k$, the M/M/$k$ queue gives rise to the following steady-state equations:

$\pi_1 = A\pi_0$
$\pi_2 = A\pi_1/2 = A^2\pi_0/2$
$\pi_3 = A\pi_2/3 = A^3\pi_0/(3!)$
...
$\pi_k = A\pi_{k-1}/k = A^k\pi_0/(k!)$
$\pi_{k+1} = A\pi_k/k = A^{k+1}\pi_0/(k!k)$
$\pi_{k+2} = A\pi_{k+1}/k = A^{k+2}\pi_0/(k!k^2)$
...
$\pi_{k+j} = A\pi_{k+j-1}/k = A^{k+j}\pi_0/(k!k^j)$ for $j = 1,\ 2,\ 3,\ \ldots$

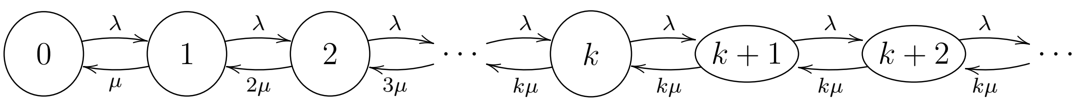
and in general:

$$\pi_n = \frac{A^n\pi_0}{n!} \text{ for } n = 0,\ 1,\ 2,\ \ldots,\ k-1 \tag{347}$$

and

$$\pi_n = \frac{A^n\pi_0}{k!k^{n-k}} \text{ for } n = k,\ k+1,\ k+2,\ \ldots\ . \tag{348}$$

These balance equations can also be described by the following state transition diagram of M/M/$k$:




To obtain $\pi_0$, we sum up both sides of Eqs. (347) and (348), and because the sum of the $\pi_n$s equals one, we obtain an equation for $\pi_0$, which its solution is

$$\pi_0 = \left(\sum_{n=0}^{k-1}\frac{A^n}{n!} + \frac{A^k}{k!}\frac{k}{(k-A)}\right)^{-1}. \tag{349}$$

Substituting the latter in Eqs. (347) and (348), we obtain the steady-state probabilities $\pi_n,\ n = 0,\ 1,\ 2,\ \ldots\ .$

## 9.2 Erlang C Formula

Of special interest is the so called Erlang C formula. It represents the proportion of time that all $k$ servers are busy and is given by:

$$C_k(A) = \sum_{n=k}^{\infty} \pi_n = \frac{A^k}{k!} \frac{k}{(k-A)} \pi_0 = \frac{\frac{A^k}{k!} \frac{k}{(k-A)}}{\sum_{n=0}^{k-1} \frac{A^n}{n!} + \frac{A^k}{k!} \frac{k}{(k-A)}}. \tag{350}$$

**Homework 9.1**

Derive Eq. (350). □

By Eqs. (320) and (350) we obtain the following relationship:

$$C_k(A) = \frac{kE_k(A)}{k - A[1 - E_k(A)]}. \tag{351}$$

**Homework 9.2**

1. Derive Eq. (351);
2. Show that $C_k(A) \geq E_k(A)$. □

An elegant result for $C_k(A)$ is the following

$$\frac{1}{C_k(A)} = \frac{1}{E_k(A)} - \frac{1}{E_{k-1}(A)}. \tag{352}$$

**Homework 9.3**

Prove Eq. (352).

In the following table, we add the corresponding $C_k(A)$ values to the table of the previous section. We can observe the significant difference between $C_k(A)$ and $E_k(A)$ as the ratio $A/k$ increases. Clearly, when $A/k > 1$, the M/M/$k$ queue is unstable.

| $A$ | $k$ | $E_k(A)$ | $C_k(A)$ |
|---|---|---|---|
| 20 | 30 | 0.0085 | 0.025 |
| 100 | 117 | 0.0098 | 0.064 |
| 500 | 527 | 0.0095 | 0.158 |
| 1000 | 1029 | 0.0099 | 0.262 |
| 5000 | 5010 | 0.0100 | 0.835 |
| 10000 | 9970 | 0.0099 | unstable |

**Homework 9.4**

Reproduce the results of the above table. □

## 9.3 Mean Queue Size, Delay, Waiting Time and Delay Factor

Let us reuse the following notation:
$Q$ = a random variable representing the total number of customers in the system (waiting in the queue and being served);
$N_Q$ = a random variable representing the total number of customers waiting in the queue (this does not include those customers being served);
$N_s$ = a random variable representing the total number of customers that are being served;
$D$ = a random variable representing the total delay in the system (this includes the time a customer waits in the queue and in service);
$W_Q$ = a random variable representing the time a customer waits in the queue (this excludes the time a customer spends in service);
$S$ = a random variable representing the service time.
$\hat{D}$ = The delay of a delayed customer including the service time
$\hat{W}_Q$ = The delay of a delayed customer in the queue excluding the service time.

Using the above notation, we have

$$E[Q] = E[N_Q] + E[N_s] \tag{353}$$

and

$$E[D] = E[W_Q] + E[S]. \tag{354}$$

Clearly,

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

To obtain $E[N_s]$ for the M/M/$k$ queue, we use Little's formula for the system made of servers. If we consider the system of servers (without considering the waiting room outside the servers), we notice that since there are no losses, the arrival rate into this system is $\lambda$ and the mean waiting time of each customer in this system is $E[S] = 1/\mu$. Therefore, by Little's formula the mean number of busy servers is given by

$$E[N_s] = \frac{\lambda}{\mu} = A. \tag{355}$$

To obtain $E[N_Q]$, we consider two mutually exclusive and exhaustive events: $\{Q \geq k\}$, and $\{Q < k\}$. Recalling the Law of Iterated Expectation (95), we have

$$E[N_Q] = E[N_Q \mid Q \geq k]P(Q \geq k) + E[N_Q \mid Q < k]P(Q < k). \tag{356}$$

To derive $E[N_Q \mid Q \geq k]$, we notice that the evolution of the M/M/$k$ queue during the time when $Q \geq k$ is equivalent to that of an M/M/1 queue with arrival rate $\lambda$ and service rate $k\mu$. The mean queue size of such M/M/1 queue is equal to $\rho/(1-\rho)$ where $\rho = \lambda/(k\mu) = A/k$. Thus,

$$E[N_Q \mid Q \geq k] = \frac{A/k}{1 - A/k} = \frac{A}{k - A}.$$

Therefore, since $E[N_Q \mid Q < k] = 0$ and $P(Q \geq k) = C_k(A)$, we obtain by (356) that

$$E[N_Q] = C_k(A)\frac{A}{k-A}. \tag{357}$$

**Homework 9.5**

Derive Eq. (357) by a direct approach using $E[N_Q] = \sum_{n=k}^{\infty}(n-k)\pi_n$.

**Guide**

By (348),

$$E[N_Q] = \sum_{n=k}^{\infty}(n-k)\pi_n = \sum_{n=k}^{\infty}(n-k)\frac{A^n\pi_0}{k!k^{n-k}}$$

Set $i = n - k$, to obtain

$$E[N_Q] = \sum_{i=0}^{\infty} i\frac{A^{i+k}\pi_0}{k!k^i} = \frac{\pi_0 A^k}{k!}\sum_{i=0}^{\infty} i\left(\frac{A}{k}\right)^i = C_k(A)\frac{k-A}{k}\frac{A/k}{(1-A/k)^2},$$

and (357) follows. □

**Homework 9.6**

Confirm consistence between (357) and (283). □

By (353), (355) and (357), we obtain

$$E[Q] = C_k(A)\frac{A}{k-A} + A. \tag{358}$$

Therefore, by Little's formula

$$E[W_Q] = \frac{C_k(A)\frac{A}{k-A}}{\lambda} = \frac{C_k(A)}{\mu k - \lambda}. \tag{359}$$

Notice the physical meaning of $E[W_Q]$. It is the ratio between the probability of having all servers busy and the spare capacity of the system.

The mean delay is readily obtained by adding the mean service time to $E[W_Q]$. Thus,

$$E[D] = \frac{C_k(A)}{\mu k - \lambda} + \frac{1}{\mu}. \tag{360}$$

Another useful measure is the so-called *delay factor* [19]. It is defined as the ratio of the mean waiting time in the queue to the mean service time. Namely, it is given by

$$D_F = \frac{E[W_Q]}{1/\mu} = \frac{\frac{C_k(A)}{\mu k-\lambda}}{\frac{1}{\mu}} = \frac{C_k(A)}{k-A}. \tag{361}$$

The rationale to use delay factor is that in some applications users that require long service time may be willing to wait longer time in the queue in direct proportion to the service time.

## 9.4 Mean Delay of Delayed Customers

In Section 6.4, we have shown how to derive, for the case of M/M/1, $E[\hat{D}]$ and $E[\hat{W}_Q]$, namely, mean delay of a delayed customer including the service time and excluding the service time, respectively. We now extend the same ideas to the case of M/M/$k$. As in the previous case, to obtain $E[\hat{W}_Q]$, we use Little's formula where we consider the queue (without the servers) as the system and the arrival rate of the delayed customers which in the present case is $\lambda C_k(A)$.

Therefore,

$$E[\hat{W}_Q] = \frac{AC_k(A)}{\lambda C_k(A)(k-A)} = \frac{1}{k\mu - \lambda}.$$

and

$$E[\hat{D}] = E[\hat{W}_Q] + \frac{1}{\mu} = \frac{1}{k\mu - \lambda} + \frac{1}{\mu}.$$

As in Section 6.4, we can check the latter using the Law of Iterated Expectation as follows:

$$\begin{aligned} E[D] &= (1 - C_k(A))E[S] + C_k(A)E[\hat{D}] \\ &= (1 - C_k(A))\frac{1}{\mu} + C_k(A)\left(\frac{1}{k\mu - \lambda} + \frac{1}{\mu}\right) = \frac{C_k(A)}{\mu k - \lambda} + \frac{1}{\mu}, \end{aligned}$$

and again we observe that consistency is achieved and note that this consistency check is an alternative way to obtain $E[\hat{D}]$.

## 9.5 Dimensioning

One could solve the dimensioning problem of finding, for a given $A$, the smallest $k$ such that $C_k(A)$ or the mean delay is lower than a given value. Using Eq. (351), and realizing that the value of $C_k(A)$ decreases as $k$ increases, the dimensioning problem with respect to $C_k(A)$ can be solved in an analogous way to the M/M/$k$/$k$ dimensioning problem. Having the $C_k(A)$ values for a range of $k$ value one can also obtain the minimal $k$ such that the mean delay is bounded using Eq. (360). A similar procedure can be used to find the minimal $k$ such that delay factor requirement is met.

## 9.6 Utilization

The utilization of an M/M/$k$ queue is the ratio of the mean number of busy servers to $k$, therefore the utilization of an M/M/$k$ queue is obtained by

$$\hat{U} = \frac{E[N_s]}{k} = \frac{A}{k}. \tag{362}$$

**Homework 9.7**

Write a computer program that computes the minimal $k$, denoted $k^*$, subject to a bound on $E[D]$. Run the program for a wide range of parameter values and plot the results. Try to consider meaningful relationships, e.g., plot the spare capacity $k^*\mu - \lambda$ and utilization as a function of various parameters and discuss implications. □

**Homework 9.8**

Consider the M/M/2 queue with arrival rate $\lambda$ and service rate $\mu$ of each server.

1. Show that
$$\pi_0 = \frac{2-A}{2+A}.$$
2. Derive formulae for $\pi_i$ for $i = 1, 2, 3, \ldots$ .
3. Show that
$$C_2(A) = \frac{A^2}{2+A}.$$
Note that for $k = 2$, it is convenient to use $C_2(A) = 1 - \pi_0 - \pi_1$.
4. Derive a formula for $E[N_s]$ using the sum: $\pi_1 + 2C_2(A)$ and show that
$$E[N_s] = \pi_1 + 2C_2(A) = A.$$
5. Derive $E[Q]$ in two ways, one using the sum $\sum_{i=0}^{\infty} i\pi_i$ and the other using Eqs. (356) – (358), and show that in both ways you obtain
$$E[Q] = \frac{4A}{4-A^2}.$$

□

**Homework 9.9**

Queueing theory is a useful tool for decisions on hospital resource allocation [18, 35, 36, 88]. In particular, the M/M/$k$ model has been considered [35, 36]. Consider the following example from [88]. Assume that a patient stays at an Intensive Care Unit (ICU) for an exponentially distributed period of time with an average time of 2.5 days. Consider two hospitals. Patients arrivals at each of the hospitals follow a Poisson process. They arrive at Hospital 1 at the rate of one patient per day and the arrival rate at Hospital 2 is 2 patients per day. Assume that Hospital 2 has 10 ICU beds. Then the management of Hospital 1 that has never studied queueing theory believes that they need only 5 beds, because they think that if they have half the traffic load they need half the number of beds. Your job is to evaluate and criticize their decision. Assuming an M/M/$k$ model, calculate the mean delay and the probability of having all servers busy for each of the two systems. Which one performs better? If you set the probability of having all servers busy in Hospital 2 as the desired QoS standard, how many beds Hospital 1 will need? Maintaining the same QoS standard, provide a table with the number of beds needed in Hospital 1 if it has traffic arrival rates $\lambda_1 = 4, 8, 16, 32, 64$ patients per day. For each of the $\lambda_1$ values, provide a simple rule to estimate the number of beds $n$ in Hospital 1, maintaining the same QoS standard. Provide rigorous arguments to justify this rule for large values of $\lambda_1$ and $n$.

**Hint:** Observe that as $n$ grows with $\lambda_1$, $n - \lambda_1$, approaches $C\sqrt{n}$ for some constant $C$ (find that constant!). For rigorous arguments, study [39]. □

# 10 Engset Loss Formula

The Engset loss formula applies to telephony situations where the number of customers is not too large relative to the number of available circuits. Such situations include: an exchange in a small rural community, PABX, or a lucrative satellite service to a small number of customers. Let the call holding times be IID exponentially distributed with mean $1/\mu$ and the time until an idle source attempts to make a call is also exponentially distributed but with mean $1/\hat{\lambda}$. We also assume that there is no dependence between the holding times and the idle periods of the sources. Let the number of customers (sources of traffic) be $M$, the number of circuits $k$ and the blocking probability $P_b$.

This gives rise to a finite source loss model as oppose to the M/M/$k$/$k$ model where the arrival process follows a Poisson process which can describe traffic generated by an infinite number of sources (customers). In this book we only consider finite source loss models that do not involve queueing. For finite source models that involve queueing, the reader is referred to [86].

The reader will recall that in M/M/1, the arrival rate as well as the service rate are independent of the state of the system, and in M/M/$\infty$, the arrival rate is also independent of the number of customers in the system, but the service rate is state dependent. In the present case, when the number of customers is limited, we have a case where both the arrival rate and the service rate are state dependent.

As in M/M/$k$/$k$, in our Engset finite source loss model, the service rate is $n\mu$ when there are $n$ busy circuits (namely $n$ customers are making phone calls). However, unlike M/M/$k$/$k$, where the arrival rate is state independent, in the present case, the arrival rate depends on the number of the customers that are being served. Under the Engset finite source model, busy customers do not make new phone calls thus they do not contribute to the arrival rate. Therefore, if $n$ circuits are busy, the arrival rate is $(M-n)\hat{\lambda}$. As a result, considering both arrival and service processes, at any point in time, given that there are $n$ customers in the system, and at the same time, $n$ servers/circuits are busy, the time until the next event is exponentially distributed with parameter $(M-n)\hat{\lambda}+n\mu$, because it is a competition between $M$ exponential random variables: $n$ with parameter $\mu$ and $M-n$ with parameter $\hat{\lambda}$.

An important question we must always answer in any Markov-chain analysis is how many states do we have. If $M > k$, then the number of states is $k+1$, as in M/M/$k$/$k$. However, if $M < k$, the number of states is $M+1$ because no more than $M$ calls can be in progress at the same time. Therefore, the number if states is $\min\{M,k\}+1$.

## 10.1 Steady-State Equations and Their Solution

Considering a finite state birth-and-death process that represents the queue evolution of the above described queueing system with $M$ customers (sources) and $K$ servers, we obtain the following steady-state equations:

$\pi_0 M\hat{\lambda} = \pi_1\mu$

$\pi_1(M-1)\hat{\lambda} = \pi_2 2\mu$

$\pi_2(M-2)\hat{\lambda} = \pi_3 3\mu$

...

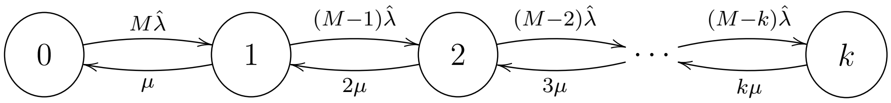
and in general:

$$\pi_n(M-n)\hat{\lambda} = \pi_{n+1}(n+1)\mu, \text{ for } n = 0,\ 1,\ 2,\ \ldots, \min\{M,k\} - 1. \tag{363}$$

Therefore, after standard algebraic manipulations of (363), we can write $\pi_n$, for $n = 0,\ 1,\ 2,\ \ldots, \min\{M,k\}$, in terms of $\pi_0$ as follows:

$$\pi_n = \binom{M}{n}\left(\frac{\hat{\lambda}}{\mu}\right)^n \pi_0, \text{ for } n = 0,\ 1,\ 2,\ \ldots, \min\{M,k\}, \tag{364}$$

These local balance steady-state equations are described by the following state transition diagram for the case $M > k$:




Using the notation $\hat{\rho} = \hat{\lambda}/\mu$, we obtain

$$\pi_n = \binom{M}{n}\hat{\rho}^n \pi_0, \text{ for } n = 0,\ 1,\ 2,\ \ldots, \min\{M,k\}. \tag{365}$$

**Homework 10.1**

Derive Eqs. (364) and (365). □

Of course, the sum of the steady-state probabilities must be equal to one, so we again have the additional normalizing equation

$$\sum_{j=0}^{\min\{M,k\}} \pi_j = 1. \tag{366}$$

By (365) together with the normalizing Eq. (366), we obtain

$$\pi_0 = \frac{1}{\sum_{j=0}^{\min\{M,k\}} \binom{M}{j}\hat{\rho}^j}.$$

Therefore, by (365), we obtain

$$\pi_n = \frac{\binom{M}{n}\hat{\rho}^n}{\sum_{j=0}^{\min\{M,k\}} \binom{M}{j}\hat{\rho}^j}, \text{ for } n = 0,\ 1,\ 2,\ \ldots, \min\{M,k\}. \tag{367}$$

## 10.2 Blocking Probability

Now, what is the blocking probability $P_b$? When $k \geq M$, clearly $P_b = 0$, as there is never a shortage of circuits.

To derive the blocking probability for the case when $k < M$, we first realize that unlike in the case of Erlang Formula, $\pi_k$ does not give the blocking probability. Still, $\pi_k$ is the probability of having $k$ busy circuits, or the proportion of time that all circuits are busy which is the so-called *time-congestion*, but it is not the probability that a call is blocked – the so-called *call-congestion*. Unlike the case of Erlang B Formula, here, call-congestion is not equal to time congestion. This is because in the Engset model, the arrival process does not follow a Poisson process. In fact, the arrival rate is dependent on the state of the system. When the system is full the arrival rate is lower, and could be much lower, than when the system is empty.

In particular, when $i$ circuits are busy, the arrival rate is $\hat{\lambda}(M-i)$, therefore to find the proportion of calls blocked, or the blocking probability denoted $P_b$, we compute the ratio between calls arrive when there are $k$ circuits busy and the total calls arrive. This gives

$$P_b = \frac{\hat{\lambda}(M-k)\pi_k}{\hat{\lambda}\sum_{i=0}^{k}(M-i)\pi_i}. \tag{368}$$

Substituting (364) and (365) in (368) and performing few algebraic manipulations, we obtain the Engset loss formula that gives the blocking probability for the case $M > k$ as follows.

$$P_b = \frac{\binom{M-1}{k}\hat{\rho}^k}{\sum_{i=0}^{k}\binom{M-1}{i}\hat{\rho}^i}. \tag{369}$$

Notice that $\hat{\rho}$, defined above by $\hat{\rho} = \hat{\lambda}/\mu$, is the traffic generated by a **free** customer. An interesting interpretation of (369) is that the call congestion, or the blocking probability, when there are $M$ sources is equal to the time congestion when there are $M-1$ sources. This can be intuitively explained as follows. Consider an arbitrary tagged source (or customer). For this particular customer, the proportion of time it cannot access is equal to the proportion of time the $k$ circuits are all busy by the other $M-1$ customers. During the rest of the time our tagged source can successfully access a circuit.

**Homework 10.2**

Perform the derivations that lead to Eq. (369). □

## 10.3 Obtaining the Blocking Probability by a Recursion

Letting $B_i$ be the blocking probability given that the number of circuits (servers) is $i$, the Engset loss formula can be solved numerically by the following recursion:

$$B_i = \frac{\hat{\rho}(M-i)B_{i-1}}{i + \hat{\rho}(M-i)B_{i-1}} \qquad i = 1, 2, 3, \ \ldots, \ k \tag{370}$$

with the initial condition

$$B_0 = 1. \tag{371}$$

**Homework 10.3**

Derive Eqs. (370) and (371).

**Guide**

By (369) and the definition of $B_i$ we have

$$B_i = \frac{\binom{M-1}{i}\hat{\rho}^i}{\sum_{j=0}^{i}\binom{M-1}{j}\hat{\rho}^j}$$

and

$$B_{i-1} = \frac{\binom{M-1}{i-1}\hat{\rho}^{i-1}}{\sum_{j=0}^{i-1}\binom{M-1}{j}\hat{\rho}^j}.$$

Consider the ratio $B_i/B_{i-1}$ and after some algebraic manipulations (that are somewhat similar to the derivations of the Erlang B recursion) you will obtain

$$\frac{B_i}{B_{i-1}} = \frac{\rho(M-i)}{i}(1-B_i)$$

which leads to (370). Notice that $B_0 = 1$ is equivalent to the statement that if there are no circuits (servers) (and $M > 0, \hat{\rho} > 0$) the blocking probability is equal to one. □

## 10.4 Insensitivity

In his original work [23], Engset assumed that the idle time as well as the holding time are exponentially distributed. These assumptions have been relaxed in [20] and now it is known that Engset formula applies also to arbitrary idle and holding time distributions (see also [44]).

## 10.5 Load Classifications and Definitions

An important feature of Engset setting is that a customer already engaged in a conversation does not originate calls. This leads to an interesting peculiarity that if we fix the number of customers (assuming $M > k$) and reduce $k$, the offered traffic increases because reduction in $k$ leads to increase in $P_b$ and reduction in the average number of busy customers which in turn leads to increase in idle customers each of which offer more calls, so the offered load increases.

Let us now discuss the concept of the so-called *intended* offered load [6] under the Engset setting. We know that $1/\hat{\lambda}$ is the mean time until a free customer makes a call (will attempt to seize a circuit). Also, $1/\mu$ is the mean holding time of a call. If a customer is never blocked, it is behaving like an on/off source, alternating between on and off states, being on for an exponentially distributed period of time with mean $1/\mu$, and being off for an exponentially distributed period of time with mean $1/\hat{\lambda}$. For each cycle of average length $1/\hat{\lambda}+1/\mu$, a source will be busy, on average, for a period of $1/\mu$. Therefore, in steady-state, the proportion of time a source is busy is $\hat{\lambda}/(\hat{\lambda}+\mu)$, and since we have $M$ sources, the *intended* offered load is given by

$$T = M\frac{\hat{\lambda}}{\hat{\lambda}+\mu} = \frac{\hat{\rho}M}{(1+\hat{\rho})}. \tag{372}$$

This *intended* offered load is equal to the offered traffic load and the carried traffic load if $M \leq k$, namely, when $P_b = 0$. However, when $M > k$ (thus $P_b > 0$), the offered traffic load and the carried traffic load are not equal. Let $T_c$ and $T_o$ be the *carried* and the *offered* traffic load respectively. The carried traffic is the mean number of busy circuits and it is given by

$$T_c = \sum_{i=0}^{k} i\pi_i. \tag{373}$$

The offered traffic is obtained by averaging the intensities of the free customers weighted by the corresponding probabilities of their numbers, as follows.

$$T_o = \sum_{i=0}^{k} \hat{\rho}(M-i)\pi_i. \tag{374}$$

To compute the values for $T_c$ and $T_o$ in terms of the blocking probability $P_b$, we first realize that

$$T_c = T_o(1-P_b), \tag{375}$$

and also,

$$T_o = \sum_{i=0}^{k} \hat{\rho}(M-i)\pi_i = \hat{\rho}M - \hat{\rho}\sum_{i=0}^{k} i\pi_i = \hat{\rho}(M-T_c) \tag{376}$$

and by (375) – (376) we obtain

$$T_c = \frac{\hat{\rho}M(1-P_b)}{[1+\hat{\rho}(1-P_b)]} \tag{377}$$

and

$$T_o = \frac{\hat{\rho}M}{[1+\hat{\rho}(1-P_b)]}. \tag{378}$$

Notice that when $P_b = 0$, we have

$$T_o = T = T_c, \tag{379}$$

and when $P_b > 0$, we obtain by (372), (377) and (378) that

$$T_o > T > T_c. \tag{380}$$

**Homework 10.4**

Using (372), (377) and (378), show (379) and (380). □

Notice also that the above three measures may be divided by $k$ to obtain the relevant traffic load per server.

## 10.6 The Small Idle Period Limit

Let $\hat{\lambda}$ approach infinity, while $M$, $k$ and $\mu$ stay fixed and assume $M > k$. Considering the steady-state equations (363), their solution at the limit is $\pi_i = 0$ for $i = 0,\ 1,\ 2, \ldots, k-1$ and $\pi_k = 1$. To see that consider the equation

$$\pi_0 M \hat{\lambda} = \pi_1 \mu.$$

Assume $\pi_0 > 0$, so as $\hat{\lambda} \rightarrow \infty$, we must have $\pi_1 > 1$ which leads to contradiction; thus, $\pi_0 = 0$, repeating the same argument for the steady-state equations (363) for $n = 1,\ 2, \ldots, k-1$, we obtain that $\pi_i = 0$ for $i = 0,\ 1,\ 2, \ldots, k-1$. Then because $\sum_{i=0}^{k} \pi_i = 1$, we must have $\pi_k = 1$. Therefore by (373),

$$T_c = k.$$

and by (374),

$$T_o = \hat{\rho}(M-k) \rightarrow \infty.$$

Intuitively, this implies that as $k$ channels (circuits) are constantly busy serving $k$ customers, the remaining $M-k$ sources (customers) reattempt to make calls at infinite rate. In this case, by (372), the intended traffic load is

$$T = \frac{\hat{\rho}M}{(1+\hat{\rho})} \rightarrow M.$$

Note also that under this condition, by (375), noticing that $T_o \rightarrow \infty$, and that $T_c = k$, we obtain that the blocking probability is approaches 1.

## 10.7 The Many Sources Limit

Let $M$ approach infinity and $\hat{\lambda}$ approach zero in a way that maintains the intended offered load constant. In this case, since $\hat{\lambda} + \mu \rightarrow \mu$, the limit of the intended load will take the form

$$\lim T = M\frac{\hat{\lambda}}{\mu} = \hat{\rho}M. \tag{381}$$

Furthermore, under this limiting condition, the terms $\hat{\rho}(M-i)$, $i = 1, 2, 3, \ldots, k$, in (370) can be substituted by $\hat{\rho}M$ which is the limit of the intended traffic load. It is interesting to observe that if we substitute $A = \hat{\rho}M$ for the $\hat{\rho}(M-i)$ terms in (370), equations (322) and (370) are equivalent. This means that if the number of sources increases and the arrival rate of each source decreases in a way that the intended load stays fixed, the blocking probability obtained by Engset loss formula approaches that of Erlang B formula.

## 10.8 Obtaining the Blocking Probability by Successive Iterations

In many cases, $\hat{\rho}$ is not available and instead the offered load $T_o$ is available. Then it is convenient to obtain the blocking probability $P_b$ in terms of $T_o$. By Eq. (378) we obtain,

$$\hat{\rho} = \frac{T_o}{M - T_o(1 - P_b)}. \tag{382}$$

The latter can be used together with Eq. (369) or (370) to obtain $P_b$ by an iterative process. One begin by setting an initial estimate value to $P_b$ (e.g. $P_b = 0.1$). Then this initial estimate is substituted into Eq. (382) to obtain an estimate for $\hat{\rho}$ then the value you obtain for $\hat{\rho}$ is substituted in Eq. (369), or use the recursion (370), to obtain another value for $P_b$ which is then substituted in Eq. (382) to obtain another estimate for $\hat{\rho}$. This iterative process continues until the difference between two successive estimations of $P_b$ is arbitrarily small.

**Homework 10.5**

Consider the case $M = 20$, $k = 10$, $\hat{\lambda} = 2$, $\mu = 1$. Compute $P_b$ using the recursion of Eq. (370). Then compute $T_o$ and assuming $\rho$ is unknown, compute $P_b$ using the iterative processes starting with various initial estimations. Compare the results and the running time of your program. □

# 11 State Dependent SSQ

In the queueing model discussed in the previous chapter, the arrival and service rates vary based on the state of the system. In this section we consider a general model of a Markovian queue where the arrival and service rates depend on the number of customers in the system. Having this general model, we can apply it to many systems whereby capacity is added (service rate increases) and/or traffic is throttled back as queue size increases.

In particular, we study a model of a single-server queue in which the arrival process is a state dependent Poisson process. This is a Poisson process that its rate $\lambda_i$ fluctuates based on the queue size $i$. The service rate $\mu_i$ also fluctuates based on $i$. That is, when there are $i$ customers in the system, the service is exponentially distributed with parameter $\mu_i$. If during service, before the service is complete, the number of customers changes from $i$ to $j$ ($j$ could be either $i+1$ or $i-1$) then the remaining service time changes to exponentially distributed with parameter $\mu_j$. We assume that the number of customers in the queue is limited by $k$.

This model gives rise to a birth-and-death process described in Section 2.5. The state dependent arrival and service rates $\lambda_i$ and $\mu_i$ are equivalent to the birth-and-death rates $a_i$ and $b_i$, respectively.

Following the birth-and-death model of Section 2.5 the infinitesimal generator for our Markovian state dependent queue-size process is given by

$Q_{i,i+1} = \lambda_i$ for $i = 0, 1, 2, 3, \ldots, k$
$Q_{i,i-1} = \mu$ for $i = 1, 2, 3, 4, \ldots, k$
$Q_{0,0} = -\lambda_0$
$Q_{i,i} = -\lambda_i - \mu_i$ for $i = 1, 2, 3, \ldots, k-1$
$Q_{k,k} = -\mu_k$.

Then the steady-state equations $0 = \mathbf{\Pi Q}$, can be written as:

$$0 = -\pi_0\lambda_0 + \pi_1\mu_1 \tag{383}$$

and

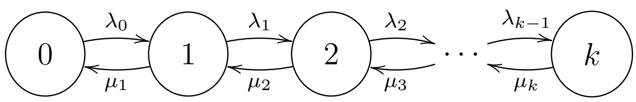
$$0 = \pi_{i-1}\lambda_{i-1} - \pi_i(\lambda_i + \mu_i) + \pi_{i+1}\mu_{i+1} \quad \text{for} \quad i = 1, 2, 3, \ldots, k-1. \tag{384}$$

There is an additional last dependent equation

$$0 = \pi_{k-1}\lambda_{k-1} - \pi_k(\mu_k) \tag{385}$$

which is redundant.

These balance equations can also be described by the following state transition diagram:




The normalizing equation

$$\sum_{i=0}^{k} \pi_i = 1 \tag{386}$$

must also be satisfied.

Notice that the equation

$$0 = -\pi_0\lambda_0 + \pi_1\mu_1$$

and the first equation of the set (384), namely,

$$0 = \pi_0\lambda_0 - \pi_1(\lambda_1 + \mu_1) + \pi_2\mu_2$$

gives

$$0 = -\pi_1\lambda_1 + \pi_2\mu_2$$

which together with the second equation of the set (384), namely,

$$0 = \pi_1\lambda_1 - \pi_2(\lambda_2 + \mu_2) + \pi_3\mu_3$$

gives

$$0 = -\pi_2\lambda_2 + \pi_3\mu_3$$

and in general, we obtain the set of $k$ equations:

$$0 = -\pi_{i-1}\lambda_{i-1} + \pi_i\mu_i \qquad i = 1, 2, 3, \ldots, k$$

or the recursive equations:

$$\pi_i = \rho_i\pi_{i-1} \text{ for } i = 1, 2, 3, \ldots, k \tag{387}$$

where

$$\rho_i = \frac{\lambda_{i-1}}{\mu_i} \text{ for } i = 1, 2, 3, \ldots k.$$

Defining also $\rho_0 \equiv 1$, by (387), we obtain

$$\pi_i = \rho_i\rho_{i-1}\rho_{i-2}\ldots\rho_1\pi_0 \text{ for } i = 0, 1, 2, 3, \ldots, k. \tag{388}$$

**Homework 11.1**

Drive $\pi_i$ for $i = 0, 1, 2 \ldots k$.

**Guide**

Summing up equations (387) will give an equation with $1 - \pi_0$ on its left-hand side and a constant times $\pi_0$ on its right-hand side. This linear equation for $\pi_0$ can be readily solved for $\pi_0$. Having $\pi_0$, all the other $\pi_i$ can be obtained by (387). □

Having obtained the $\pi_i$ values, let us derive the blocking probability. As in the case of M/M/$k$/$k$, the proportion of time that the buffer is full is given by $\pi_k$. However, the proportion of time that the buffer is full is not the blocking probability. This can be easily see in the case $\lambda_k = 0$. In this case, no packets arrive when the buffer is full, so no losses occur, but we may still have $\rho_i > 0$ for $i = 1, 2, 3, \ldots, k$, so $\pi_k > 0$.

As in the case of the Engset model, the blocking probability is ratio of the number of arrivals during the time that the buffer is full to the total number of arrivals. Therefore,

$$P_b = \frac{\lambda_k\pi_k}{\sum_{i=0}^{k}\lambda_i\pi_i}. \tag{389}$$

Notice that, as in the Engset model, the PASTA principle does not apply here since the arrival process is not Poisson. However, if the arrival rates do not depend on the state of the system, even if the service rates do, the arrival process becomes Poisson and the blocking probability is equal to $\pi_k$. To see this simply set $\lambda_i = \lambda$ for all $i$ in Eq. (389) and we obtain $P_b = \pi_k$.

**Homework 11.2**

Consider a single-server Markovian queue with state dependent arrivals and service. You are free to choose the $\lambda_i$ and $\mu_i$ rates, but make sure they are different for different $i$ values. Set the buffer size at $k = 200$. Solve the steady-state equations using the successive relaxation method and using a standard method. Compare the results and the computation time. Then obtain the blocking probability by simulation and compare with the equivalent results obtained by solving the state equations. Repeat the results for a wide range of parameters by using various $\lambda_i$ vectors. $\square$

# 12 Queueing Models with Finite Buffers

We have encounters already several examples of queueing systems where the number of customers/packets in the system in limited. Examples include the M/M/$k$/$k$ system, the Engset system and the state-dependent SSQ described in the previous chapter. Given that in real life all queueing systems have a limited capacity, it is important to understand the performance behavior of such queues. A distinctive characteristic of a queue with finite capacity is the possibility of a blocking event. In practice, blocking probability evaluation is an important performance measure of a queueing system. For example, depending of the type of service and protocol used, packets lost in the Internet due to buffer overflow are either retransmitted which increases delay, or never arrive at their destination which may adversely affect QoS perceived by users.

In a loss system such as M/M/$k$/$k$ an arriving customer may either experience blocking, or is immediately admitted to service. Then, in a delay system such the M/M/1 queue, an arriving customer may either experience delay, or is immediately admitted to service. By comparison, in a system with a finite queue discussed here, where the total number of buffer places $N$ is larger than the number of servers $k$ (which is equal to the number of buffer places at the service system), an arriving customer may encounter either one of the following three possible events:

1. It may be admitted to service immediately.
2. It may be place in the queue until a server is available.
3. It may be blocked because all servers are busy and all buffer places are occupied.

We begin the chapter by considering two extreme SSQ systems with finite buffer. The first is a D/D/1/$N$ system where the blocking probability is equal to zero as long as the arrival rate is not higher than the service rate and the second one is a model where a single large burst (SLB) arrives at time zero. We call it an SLB/D/1/$N$ queue. In such a queue, for an arbitrarily small arrival rate, the blocking probability approaches unity. These two extreme examples signifies the importance of using the right traffic model, otherwise the blocking probability estimation can be very inaccurate. These two extreme cases will be followed by four other cases of Markovian queues with finite buffers: the M/M/1/$N$, the M/M/$k$/$N$ for $N > k$, the MMPP(2)/M/1/$N$ and the M/$E_m$/1/$N$ Queues.

## 12.1 D/D/1/$N$

As in our discussion on deterministic queue, we assume that if an arrival and a departure occur at the same point in time, the departure occurs before the arrival. For the case of $\rho = \lambda/\mu < 1$, the evolution of the D/D/1/$N$, $N \geq 1$ is the same as that of a D/D/1 queue. In such a case, there is never more than one packet in the system, thus no losses occur and an arriving customer always finds the server available to serve it. That is, out of the three possible events mentioned above, only the first event occurs in this particular case. Let us now consider the case $\rho = \lambda/\mu > 1$. In this case, the queue reaches a persistent congestion state where the queue size fluctuates between $N$ and $N - 1$. The case $N = 1$ was already considered in previous discussions, so we assume $N > 1$. In this case, whenever a packet completes its service, there is always another packet queued which enters service immediately after the previous one left the system. Therefore, the server generates output at a constant rate of $\mu$. We also know that the arrival rate is $\lambda$, therefore the loss rate is $\lambda$ - $\mu$ so the blocking probability is given by

$$P_B = \frac{\lambda - \mu}{\lambda}. \tag{390}$$

In this case, where $\lambda > \mu$, if $N > 1$, of the three cases mentioned above only the second and third events occur, and no customer will enter service upon its arrival. However, if $N = 1$ and $\lambda > \mu$, only the first and third events will occur and no customer will experience delay.

## 12.2 SLB/D/1/$N$

In this case we have an arbitrarily large burst $L_B >> N$ [packets] arrives at time 0, and no further packets ever arrive. For this case the blocking probability is

$$P_B = \frac{L_B - N}{L_B}. \tag{391}$$

and since $L_B >> N$, we have that $P_B \approx 1$. Notice that in this case $L_B$ packets arrive during a period of time $T$, with $T \to \infty$, so the arrival rate approaches zero. This case demonstrates that we can have arbitrarily small arrival rate with very high blocking probability.

## 12.3 M/M/1/$N$

As in the M/M/1 case, the M/M/1/$N$ queue-size process increases by only one and decreases by only one, so it is also a birth-and-death process. However, unlike the case of the M/M/1 birth-and-death process where the state-space is infinite, in the case of the M/M/1/$N$ birth-and-death process, the state-space is finite limited by the buffer size.

The M/M/1/$N$ queue is a special case of the state dependent SSQ considered in the previous section. If we set $\lambda_i = \lambda$ for all $i = 0, 1, 2 \ldots N-1$, $\lambda_i = 0$ for all $i \geq N$ and $\mu_i = \mu$, for all $i = 1, 2 \ldots N$, in the model of the previous section, that model is reduced to M/M/1/$N$.

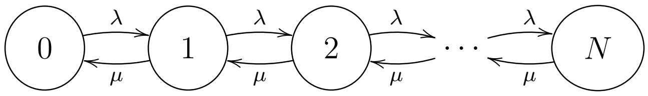
As $N$ is the buffer size, the infinitesimal generator for the M/M/1/$N$ queue-size process is given by

$Q_{i,i+1} = \lambda$ for $i = 0, 1, 2, 3, \ \ldots, \ N-1$
$Q_{i,i-1} = \mu$ for $i = 1, 2, 3, 4, \ \ldots, \ N$
$Q_{0,0} = -\lambda$
$Q_{i,i} = -\lambda - \mu$ for $i = 1, 2, 3, \ \ldots, \ N-1$
$Q_{k,k} = -\mu$.

Substituting this infinitesimal generator in Eq. (235) and performing some simple algebraic operations, we obtain the following steady-state equations for the M/M/1/$N$ queue.
$\pi_0\lambda = \pi_1\mu$
$\pi_1\lambda = \pi_2\mu$
...
and in general:

$$\pi_i\lambda = \pi_{i+1}\mu, \text{ for } i = 0, \ 1, \ 2, \ \ldots, \ N-1. \tag{392}$$

These balance equations can also be described by the following state transition diagram of M/M/1/$N$:




The normalizing equation is:

$$\sum_{j=0}^{N} \pi_j = 1. \tag{393}$$

Setting $\rho = \lambda/\mu$, so we obtain,

$\pi_1 = \rho\pi_0$
$\pi_2 = \rho\pi_1 = \rho^2\pi_0$
$\pi_3 = \rho\pi_2 = \rho^3\pi_0$

and in general:

$$\pi_i = \rho^i\pi_0 \text{ for } i = 0,\ 1,\ 2,\ \ldots N. \tag{394}$$

Summing up both sides of (394), we obtain (for the case $\rho \neq 1$)

$$1 = \sum_{i=0}^{N} \rho^i\pi_0 = \pi_0 \frac{1-\rho^{N+1}}{1-\rho}. \tag{395}$$

Therefore,

$$\pi_0 = \frac{1-\rho}{1-\rho^{N+1}}. \tag{396}$$

Substituting the latter in (394), we obtain (for the case $\rho \neq 1$)

$$\pi_i = \rho^i \frac{1-\rho}{1-\rho^{N+1}} \text{ for } i = 0,\ 1,\ 2,\ \ldots N. \tag{397}$$

Of particular interest is the blocking probability $\pi_N$ given by

$$\pi_N = \rho^N \frac{1-\rho}{1-\rho^{N+1}} = \frac{\rho^N - \rho^{N+1}}{1-\rho^{N+1}} = \frac{\rho^N(1-\rho)}{1-\rho^{N+1}}. \tag{398}$$

Notice that since M/M/1/$N$ has a finite state-space, stability is assured even if $\rho > 1$.

Considering the probabilities of the three events encountered by an arriving customer, the steady state probability of the first and the third events are given by $\pi_0$ and $\pi_N$, respectively, and the probability of the second event, also called the *delay probability* is given simply by $1 - \pi_0 - \pi_N$.

Clearly, if $N = 1$, the delay probability is equal to zero, because for $N = 1$: $1 - \pi_0 - \pi_N = 1 - \pi_0 - \pi_1 = 0$.

**Homework 12.1**

Complete the above derivations for the case $\rho = 1$, noticing that equation (395) for this case is:

$$1 = \sum_{i=0}^{N} \rho^i \pi_0 = \pi_0(N+1).$$

Alternatively, use L'Hopital rule to obtain the limit:

$$\lim_{\rho \to 1} \frac{1-\rho^{N+1}}{1-\rho}.$$

Make sure that the results are consistent. □

A numerical solution for the M/M/1/$N$ queue steady-state probabilities follows. Set an initial value for $\pi_0$ denoted $\hat{\pi}_0$ at an arbitrary value. For example, $\hat{\pi}_0 = 1$; then compute the initial value for $\pi_1$ denoted $\hat{\pi}_1$, using the equation $\hat{\pi}_0 \lambda = \hat{\pi}_1 \mu$, substituting $\hat{\pi}_0 = 1$. Then use your result for $\hat{\pi}_1$ to compute the initial value for $\pi_2$ denoted $\hat{\pi}_2$ using $\hat{\pi}_1 \lambda = \hat{\pi}_2 \mu$, etc. until all the initial values $\hat{\pi}_N$ are obtained. To obtain the corresponding $\pi_N$ values, we normalize the $\hat{\pi}_N$ values as follows.

$$\pi_N = \frac{\hat{\pi}_N}{\sum_{i=0}^{N} \hat{\pi}_i}. \qquad (399)$$

**Homework 12.2**

Consider an M/M/1/$N$ queue with $N = \rho = 1000$, estimate the blocking probability. **Answer:** 0.999. □

**Homework 12.3**

Consider an M/M/1/2 queue.

1. Show that the blocking probability is equal to $E[N_Q]$ i.e., the mean number of customers in the queue (excluding the one in service).
2. Derive $E[N_s]$ (the mean number of customers at the server), $E[Q]$ (the mean number of customers in the system) and show that $E[N_Q] + E[N_s] = E[Q]$.

**Guide**

First notice that $E[N_Q] = 0(\pi_0 + \pi_1) + 1(\pi_2) = \pi_2$ which is the blocking probability.

Next,

$$\pi_0 = \frac{1}{1+\rho+\rho^2} \qquad \pi_1 = \frac{\rho}{1+\rho+\rho^2} \qquad \pi_2 = \frac{\rho^2}{1+\rho+\rho^2}$$

Therefore,

$$E[Q] = 0\pi_0 + 1\pi_1 + 2\pi_2 = \frac{\rho + 2\rho^2}{1 + \rho + \rho^2}.$$

By Little's formula,

$$E[N_s] = \lambda(1 - \pi_2)\frac{1}{\mu} = \rho(1 - \pi_2) = \frac{\rho(1 + \rho)}{1 + \rho + \rho^2} = \frac{\rho + \rho^2}{1 + \rho + \rho^2}.$$

Clearly, $E[N_Q] + E[N_s] = E[Q]$.

□

**Homework 12.4**

A well known approximate formula that links TCP's flow rate $R_{TCP}$ [packets/sec], its round trip time (RTT), denoted $RTT$, and TCP packet loss rate $L_{TCP}$ is [61]:

$$R_{TCP} = \frac{1.22}{RTT\sqrt{L_{TCP}}}. \tag{400}$$

Consider a model of TCP over an M/M/1/$N$. That is, consider many TCP connections with a given RTT all passing through a bottleneck modeled as an M/M/1/$N$ queue. Assuming that packet sizes are exponentially distributed, estimate TCP throughput, using Equations (398) and (400) for a given RTT, mean packet size and service rate of the M/M/1/$N$ queue. Compare your results with those obtained by ns2 simulations [62].

**Guide**

Use the method of iterative fixed-point solution. See [26] and [32]. □

**Homework 12.5**

Consider a state dependent Markovian SSQ described as follows.
$\lambda_i = \lambda$ for $i = 0, 1, 2, \ldots, N - 1$
$\lambda_N = \alpha\lambda$ where $0 \leq \alpha \leq 1$
$\mu_i = \mu$ for $i = 1, 2, 3 \ldots, N$.
This represents a congestion control system (like TCP) that reacts to congestion by reducing the arrival rate. Derive the blocking probability and compare it with that of an M/M/1/$N$ SSQ with arrival rate of $\lambda$ and service rate of $\mu$. □

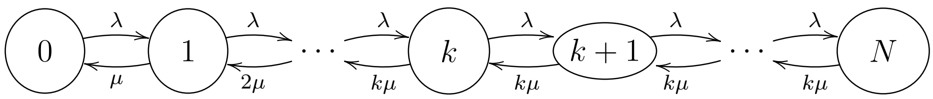
## 12.4 M/M/$k$/$N$

This Markovian queue can also be viewed as a special case of the state dependent SSQ considered in the previous section setting $\lambda_i = \lambda$ for all $i = 0, 1, 2 \ldots N - 1$, $\lambda_i = 0$ for all $i \geq N$, and setting and $\mu_i = i\mu$, for all $i = 1, 2 \ldots k$, and $\mu_i = k\mu$, for all $i = k + 1, k + 2 \ldots N$.

We can observe that the rates between states $0, 1, 2, \ldots k, k+1, k+2, \ldots N$ are the same as in the M/M/$k$ queue. Therefore, the detailed balance equations of M/M/$k$/$N$ are the same as the first $N-1$ equations of the M/M/$k$ queue. They are:

$$\begin{aligned}
\lambda\pi_0 &= \mu\pi_1 \\
\lambda\pi_1 &= 2\mu\pi_2 \\
&\ldots\ldots\ldots \\
\lambda\pi_{i-1} &= i\mu\pi_i \quad i \le k \\
\lambda\pi_{i-1} &= k\mu\pi_i \quad i = k+1, k+2, \ldots, N
\end{aligned}$$

The state transition diagram that describes these local balance equations is:




The normalizing equation is:

$$\sum_{j=0}^{N} \pi_j = 1. \tag{401}$$

Consider the notation used previously:

$$A = \frac{\lambda}{\mu},$$

and

$$\rho = \frac{A}{k}.$$

From the local balance equations, we obtain

$$\pi_i = \begin{cases} \frac{A^i}{i!}\pi_0 & \text{for } 0 \le i \le k \\ \frac{A^k}{k!}\left(\frac{A}{k}\right)^{i-k}\pi_0 & \text{for } k < i \le N. \end{cases}$$

Notice that $\pi_k$ is the probability that there are $k$ customers in the system, namely all servers are busy and the queue is empty. This probability is given by:

$$\pi_k = \frac{A^k}{k!}\pi_0. \tag{402}$$

Also, $\pi_N$, the probability that an arriving customer is blocked, is given by

$$\pi_N = \pi_0 \frac{A^k}{k!}\left(\frac{A}{k}\right)^{N-k} = \pi_k \left(\frac{A}{k}\right)^{N-k}. \tag{403}$$

Summing up the $\pi_i$s, using the normalising equation, and isolating $\pi_0$, we obtain:

$$\pi_0 = \left( \sum_{j=0}^{k-1} \frac{A^j}{j!} + \frac{A^k}{k!} \sum_{j=k}^{N} \left(\frac{A}{k}\right)^{j-k} \right)^{-1} \tag{404}$$

Summing up the second sum (the geometrical series) in (404), we obtain for the case $\rho \neq 1$ the following:

$$\begin{aligned}\sum_{j=k}^{N}\left(\frac{A}{k}\right)^{j-k} &= \sum_{j=k}^{N}\rho^{j-k} \\ &= 1+\rho+\rho^2+\ldots,+\rho^{N-k} \\ &= \frac{(1-\rho^{N-k+1})}{1-\rho}.\end{aligned}$$

This leads, in the case of $\rho \neq 1$, to the following result for $\pi_0$.

$$\pi_0 = \left(\sum_{j=0}^{k-1}\frac{A^j}{j!}+\frac{A^k}{k!}\frac{(1-\rho^{N-k+1})}{1-\rho}\right)^{-1}. \tag{405}$$

For the case $\rho = 1$, $\pi_0$ can be derived by observing that the second sum in Eq. (404), can be simplified, namely,

$$\begin{aligned}\sum_{j=k}^{N}\left(\frac{A}{k}\right)^{j-k} &= \sum_{j=k}^{N}\rho^{j-k} \\ &= 1+\rho+\rho^2+\ldots,+\rho^{N-k} \\ &= N-k+1.\end{aligned}$$

Therefore, for $\rho = 1$,

$$\pi_0 = \left(\sum_{j=0}^{k-1}\frac{A^j}{j!}+\frac{A^k}{k!}(N-k+1)\right)^{-1}. \tag{406}$$

Notice also that using L'Hopital law we obtain

$$\lim_{\rho\to 1}\frac{(1-\rho^{N-k+1})}{1-\rho} = \frac{-(N-k+1)}{-1} = N-k+1$$

which is consistent with Eq. (406).

Noticing that by Eq. (404), the expression for $\pi_0$ can be rewritten as

$$\pi_0 = \left(\sum_{j=0}^{k}\frac{A^j}{j!}+\frac{A^k}{k!}\sum_{j=k+1}^{N}\left(\frac{A}{k}\right)^{j-k}\right)^{-1}. \tag{407}$$

Then by (407) and (402), in the case $\rho \neq 1$, we obtained

$$\pi_k = \left(E_k^{-1}(A)+\rho\frac{(1-\rho^{N-k})}{1-\rho}\right)^{-1}$$

where $E_k(A)$ = the Erlang B blocking probability for an M/M/$k$/$k$ system with offered traffic $A$.

For the case $\rho = 1$, we obtain

$$\pi_k = \left(E_k^{-1}(A) + N - k\right)^{-1}.$$

A call is delayed if it finds all servers busy and there is a free place in the queue. Notice that in our discussion on the M/M/$k$ queue a call is delayed if it arrives and find all servers busy and the probability of an arriving call to be delayed is, for M/M/$k$, by the PASTA principle, the proportion of time all servers are busy. For the M/M/$k$/$N$ queue, there is the additional condition that the queue is not full as in such a case, an arriving call will be blocked. Therefore, the probability that an arriving call is delayed (the delay probability) is:

$$\begin{aligned} P(\text{delay}) &= \sum_{j=0}^{N-k-1} \pi_{j+k} \\ &= \pi_k \sum_{j=0}^{N-k-1} \rho^j \\ &= \pi_k \frac{1-\rho^{N-k}}{1-\rho}. \end{aligned}$$

We can observe that under the condition $\rho < 1$, and $N \to \infty$ the M/M/$k$/$N$ reduces to M/M/$k$. We can also observe that the M/M/1/$N$ and M/M/$k$/$k$ are also special cases of M/M/$k$/$N$, in the instances of $k = 1$ and $N = k$, respectively.

**Homework 12.6**

Show that the results of M/M/$k$, M/M/1/$N$, and M/M/$k$/$k$ for $\pi_0$ and $\pi_k$ are consistent with the results obtained of M/M/$k$/$N$. □

Next we derive the mean number of customers waiting in the queue $E[N_Q]$.

$$\begin{aligned} E[N_Q] &= \sum_{j=k+1}^{N} (j-k)\pi_j \\ &= \pi_k \sum_{j=k+1}^{N} (j-k)\rho^{j-k} \\ &= \frac{A^k}{k!}\pi_0 \sum_{j=k+1}^{N} (j-k)\rho^{j-k} \\ &= \frac{A^k \rho}{k!}\pi_0 \sum_{j=k+1}^{N} (j-k)\rho^{j-k-1} \end{aligned}$$

$$= \frac{A^k\rho}{k!}\pi_0 \sum_{i=1}^{N-k} i\rho^{i-1}.$$

Now, as before, we consider two cases: $\rho = 1$ and $\rho \neq 1$. In the case of $\rho = 1$, we have:

$$\sum_{i=1}^{N-k} i\rho^{i-1} = 1 + 2 + \ldots + N - k = \frac{(1 + N - k)(N - k)}{2}.$$

Therefore,

$$E[N_Q]_{\rho=1} = \frac{\pi_0 A^k \rho(1 + N - k)(N - k)}{2k!}.$$

In the case of $\rho \neq 1$, the mean number of customers in the queue is derived as follows:

$$\begin{aligned} E[N_Q]_{\rho\neq 1} &= \frac{\pi_0 A^k \rho}{k!}\frac{d}{d\rho}\left(\sum_{i=0}^{N-k}\rho^i\right) \\ &= \frac{\pi_0 A^k \rho}{k!}\frac{d}{d\rho}\left(\frac{1-\rho^{N-k+1}}{1-\rho}\right) \\ &= \frac{\pi_0 A^k \rho[1-\rho^{N-k+1} - (1-\rho)(N-k+1)\rho^{N-k}]}{k!(1-\rho)^2}. \end{aligned}$$

As in our previous discussion on the M/M/$k$ queue, we have

$$E[Q] = E[N_Q] + E[N_s] \tag{408}$$

and

$$E[D] = E[W_Q] + E[S]. \tag{409}$$

We know that,

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

To obtain $E[N_s]$ for the M/M/$k$/$N$ queue, we again use Little's formula for the system made of servers. recall that in the case of the M/M/$k$ queue the arrival rate into this system was $\lambda$, but now the arrival rate should exclude the blocked customers, so now in the case of the M/M/$k$/$N$ queue the arrival rate that actually access the system of servers is $\lambda(1 - \pi_N)$. The mean waiting time of each customer in that system is $E[S] = 1/\mu$ (as in M/M/$k$). Therefore, by Little's formula the mean number of busy servers is given by

$$E[N_s] = \frac{\lambda(1 - \pi_N)}{\mu} = A(1 - \pi_N). \tag{410}$$

Having $E[N_Q]$ and $E[N_s]$ we can obtain the mean number of customers in the system $E[Q]$ by Eq. (408).

Then by Little's formula we obtain

$$E[D] = \frac{E[Q]}{\lambda(1 - \pi_N)}$$

and

$$E[W_Q] = \frac{E[N_Q]}{\lambda(1-\pi_N)}$$

Also, since

$$E[S] = \frac{1}{\mu},$$

by (409), we also have the relationship:

$$E[D] = E[W_Q] + \frac{1}{\mu}.$$

If we are interested in the mean delay of the delayed customers, denoted $E[\hat{D}] = E[D|Delayed]$, we notice that.

$$E[\hat{D}] = E[D|Delayed] = E[W_Q|Delayed] + \frac{1}{\mu},$$

where $E[W_Q|Delayed]$ is the the mean waiting time in the queue of delayed customers. To obtain $E[W_Q|Delayed]$ we again use Little's formula considering the system that includes only the delayed customers, as follows:

$$E[W_Q|Delayed] = \frac{E[N_Q]}{\lambda P(\text{delay})},$$

where $P(\text{delay}) = \pi_k + \pi_{k+1} + \ldots + \pi_{N-1}$.

An alternative approach is to use the law of iterated expectation by solving the following equation for $E[\hat{D}]$.

$$E[D] = \frac{\pi_0 + \pi_1 + \ldots + \pi_{k-1}}{1-\pi_N} \times \frac{1}{\mu} + \frac{\pi_k + \pi_{k+1} + \ldots + \pi_{N-1}}{1-\pi_N} \times E[\hat{D}].$$

**Homework 12.7**

Show that the lower bound for the M/M/$k/k$ blocking probability obtained in (330), namely

$$\max\left(0, 1 - \frac{k}{A}\right)$$

is also applicable to the M/M/$k/N$ queue. □

## 12.5 MMPP(2)/M/1/$N$

In Section 2.3, we described the MMPP and its two-state special case – the MMPP(2). Here we consider an SSQ where the MMPP(2) is the arrival process.

The MMPP(2)/M/1/$N$ Queue is an SSQ with buffer size $N$ characterized by an MMPP(2) arrival process with parameters $\lambda_0$, $\lambda_1$, $\delta_0$, and $\delta_1$, and exponentially distributed service time with parameter $\mu$. The service times are mutually independent and are independent of the arrival process. Unlike the Poisson arrival process, the inter-arrival times in the case of the MMPP(2) process are not independent. As will be discussed, such dependency affects queueing performance, packet loss and utilization.

The MMPP(2)/M/1 queue process is a continuous-time Markov-chain, but its states are two-dimensional vectors and not scalars. Each state is characterized by two scalars: the mode $m$ of the arrival process that can be either $m = 0$ or $m = 1$ and the queue size. Notice that all the other queueing systems we considered so far were based on a single dimensional state-space.

Let $p_{im}$ for $i = 0, 1, 2 \ldots, N$ be the probability that the arrival process is in mode $m$ and that there are $i$ packets in the system. After we obtain the $\pi_{im}$ values, the steady-state queue size probabilities can then be obtained by

$$\pi_i = \pi_{i0} + \pi_{i1} \quad \text{for} \quad i = 0, 1, 2, \ldots, N.$$

Note that the mode process itself is a two-state continues-time Markov-chain, so the probabilities of the arrival mode being in state $j$, denoted $P(m = j)$, for $j = 0, 1$, can be solved using the following equations:

$$P(m = 0)\delta_0 = P(m = 1)\delta_1$$

and the normalizing equation

$$P(m = 0) + P(m = 1) = 1.$$

Solving these two equations gives the steady-state probabilities $P(m = 0)$ and $P(m = 1)$ as functions of the mode duration parameters $\delta_0$ and $\delta_1$, as follows:

$$P(m = 0) = \frac{\delta_1}{\delta_0 + \delta_1} \tag{411}$$

$$P(m = 1) = \frac{\delta_0}{\delta_0 + \delta_1}. \tag{412}$$

Because the probability of the arrival process to be in mode $m$ (for $m = 0, 1$) is equal to $\sum_{i=0}^{N} \pi_{im}$, we obtain by (411) and (412)

$$\sum_{i=0}^{N} \pi_{im} = \frac{\delta_{1-m}}{\delta_{1-m} + \delta_m} \quad \text{for} \quad m = 0, 1. \tag{413}$$

The average arrival rate, denoted $\lambda_{av}$, is given by

$$\lambda_{av} = P(m = 0)\lambda_0 + P(m = 1)\lambda_1 = \frac{\delta_1}{\delta_0 + \delta_1}\lambda_0 + \frac{\delta_0}{\delta_0 + \delta_1}\lambda_1. \tag{414}$$

Denote

$$\rho = \frac{\lambda_{av}}{\mu}.$$

The MMPP(2)/M/1/$N$ queueing process is a stable, irreducible and aperiodic continuous-time Markov-chain with finite state-space (because the buffer size $N$ is finite). We again remind the reader that the condition $\rho < 1$ is not required for stability in a finite buffer queueing system, or more generally, in any case of a continuous-time Markov-chain with finite state-space. Such a system is stable even if $\rho > 1$.

An important performance factor in queues with MMPP(2) input is the actual time the queue stays in each mode. Even if the apportionment of time between the modes stays fixed, the actual time can make a big difference. This is especially true for the case $\rho_1 = \lambda_1/\mu > 1$ and $\rho_2 = \lambda_1/\mu < 1$, or vise versa. In such a case, if the actual time of staying in each mode is long, there will be a long period of overload when a long queue is built up and/or many packets lost, followed by long periods of light traffic during which the queues are cleared. In such a case we say that the traffic is *bursty* or strongly correlated. (As mentioned above here inter-arrival times are not independent.) On the other hand, if the time of staying in each mode is short; i.e., the mode process exhibits frequent fluctuations, the overall traffic process is smoothed out and normally long queues are avoided. To see this numerically one could set initially $\delta_0 = \delta_0^*$ $\delta_1 = \delta_1^*$ where, for example, $\delta_0 = 1$ and $\delta_1^* = 2$, or $\delta_0^* = \delta_1^* = 1$, and then set $\delta_m = \psi\delta_m^*$ for $m = 0, 1$. Letting $\psi$ move towards zero will mean infrequent fluctuations of the mode process that may lead to bursty traffic (long stay in each mode) and letting $\psi$ move towards infinity means frequent fluctuations of the mode process. The parameter $\psi$ is called *mode duration parameter*. In the exercises below the reader is asked to run simulations and numerical computations to obtain blocking probability and other measures for a wide range of parameter values. Varying $\psi$ is one good way to gain insight into performance/burstiness effects.

Therefore, the $\pi_{im}$ values can be obtain by solving the following finite set of steady-state equations:

$$0 = \mathbf{\Pi}\mathbf{Q} \tag{415}$$

where $\mathbf{\Pi} = [\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}, \pi_{20}, \pi_{21}, \ldots, \pi_{N-1,0}, \pi_{N-1,1}, \pi_{N0}, \pi_{N1}]$, and the infinitesimal generator $2N \times 2N$ matrix is $\mathbf{Q} = [Q_{\mathbf{i},\mathbf{j}}]$, where $\mathbf{i}$ and $\mathbf{j}$ are two-dimensional vectors. Its non-zero entries are:

$Q_{00,00} = -\lambda_0 - \delta_0$;  $Q_{00,01} = \delta_0$;  $Q_{00,10} = \lambda_0$;

$Q_{01,00} = \delta_1$;  $Q_{01,01} = -\lambda_1 - \delta_1$;  $Q_{01,11} = \lambda_1$;

For $N > i > 0$, the non-zero entries are:

$Q_{i0,i0} = -\lambda_0 - \delta_0 - \mu$;  $Q_{i0,i1} = \delta_0$;  $Q_{i0,(i+1,0)} = \lambda_0$;

$Q_{i1,i0} = \delta_1$;  $Q_{01,01} = -\lambda_1 - \delta_1 - \mu$;  $Q_{i1,(i+1,1)} = \lambda_1$;

and

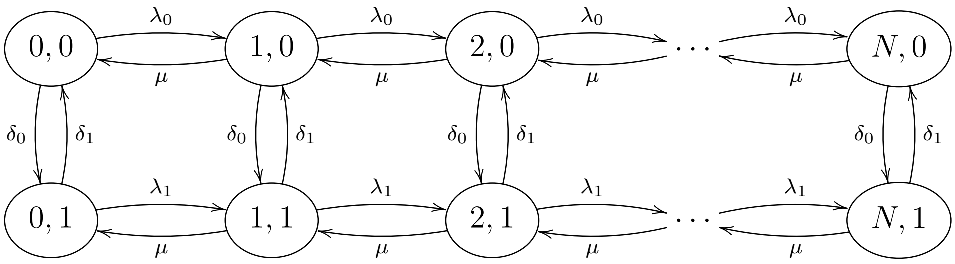
$Q_{N0,(N-1,0)} = \mu$;  $Q_{N0,N0} = -\delta_0 - \mu$;  $Q_{N0,N1} = \delta_0$;

$Q_{N1,(N-1,1)} = \mu$;  $Q_{N1,N1} = -\delta_1 - \mu$;  $Q_{N1,N0} = \delta_1$.

In addition we have the normalizing equation

$$\sum_{i=0}^{N}\sum_{m=0}^{1} \pi_{im} = 1. \tag{416}$$

The state transition diagram for the MMPP(2)/M/1/$N$ queue is:




An efficient way, that normally works well for solving the steady-state equations of the MMPP(2)/M/1/$N$ queue is the so called *successive substitution* method (it is also known as Gauss-Seidel, successive approximation or iterations) [22]. It can be described as follows. Consider a set of equation of the form of (415). First, isolate the first element of the vector $\mathbf{\Pi}$, in this case it is the variable $\pi_{00}$ in the first equation. Next, isolate the second element of the vector $\mathbf{\Pi}$, namely $\pi_{01}$ in the second equation, and then keep isolation all the variables of the vector $\mathbf{\Pi}$. This leads to the following vector equation for $\mathbf{\Pi}$

$$\mathbf{\Pi} = f(\mathbf{\Pi}). \tag{417}$$

where $f(\mathbf{\Pi})$ is of the form

$$f(\mathbf{\Pi}) = \mathbf{\Pi}\hat{\mathbf{Q}}$$

where $\hat{\mathbf{Q}}$ is different from the original $\mathbf{Q}$ because of the algebraic operations we performed when we isolated the elements of the $\mathbf{\Pi}$ vector. Then perform the successive substitution operations by setting arbitrary initial values to the vector $\mathbf{\Pi}$; substitute them in the right-hand side of (417) and obtain different values at the left-hand side which are then substituted back in the right-hand side, etc. For example, the initial setting can be $\mathbf{\Pi} = 1$ without any regards to the normalization equation (416). When the values obtain for $\mathbf{\Pi}$ are sufficiently close to those obtained in the previous subsection, say, within a distance no more than $10^{-6}$, stop. Then normalize the vector $\mathbf{\Pi}$ obtained in the last iteration using (416). This is the desired solution.

After obtaining the solution for Eq. (415) and (416), one may verify that (413) holds.

Since the arrival process is not Poisson, so the PASTA principle does not apply here. Therefore, to obtain the blocking probability $P_b$ we again notice that $\pi_N = \pi_{N0} + \pi_{N1}$ is the proportion of time that the buffer is full. The proportion of packets that are lost is therefore the ratio of the number of packets arrive during the time that the buffer is full to the total number of packets that arrive. Therefore,

$$P_b = \frac{\lambda_0 \pi_{N0} + \lambda_1 \pi_{N1}}{\lambda_{av}}. \tag{418}$$

As an example, we hereby provide the infinitesimal generator for $N = 2$:

| | **00** | **01** | **10** | **11** | **20** | **21** |
|---|---|---|---|---|---|---|
| **00** | $-\lambda_0 - \delta_0$ | $\delta_0$ | $\lambda_0$ | 0 | 0 | 0 |
| **01** | $\delta_1$ | $-\lambda_1 - \delta_1$ | 0 | $\lambda_1$ | 0 | 0 |
| **10** | $\mu$ | 0 | $-\lambda_0 - \delta_0 - \mu$ | $\delta_0$ | $\lambda_0$ | 0 |
| **11** | 0 | $\mu$ | $\delta_1$ | $-\delta_1 - \mu$ | 0 | $\lambda_1$ |
| **20** | 0 | 0 | $\mu$ | 0 | $-\lambda_0 - \delta_0 - \mu$ | $\delta_0$ |
| **21** | 0 | 0 | 0 | $\mu$ | $\delta_1$ | $-\delta_1 - \mu$ |

**Homework 12.8**

Consider an MMPP(2)/M/1/1 queue with $\lambda_0 = \delta_0 = 1$ and $\lambda_1 = \delta_1 = 2$ and $\mu = 2$.

1. Without using a computer solve the steady-state equations by standard methods to obtain $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$ and verify that (413) holds.
2. Obtain the blocking Probability.
3. Find the proportion of time that the server is idle.
4. Derive an expression and a numerical value for the utilization.
5. Find the mean queue size. □

**Homework 12.9**

Consider an MMPP(2)/M/1/200 queue with $\lambda_0 = 1$, $\delta_0 = 10^{-3}$, $\lambda_1 = 2$, $\delta_1 = 2 \times 10^{-3}$ and $\mu = 1.9$.

1. Solve the steady-state equations by sucessive substitutions to obtain the $\pi_{im}$ values and verify that (413) holds.
2. Obtain the blocking Probability.
3. Find the proportion of time that the server is idle.
4. Obtain numerical value for the utilization.
5. Find the mean queue size.
6. Compare the results obtained with those obtained before for the case $N = 1$ and discuss the differences. □

**Homework 12.10**

Consider again the MMPP(2)/M/1/200 queue. Using successive substitutions, obtain the mean queue size for a wide range of parameter values and discuss differences. Confirm your results by simulations with confidence intervals. Compare the results with those obtained by successive substitution and simulation of an equivalent M/M/1/200 queue that has the same service rate and its arrival rate is equal to $\lambda_a v$ of the MMPP(2)/M/1/200. Provide interpretations and explanations to all your results. □

**Homework 12.11**

Consider again the MMPP(2)/M/1/200 queue and its M/M/1/200 equivalence. For a wide range of parameter values, compute the minimal service rate $\mu$ obtained such that the blocking probability is no higher than $10^{-4}$ and observe the utilization. Plot the utilization as a function of the mode duration parameter $\psi$ to observe the effect of burstiness on the utilization. Confirm your results obtained by successive substitution by simulations using confidence intervals. Demonstrate that as $\psi \rightarrow \infty$ the performance (blocking probability and utilization) achieved approaches that of the M/M/1/200 equivalence. Discuss and explain all the results you obtained. □

## 12.6 M/E$_\mathrm{m}$/1/$N$

We consider here an M/E$_\mathrm{m}$/1/$N$ SSQ model characterized by a Poisson arrival process with parameter $\lambda$, buffer size of $N$, and service time that has Erlang distribution with $m$ phases (E$_\mathrm{m}$) with mean $1/(\mu)$. Such service time model arises is situations when the standard deviation to mean ratio of the service time is lower than one (recall that for the exponential random variable this ratio is equal to one).

**Homework 12.12**

Derive and plot the standard deviation to mean ratio as a function of $m$ for an E$_\mathrm{m}$ random variable. □

This queueing system can be analyzed using a two-dimensional state-space representing the number of customers and the number of phases still remained to be served for the customer in service. However, it is simpler if we are able to represent the system by a single dimension state-space. In the present case this can be done by considering the total number of phases as the state, where each of the items (phases) in the queue is served at the rate of $m\mu$ and an arrival adds $m$ items to the queue. The total number of items (phases) is limited to $m \times N$ because the queue size is limited to $N$ customers each of which required $m$ service phases. Notice the one-to-one correspondence between the single dimension vector $(0, 1, 2, \ldots m \times N)$ and the ordered set $(0, 11, 12, \ldots, 1m, 21, 22, \ldots 2m, 31, \ldots Nm)$ where the first element is 0, and the others are 2-tuples where the first is the number of customers and the second is the number of phases remains for the customer in service.

Let $\pi_i$ be the probability that there are $i$ items (phases) in the queue $i = 0, 1, 2, 3, \ldots, m \times N$. For clarity of presentation also define

$$\pi_i = 0 \quad \text{for} \quad i < 0.$$

This model is a continuous-time Markov-chain, so the steady-state probabilities $\pi_i$, $i = 1, 2, 3, \ldots, m \times N$ satisfy the following local-balance steady-state equations:

$$\begin{aligned}\lambda\pi_0 &= m\mu\pi_1 \\ (\lambda+m\mu)\pi_i &= m\mu\pi_{i+1}+\lambda\pi_{i-m} \;\text{ for }\; i=2,3,\ldots,m\times N-1\end{aligned}$$

The first equation equates the probability flux of leaving state 0 (to state $m$) with the probability flux of entering state 0 only from state 1 - where there is only one customer in the system who is in its last service phase (one item). The second equation equates the probability flux of leaving state i (either by an arrival or by completion of the service phase) with the probability flux of entering state i (again either by an arrival, i.e., a transition from below from state $i-m$, or from above by phase service completion from state i+1).

The probability of having $i$ customers in the system, denoted $P_i$, is obtained by

$$P_i=\sum_{j=1}^{m}\pi_{(i-1)m+j}.$$

The blocking probability is the probability that the buffer is full namely $P_N$. The mean queue size is obtained by

$$E[Q]=\sum_{i=1}^{N}i\pi_i.$$

The mean delay is obtained by Little's formula:

$$E[D]=\frac{E[Q]}{\lambda}.$$

**Homework 12.13**

Plot the state transition diagram for the M/$E_m$/1/$N$ considering the number of phases as the state. □

**Homework 12.14**

Consider an M/$E_m$/1/$N$ queue. For a wide range of parameter values (varying $\lambda, \mu, m, N$) using successive substitutions, obtain the mean queue size, mean delay and blocking probability and discuss the differences. Confirm your results by simulations using confidence intervals. Provide interpretations and explanations to all your results. □

## 12.7 Saturated Queues

Saturated queues are characterized by having all the servers busy all the time (or almost all the time). In such a case it is easy to estimate the blocking probability for queues with finite buffers, by simply considering the so-called *fluid flow model*. Let us consider, for example, An M/M/$k$/$N$ queue, and assume that either the arrival rate $\lambda$ is much higher than the total service rate of all $k$ servers $k\mu$, i.e., $\lambda >> k\mu$, or that $\lambda > k\mu$ and $N >> 0$. Such conditions will guarantee that the servers will be busy all (or most of) the time. Since all $k$ servers are busy all the time, the output rate of the system is $k\mu$ packets/s and since the input is $\lambda$ packets/s during a very long period of time $L$, there will be $\lambda L$ arrivals and $k\mu L$ departures. Allowing $L$ to be arbitrarily large, so that the initial transient period during which the buffer is filled can be ignored, the blocking probability can be evaluated by

$$P_b = \frac{\lambda L - k\mu L}{\lambda L} = \frac{\lambda - k\mu}{\lambda} = \frac{A - k}{A}, \tag{419}$$

where $A = \lambda/\mu$.

Another way to see (419) is by recalling that the overflow traffic is equal to the offered traffic minus the carried traffic. The offered traffic is $A$, the carried traffic in a saturated M/M/$k$/$N$ queue is equal to $k$ because all $k$ servers are continuously busy so the mean number of busy servers is equal to $k$ and the overflow traffic is equal to $AP_b$. Thus,

$$A - k = AP_b$$

and (419) follows.

**Homework 12.15**

Consider an M/M/$k$/$N$ queue. Write and solve the steady-state equations to obtain exact solution for the blocking probability. A numerical solution is acceptable. Validate your results by both Markov-chain and discrete event simulations using confidence intervals. Then demonstrate that as $\lambda$ increases the blocking probability approaches the result of (419). Present your results for a wide range of parameter values (varying $\lambda, \mu, N, k$). Provide interpretation of your results. □

**Homework 12.16**

Consider again an M/M/1/$N$ queue with $N = \rho = 1000$ and estimate the blocking probability, but this time use the saturated queue approach. **Answer:** 0.999. □

# 13 Processor Sharing

In a processor sharing (PS) queueing system the server capacity is shared equally among all the customers that are present in the system. This model is applicable to a time-shared computer system where a central processor serves all the jobs present in the system simultaneously at an equal service rate. Another important application of PS is for a multiplicity of TCP connections that share a common bottleneck. The Internet router at the bottleneck simultaneously switches (serves) the flows generated by the users, while TCP congestion control mechanism guarantees that the service rate obtained by the different flows are equal. As any of the other models considered in this book, the PS model is only an approximation for the various real-life scenarios. It does not consider overheads and wastage associated with the discrete nature and various protocol operations of computer systems, and therefore it may be expected to over-estimate performance (or equivalently, underestimate delay).

If the server capacity to render service is $\mu$ [customers per time-unit] and there are $i$ customers in the system, each of the customers is served at the rate of $\mu/i$. As soon as a customer arrives, its service starts.

## 13.1 The M/M/1-PS queue

The M/M/1-PS queue is characterized by Poisson arrivals and exponentially distributed service-time requirement, as the ordinary (FIFO) M/M/1 queue), but its service regime is assumed to be processor sharing. In particular, we assume that the process of the number of customers $i$ in the system is a continuous time Markov-chain, where customers arrive according to a Poisson process with parameter $\lambda$ [customers per time-unit] and that the service time required by an arriving customer is exponentially distributed with parameter $\mu$. We also assume the stability condition of $\lambda < \mu$.

Let us consider now the transition rates of the continuous-time Markov chain for the number of customers in the system associated with the M/M/1-PS model. Firstly, we observe that the transition rates from state $i$ to state $i+1$ is $\lambda$ as in the M/M/1 model. We also observe that the rates from state $i$ to state $i+j$ for $j>1$ and from state $i$ to state $i-j$ for $j>1$ are all equal to zero (again, as in M/M/1). The latter is due to the fact that the probability of having more than one event, arrival or departure, occurred at the same time is equal to zero. To derive the rates from state $i$ to state $i-1$ for $i \geq 1$ notice that at state $i$, assuming that no arrivals occur, the time until a given customer completes its service is exponentially distributed with rate $\mu/i$. Therefore, the time until the first customer out of the $i$ customers that completes its service is the minimum of $i$ exponential random variables each of which with rate $\mu/i$, which is exponentially distributed with rate $i(\mu/i) = \mu$. Therefore, the transition rates from state $i$ to state $i-1$ is equal to $\mu$ (again, as in M/M/1). These imply that the process of number of customers in the system associated with the M/M/1-PS model is statistically the same as the continuous-time Markov chain that describes the M/M/1 (FIFO) queue. Therefore the queue size state-state distribution $\{\pi_i\}$ and the mean queue-size $E[Q]$ given by equations (282) and (283), respectively, are also applied to the M/M/1-PS model. That is,

$$\pi_i = \rho^i(1-\rho) \text{ for } i = 0,\ 1,\ 2,\ \ldots$$

and

$$E[Q] = \frac{\rho}{1-\rho}. \tag{420}$$

By Little's formula the result obtained for the mean delay $E[D]$ in Eq. (284) is also applicable to the M/M/1-PS model:

$$E[D] = \frac{1}{(1-\rho)\mu} = \frac{1}{\mu - \lambda}. \tag{421}$$

However the delay distribution of M/M/1 given by Eq. (285) does not apply to M/M/1-PS.

Having obtained the mean delay for a customer in the M/M/1-PS queue, an interesting question is what is the mean delay of a customer that requires amount of service $x$. The variable $x$ here represents the time that the customer spends in the system to complete its service assuming that there are no other customers being served and all the server capacity can be dedicated to it. By definition, $E[x] = 1/\mu$.

This is not an interesting question for the M/M/1 queue because under the FIFO discipline, the time a customer waits in the queue is not a function of $x$ because it depends only on service requirements of other customers. Only after the customer completes its waiting time in the queue, $x$ will affects its total delay simply by being added to the waiting time in the queue. By comparison, in the case of the M/M/1-PS queue, the mean delay of a customer in the system from the moment it arrives until its service it complete $D(x)$ has linear relationship with $x$ [53, 55]. That is,

$$D(x) = cx, \tag{422}$$

for some constant $c$.

We know that under our stability assumption, the process of the number of customers in the system $i$ is a stable and stationary continuous time Markov chain. In fact, it is a birth-and-death process because the transitions are only up by one or down by one. Therefore, the infinitesimal service rate obtained by a test customer will also follow a stable and stationary continuous time Markov chain. Although some customers will receive higher average service rate than others, the implication of (422) is that, on average, if a customer require twice as much service than another customer, its mean delay will be twice that of the delay of the other customer.

Taking the mean with respect to $x$ on both sides of (422) we obtain

$$E[D] = c\frac{1}{\mu},$$

and by (421) this leads to

$$\frac{1}{(1-\rho)\mu} = c\frac{1}{\mu}.$$

Thus

$$c = \frac{1}{1-\rho},$$

so by the latter and (422), we obtain

$$E[D|x] = \frac{x}{1-\rho}. \quad (423)$$

## 13.2 Insensitivity

One important property of a processor sharing queue is that the mean number of customers in the system $E[Q]$ the mean delay of a customer $E[D]$, and the mean delay of a customer with service requirement $x$, $E[D(x)]$, given by Eqs. (420) and (421), and (423), respectively, are insensitive to the shape of the distribution of the service-time requirements of the customers. In other words, these results apply also to the M/G/1-PS model characterized by Poisson arrivals, generally distributed service-time requirements, and a processor sharing service policy. The M/G/1-PS model is a generalization of the M/M/1-PS model where we relax the exponential distribution of the service time requirements of the M/M/1-PS model, but retaining the other characteristics of the M/M/1-PS model, namely, Poisson arrivals and processor sharing service discipline.

Furthermore, the insensitivity property applies also to the distribution of the number of customers in the system, but not to the delay distribution. This means that the geometric distribution of the steady-state number of customers in the system of M/M/1 applies also to the M/G/1-PS model and it is insensitive to the shape of the distribution of the service time requirement. Notice that these M/M/1 results extend to the M/M/1-PS and M/G/1-PS models, but do not extend to the M/G/1 model. See discussion on the M/G/1 queue in Chapter 16.

Although the insensitivity applies to the distribution of the number of customers in the M/G/1-PS model, it does not apply to the delay distribution of M/G/1-PS.

Finally, notice the similarity between the M/G/1-PS and the M/G/$\infty$ models. They are both insensitive to the shape of the distribution of the service time requirement in terms of mean delay and mean number of customers in the system. In both, the insensitivity applies to the distribution of the number of customers in the system, but does not apply to the delay distribution.

**Homework 13.1**

Consider packets arriving at a multiplexer where the service discipline is based on processor sharing. Assume that the service rate of the multiplexer is 2.5 Gb/s. The mean packet size is 1250 bytes. The packet arrival process is assumed to follow a Poisson process with rate of 200,000 [packet/sec] and the packet size is exponentially distributed.

1. Find the mean number of packets in the multiplexer.
2. Find the mean delay of a packet.
3. Find the mean delay of a packet of size 5 kbytes.

**Solution**

1. Find the mean number of packets in the multiplexer.

mean packet size = $1250 \times 8 = 10,000$ bits

$$\mu = \frac{2,500,000,000}{10,000} = 250,000 \quad \text{packet/s}$$

$$\rho = \frac{200,000}{250,000} = 0.8$$

$$E[Q] = \frac{0.8}{1-0.8} = 4 \quad \text{approx.}$$

2. Find the mean delay of a packet.

By Little's formula

$$E[D] = \frac{E[Q]}{\lambda} = \frac{4}{200,000} = 0.00002 \quad \text{sec.} = 20 \quad \text{microseconds}$$

3. Find the mean delay of a packet of size 5 kbytes.

Let $x$ be the time that the 5 kbytes packet is delayed if it is the only one in the system.

$$x = \frac{5000 \times 8}{2.5 \times 1,000,000,000} = 16 \quad \text{microseconds}$$

Now we will use the time units to be microseconds.

$$E[D(x)] = \frac{x}{1-\rho} = \frac{16}{1-0.8} = 80 \quad \text{microseconds}$$

A packet four times larger than the average sized packet will be delayed four times longer. □

**Homework 13.2**

Assume that packets that arrive at a processor sharing system are classified into $n$ classes of traffic, where the $i$th class is characterized by Poisson arrivals with arrival rate $\lambda_i$, $i = 1, 2, \ldots n$, and required mean holding time (assuming a packet is alone in the system) $h_i$, $i = 1, 2, \ldots n$. The server rate is $\mu$.

1. Find the mean delay of a packet.
2. Find the mean delay of a packet that requires service time $x$.
3. Find the mean number of packets in the system.
4. Find the mean number of class $i$, $i = 1, 2, \ldots n$, packets in the system.

5. Show that the mean number of packets in the system is equal to the sum of the means obtained for classes $i$, $i = 1, 2, \ \ldots \ n$.

**Guide**

The arrival rate of all packets is given by $\lambda = \sum_{i=1}^{n} \lambda_i$. The mean holding time of a packet is given by

$$h = \frac{\sum_{i=1}^{n} \lambda_i h_i}{\lambda}.$$

Then

$$\rho = \lambda h = \sum_{i=1}^{n} \lambda_i h_i.$$

Invoke insensitivity and use equations (421) and (423).

To find the mean number of packets in the system, you can use either Little's formula, or the M/M/1 model.

Next, considering the mean delay of a packet that require $h_i$ time if it is the only one in the system, obtain the mean system time of class $i$ packets. Then having the arrival rate $\lambda_i$ and the mean system time of class $i$ packets, by Little's formula, obtain the mean number of class $i$ customers in the system.

Finally, with the help of some algebra you can also show that the mean number of packets in the system is equal to the sum of the means obtained for classes $i$, $i = 1, 2, \ \ldots, \ n$. $\square$

# 14 Multi-service Loss Model

We have discussed in Section 8.10, a case of a Markovian multi-server loss system ($k$ servers without additional waiting room), involving different classes (types) of customers where customers belong to different classes (types) may be characterized by different arrival rates and holding times. There we assumed that each admitted arrival will always be served by a single server. We will now extend the model to the case where customers of some classes may require service by more than one server simultaneously. This is applicable to a telecommunications network designed to meet heterogenous service requirements of different applications. For example, it is clear that a voice call will require lower service rate than a movie download. In such a case, a movie download will belong to a class that requires more servers/channels than that of the voice call. By comparison, the M/M/$k$/$k$ system is a multi-server single-service loss model, while here we consider a multi-server multi-service loss model and we are interested in the blocking probability of each class of traffic.

As the case is with the Erlang Loss System, the blocking probability is an important performance measure also in the more general multi-service system with a finite number of servers. However, unlike the case in the M/M/$k$/$k$ system where all customers experience the same blocking probability, in the case of the present multi-service system, customers belonging to different classes experience different blocking probabilities. This is intuitively clear. Consider a system with 10 servers and assume that seven out of the 10 servers are busy. If a customer that requires one server arrives, it will not be blocked, but if a new arrival, that requires five servers, will be blocked. Therefore, in many cases, customers that belong to class that requires more servers, will experience higher blocking probability. However, there are cases, where customers of different classes experience the same blocking probability. See the relevant homework question below.

This chapter covers certain key issues on multi-service models, but it provides intuitive explanations rather than rigorous proofs. For more extensive coverage and rigorous treatments, the reader is referred to [47] and [75] and to earlier publications on the topic [30, 46, 49, 71, 73, 76, 77, 78].

## 14.1 Model Description

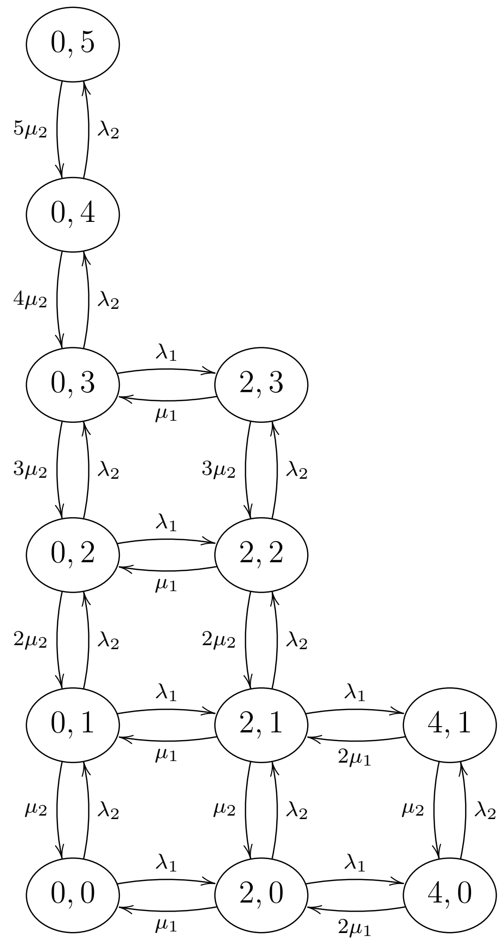
Consider a set of $k$ servers that serve arriving customers that belong to $I$ classes. Customers from class $i$ require simultaneous $s_i$ servers and their holding times are assumed exponentially distributed with mean $1/\mu_i$. (As the case is for the M/M/$k$/$k$ system, the results of the analysis presented here are insensitive to the shape of the distribution of the holding time, but since we use a continuous time Markov-chain modelling, this exponential assumption is made for now.) Class $i$ customers arrive according to an independent Poisson process with arrival rate $\lambda_i$. The holding times are independent of each other, of the arrival processes and of the state of the system.

Define

$$A_i = \frac{\lambda_i}{\mu_i}.$$

As discussed, an admitted class-$i$ customer will use $s_i$ servers for the duration of its holding time which has mean of $1/\mu_i$. After its service time is complete, all these $s_i$ servers are released and they can serve other customers. When a class-$i$ customer arrives, and cannot find $s_i$ free servers, its service is denied and it is blocked and cleared from the system. An important measure is the probability that an arriving class-$i$ customer is blocked. This is called the class-$i$ customer blocking probability denoted $B(i)$.

Next, we provide the state transition diagram for the multi-service loss model for the case $k = 5$, $s_1 = 2$, and $s_2 = 1$:




## 14.2 Attributes of the Multi-service Loss Model

The multi-service system model as defined above has the following important attributes.

1. **Accuracy and Scalability:** The process of the number of customers in the system of the various classes is reversible. This property implies that the detailed balance equations hold and together with the normalizing equation lead to exact solution. It is far easier to solve the detailed balance equations than the global balance equations and therefore the exact solution is scalable to problems of realistic size.

2. **Robustness – Insensitivity:** The blocking probabilities depend on the customers' holding time distributions only through their means. That is, they are insensitive to the shape of the holding time distributions. This insensitivity property implies that holding time (packet size, or flow size) can have any distribution. All that we need to know about the holding times of the calls of the various services are their means, and the exact block-

ing probability for each service type can be obtained using the local balance equations as if the holding times follow exponential distributions. It is known that the Internet flows follow a heavy tailed distribution such as Pareto. Due to this insensitivity property, the model is robust enough to be exact even for heavy-tailed holding time distributions. This makes the analyzes and results of multi-service systems very relevant for real life telecommunications systems and networks.

3. **Applicability:** Given the wide diversity of bandwidth requirements of Internet services, and limited capacity of communications links, there is a clear need for a model that will provide performance evaluation in terms of blocking probability. The M/M/$k$/$k$ which is a special case of this model (for the case of a single service class) has been a cornerstone in telephony used by engineers to design and dimension telephone networks for almost a century due to its accuracy, scalability and robustness. In telephony we have had one service phone calls all requiring the same link capacity. As we have entered the Internet age, the multi-service model, given its accuracy, scalability, and robustness can play an important role. As discussed, the insensitivity and scalability properties of the M/M/$k$/$k$ system extends to the multi-service system model and make it applicable to practical scenarios. For example, a transmission trunk or lightpath [94] has limited capacity which can be subdivided into many wavelength channels based on wavelength division multiplexing (WDM) and each wavelength channel is further subdivided into TDM sub-channels. Although the assumption of Poisson arrivals of Internet flows during a busy-hour that demand capacity from a given trunk or a lightpath may be justified because they are generated by a large number of sources, the actual demand generated by the different flows/connections vary significantly from a short SMS or email, through voice calls, to large movie downloads, and far larger data bursts transmitted between data centers or experimental data generated, for example, by the Large Hadron Collider (LHC). These significant variations imply a large variety of capacity allocated to the various flows/connections and also large variety in their holding times, so that the restrictive exponentially distributed holding time assumption may not be relevant. Therefore, the insensitivity property of the multi-service loss model is key to the applicability of the multi-service model.

## 14.3 A Simple Example with $I = 2$ and $k = 2$

Consider a multi-service-system with two classes of services (voice and video). Both traffic streams of voice and video calls follow a Poisson process and their holding time are exponentially distributed. The arrival rate of the voice service is $\lambda_1 = 0.3$ calls per minute and the average voice service-time $1/\mu_1$ is 3 minutes. The arrival rate of the video service is $\lambda_2 = 0.2$ calls per minute and the average video service-time $1/\mu_2$ is 5 minutes. The system has two channels (servers).

We now aim to calculate the blocking probability of the arriving voice calls and of the arriving video calls in the case where the voice service requires one channel per call and video service requires two channels per call. The system has two channels (servers), i.e, $k = 2$.

Let $j_i$ be the number of channels used to serve class-$i$ customers for $i = 1, 2$. Then the state space is all feasible pairs $\{j_1, j_2\}$, namely: (0,0), (1,0), (2,0), (0,2).

**Homework 14.1**

Plot the state transition diagram for this case with $I = 2$ and $k = 2$. □

Let $\pi_{j_1,j_2}$ be the steady-state probability of being in state $(j_1, j_2)$. Then we obtain the following global balance equations.

$$\begin{aligned}
(\lambda_1 + \lambda_2)\pi_{0,0} &= \mu_1\pi_{1,0} + \mu_2\pi_{0,2} \\
(\mu_1 + \lambda_1)\pi_{1,0} &= \lambda_1\pi_{0,0} + 2\mu_1\pi_{2,0} \\
2\mu_1\pi_{2,0} &= \lambda_1\pi_{1,0} \\
\mu_2\pi_{0,2} &= \lambda_2\pi_{0,0}.
\end{aligned}$$

Each of these equations focuses on one state and represents the balance of the total probability flux out and into the state. The first equation focuses on the state (0,0), the second on (1,0), the third on (2,0) and the fourth on (0,2).

By the first and the fourth equations we can obtain a fifth equation:

$$\mu_1\pi_{1,0} = \lambda_1\pi_{0,0}.$$

The same result is obtained by the second and third equations.

The third, fourth and fifth equations are a complete set of detailed balance equations representing the balance of probability flux between each pair of neighboring states. These three detailed balance equations together with the normalizing equation

$$\pi_{0,0} + \pi_{1,0} + \pi_{2,0} + \pi_{0,2} = 1$$

yield a unique solution for the steady-state probabilities: $\pi_{0,0}, \pi_{1,0}, \pi_{2,0}$, and $\pi_{0,2}$.

This shows that this 4-state multi-service system is reversible. As the case is with the M/M/$k$/$k$ system, the physical interpretation of the reversibility property includes the lost calls. For the system in the forward direction we have multiple of Poisson processes for different type (classes) of calls, and for the system in the reversed direction, we will also have the same processes if we include as output (input in reverse) the lost calls.

The reversibility property applies also to the general case of a multi-service system, so it is sufficient to solve the detailed balance equations together with the normalizing equation to obtain the steady-state probabilities of the process.

Having obtained the steady-state probability, we can obtain the blocking probability for the voice and for the video calls. Notice that the voice calls are only blocked when the system is completely full. Therefore the voice blocking probability is:

$$\pi_{2,0} + \pi_{0,2}.$$

However, the video calls are blocked also when there is only one channel free. Therefore, the video blocking probability is

$$\pi_{1,0} + \pi_{2,0} + \pi_{0,2}.$$

Actually, in our example, the video calls can only access in state (0,0), so the video blocking probability is also given by

$$1 - \pi_{0,0}.$$

**Homework 14.2**

Compute the blocking probability of the voice calls and of the video calls for the above small example with $I = 2$ and $k = 2$.

**Answer**

Voice blocking probability = 0.425
Video blocking probability = 0.7 □

## 14.4 Other Reversibility Criteria for Markov Chains

We have realized the importance of the reversibility property in simplifying steady-state solutions of Markov chains, where we can solve the simpler detailed balance equations and avoid the complexity of solving the global balance equations. It is therefore important to know ways that we can identify if a continuous-time Markov chain is reversible. We use here the opportunity that being presented by considering the multi-service model which is an example of the more general multi-dimensional Markov chain to discuss useful reversibility criteria for general Markov chains that have a wider scope of applicability which goes beyond multi-service systems.

Note that all the discussion on continuous-time Markov chains has analogy in discrete-time Markov chains. However, we focus here on stationary, irreducible and aperiodic continuous-time Markov chains which is the model used for multi-service systems and is also relevant to many other models in this book. Accordingly, whenever we mention a continuous-time Markov chain in this chapter, we assume that it is stationary, irreducible and aperiodic.

We already know that if the detailed balance equations together with the normalizing equation have a unique solution for the steady-state probabilities, the process is reversible. Here we describe other ways to identify if a process is reversible. However, before discussing specific reversibility criteria, we shall introduce several relevant graph theory concepts (as in [50]) to help visualize relationship associated with probability flux balances.

Consider a graph $G = G(V, E)$ where $V$ is the set on vertices and $E$ is the set of edges. We associate $G$ with a continuous-time Markov-chain as follows. Let the set $V$ represent the set of states in the continuous-time Markov chain. The graph $G$ will have an edge between nodes $\mathbf{x}$ and $\mathbf{y}$ in $G$ if there is an edge between the two states in the corresponding continuous-time Markov-chain, i.e., if there is positive rate either from state $\mathbf{x}$ to state $\mathbf{y}$, and/or from state $\mathbf{y}$ to state $\mathbf{x}$, in the corresponding continuous-time Markov-chain. We consider only cases that the continuous-time Markov chain is irreducible, therefore the corresponding graph must be connected [50]. We define a *cut* in the graph $G$ as a division of $G$ into two mutually exclusive set of nodes $A$ and $\bar{A}$ such that $A \cap \bar{A} = G$.

From the global balance equations, it can be shown (see [50] for details) that for stationary continuous-time Markov chain the probability flux across a cut in one way is equal to the probability flux in the opposite way.

Now it is easy to show that a continuous-time Markov chain that its graph is a tree must be reversible. Because in this case, every edge in $G$ is a cut and therefore the probability flux across any edge must be balanced. As a result, the detailed balanced equations hold.

Notice that all the reversible processes that we have discussed so far, including single dimension birth-and-death processes, such as the queue size processes of M/M/1, M/M/$\infty$, and M/M/$k$/$k$, and the process associated with the above discussed multi-service example with $I = 2$ and $k = 2$ are all trees. We can therefore appreciate that the tree criterion of reversibility is applicable to many useful processes. However, there are many reversible continuous-time Markov chains that are not trees and there is a need for further criteria to identify reversibility.

One important class of reversible processes is the general multi-service problem with any finite $I$ and $k$. We have already demonstrated the reversibility property for the small example with $I = 2$ and $k = 2$ that its associated graph is a tree. Let us now consider a slightly larger example where $k$ is increased from 2 to 3. All other parameter values are as before: $I = 2$, $\lambda_1 = 0.3$, $1/\mu_1 = 3$, $\lambda_2 = 0.2$, and $1/\mu_2 = 5$. The associated graph of this multi-service problem is no longer a tree, but we already know that it is reversible because the general queue size process(es) of the multi-service model is reversible.

The detailed balance equations of this multi-service problem are:

$$\begin{aligned}
(i+1)\mu_1\pi_{i+1,0} &= \lambda_1\pi_{i,0}, \qquad i = 0,1,2. \\
\mu_1\pi_{1,2} &= \lambda_1\pi_{0,2} \\
\mu_2\pi_{0,2} &= \lambda_2\pi_{0,0} \\
\mu_2\pi_{1,2} &= \lambda_2\pi_{1,0}.
\end{aligned}$$

Because the reversibility property applies to the general case of a multi-service system, it is sufficient to solve the detailed balance equations together with the normalizing equation

$$\pi_{0,0} + \pi_{1,0} + \pi_{2,0} + \pi_{3,0} + \pi_{0,2} + \pi_{1,2} = 1.$$

This yields a unique solution for the steady-state probabilities: $\pi_{0,0}, \pi_{1,0}, \pi_{2,0}, \pi_{3,0}, \pi_{0,2}$ and $\pi_{1,2}$.

Having obtained the steady-state probability, we can obtain the blocking probability for the voice and for the video calls. As in the previous case, the voice calls are only blocked when the system is completely full. Therefore, the voice blocking probability is:

$$\pi_{3,0} + \pi_{1,2}$$

and as the video calls are blocked also when the there is only one channel free, the blocking probability of the video calls is

$$\pi_{3,0} + \pi_{2,0} + \pi_{0,2} + \pi_{1,2}.$$

The associated graph of the continuous-time Markov-chain that represents this multi-service problem with $k = 3$ is not a tree, but this Markov-chain is still reversible. Using this example with $k = 3$, we will now illustrate another criterion for reversibility called **Kolmogorov criterion** that applies to a general continuous-time Markov-chain and not only to those that their associated graphs are trees. Graphs that are not trees, by definition have cycles and this criterion is based on conditions that apply to every cycle in the graph that represents the Markov chain. Furthermore, the Kolmogorov criterion has the desired feature that it establishes the reversibility property directly from the given transition rates without the need to compute other results, such as steady-state probabilities.

To establish the Kolmogorov criterion, let $i$ and $j$ be two neighboring states in a continuous-time Markov chain and define $R(i, j)$ as the transition rate from state $i$ to state $j$. The following Theorem is known as the Kolmogorov criterion.

A stationary continuous-time Markov chain is reversible if and only if for any cycle defined by the following finite sequence of states $i_1, i_2, i_3, \ldots, i_n, i_1$ its transition rates satisfy:

$$\begin{aligned} R(i_1, i_2)R(i_2, i_3) \quad \ldots, &\quad R(i_{n-1}, i_k)R(i_n, i_1) \\ = &\quad R(i_1, i_n)R(i_n, i_{n-1}) \ldots R(i_3, i_2)R(i_2, i_1). \end{aligned} \tag{424}$$

The Kolmogorov criterion essentially says that a sufficient and necessary condition for a continuous-time Markov chain to be reversible is that for every cycle in the graph associated with the Markov chain, the product of the rates in one direction of the cycle starting in a given state and ending up in the same state is equal to the product of the rates in the opposite direction.

To illustrate the Kolmogorov Criterion, consider in our example with $k = 3$, the circle composed of the states (0,0), (0,2), (1,2) and (1,0). According to the above detailed balance equations, we obtain the following rates in one direction:

$$\begin{aligned} R([0,0],[0,2]) &= \lambda_2 \\ R([0,2],[1,2]) &= \lambda_1 \\ R([1,2],[1,0]) &= \mu_2 \\ R([1,0],[0,0]) &= \mu_1 \end{aligned}$$

and in the opposite direction:

$$\begin{aligned} R([0,0],[1,0]) &= \lambda_1 \\ R([1,0],[1,2]) &= \lambda_2 \\ R([1,2],[0,2]) &= \mu_1 \\ R([0,2],[0,0]) &= \mu_2. \end{aligned}$$

We can see that the product of the rates in one direction (which is $\lambda_1\lambda_2\mu_1\mu_2$) is equal to the product of the rates in the opposite direction.

## 14.5 Computation

One simple method to compute the steady-state probabilities is to set an arbitrary initial value to one of them, to use the detailed balance equations to obtain values for the neighbors, the neighbors' neighbors etc. until they all have values that satisfy the detailed balance equations. Finally normalize all the values.

Having the steady-state probabilities, blocking probability of all classes can be found by adding up, for each class $i$ the steady-state probabilities of all the states where the server occupancy is higher than $k - s_i$.

Let $\pi_{\mathbf{i}}$ be the steady-state probability of the being in state $i$ after the normalization and $\hat{\pi}_{\mathbf{i}}$ the steady-state probability of the being in state $i$ before the normalization. Let $\mathbf{\Psi}$ be the set of all states. Therefore

$$\pi_{\mathbf{i}} = \frac{\hat{\pi}_{\mathbf{i}}}{\sum_{\mathbf{i}\in\mathbf{\Psi}} \hat{\pi}_{\mathbf{i}}}. \tag{425}$$

To illustrate this approach, let again consider the above example with $I = 2$, $k = 3$, $\lambda_1 = 0.3$, $1/\mu_1 = 3$, $\lambda_2 = 0.2$, and $1/\mu_2 = 5$.

Set $\hat{\pi_{0,0}} = 1$, then

$$\begin{aligned} \hat{\pi_{1,0}} &= \hat{\pi}_{0,0}\frac{\lambda_1}{\mu_1} \\ &= 1 \times \frac{0.3}{1/3} = 0.9. \end{aligned}$$

Next,

$$\begin{aligned} \hat{\pi_{2,0}} &= \hat{\pi}_{1,0}\frac{\lambda_1}{2\mu_1} \\ &= 0.9 \times \frac{0.3}{2/3} = 0.45 \end{aligned}$$

and

$$\begin{aligned} \hat{\pi_{3,0}} &= \hat{\pi}_{2,0}\frac{\lambda_1}{3\mu_1} \\ &= 0.45 \times \frac{0.3}{3/3} = 0.3. \end{aligned}$$

Moving on to the states (0,2) and (1,2), we obtain:

$$\begin{aligned} \hat{\pi_{0,2}} &= \hat{\pi}_{0,0}\frac{\lambda_2}{\mu_2} \\ &= 1 \times \frac{0.2}{1/5} = 1 \end{aligned}$$

and

$$\begin{aligned} \hat{\pi_{1,2}} &= \hat{\pi}_{0,2}\frac{\lambda_1}{\mu_1} \\ &= 1 \times \frac{0.3}{1/3} = 0.9. \end{aligned}$$

To normalize we compute

$$\sum_{\mathbf{i}\in\mathbf{\Psi}} \hat{\pi}_{\mathbf{i}} = 1 + 0.9 + 0.45 + 0.3 + 1 + 0.9 = 4.55$$

Therefore

$$\begin{aligned} \pi_{0,0} &= \frac{1}{4.55} = 0.21978022 \\ \pi_{1,0} &= \frac{0.9}{4.55} = 0.197802198 \\ \pi_{2,0} &= \frac{0.45}{4.55} = 0.098901099 \\ \pi_{3,0} &= \frac{0.3}{4.55} = 0.065934066 \\ \pi_{0,2} &= \frac{1}{4.55} = 0.21978022 \\ \pi_{1,2} &= \frac{0.9}{4.55} = 0.197802198. \end{aligned}$$

Therefore, the voice blocking probability is:

$$\begin{aligned} B_{voice} &= \pi_{3,0} + \pi_{1,2} \\ &= 0.065934066 + 0.197802198 = 0.263736264 \end{aligned}$$

and the video blocking probability is

$$\begin{aligned} B_{video} &= \pi_{2,0} + \pi_{3,0} + \pi_{0,2} + \pi_{1,2} \\ &= 0.098901099 + 0.065934066 + 0.21978022 + 0.197802198 = 0.582417582. \end{aligned}$$

We can see that reversibility makes it easier to solve for steady-state probabilities. However, if we consider a multi-service system where $I$ and $k$ are very large, it may be challenging to solve the problem in reasonable time.

There are two methods to improve the efficiency of the computation.

1. **The Kaufman Roberts Algorithm:** This algorithm is based on recursion on the number of busy servers. For details on this algorithm see [30, 47, 49, 71, 75].
2. **The Convolution Algorithm:** This algorithm is based on aggregation of traffic streams. In other words, if one is interested in the blocking probability of traffic type $i$, the algorithm successively aggregate by convolution all other traffic types, until we have a problem with $I = 2$, namely, traffic type $i$ and all other types together. Then the problem can be easily solved. For details on this algorithm see [46, 47, 75, 78].

## 14.6 A General Treatment

So far we discussed properties of multi-service systems through simple examples. Now we present general definitions and concepts. For notation convenience, let us consider a slightly different Markov chain than the one we considered above. Previously we considered the state space to represent the number of busy servers (channels) occupied by each of the services (traffic types). Now we will consider a continuous-time Markov chain where the state space represents the number of customers (calls) of each traffic type rather than the number of busy servers. The two approaches are equivalent because at any point in time, the number of channels for service $i$ in the system is a multiplication by a factor of $s_i$ of the number of customers (calls) of service $i$ in the system.

Let $j_i$ be the number of class-$i$ customers in the system for $i = 1, 2, \ldots, I$. Let

$$\overrightarrow{j} = (j_1, j_2, \ldots, j_I)$$

and

$$\overrightarrow{s} = (s_1, s_2, \ldots, s_I).$$

Then

$$\overrightarrow{j}\,\overrightarrow{s} = \sum_{i=1}^{I} j_i s_i$$

is the number of busy servers. Now we consider an $I$-dimensional continuous-time Markov-chain where the state space is defined by all feasible vectors $\overrightarrow{j}$ each of which represents a multi-dimensional possible state of the system. In particular, we say that state $\overrightarrow{j}$ is feasible if

$$\overrightarrow{j}\,\overrightarrow{s} = \sum_{i=1}^{I} j_i s_i \leq k.$$

Let $\mathbf{F}$ be a set of all feasible vectors $\overrightarrow{j}$.

A special case of this multi-service model is the M/M/$k$/$k$ model where $I = 1$. If we consider the M/M/$k$/$k$ model and let $k \to \infty$, we obtain the M/M/$\infty$ model described in Section 7. Accordingly, the M/M/$\infty$ model is the special case ($I = 1$) of the multi-service model with $k = \infty$. In our discussion in Section 8.2, the distribution of the number of customers in an M/M/$k$/$k$ model, given by (319) is a truncated version of the distribution of the number of customers in an M/M/$\infty$ model. As we explain there, the former distribution can be obtained using the latter by truncation.

In a similar way, we begin by describing a multi-service system with an infinite number of servers. Then using truncation, we derive the distribution of the number of customers of each class for a case where the number of servers is finite.

### 14.6.1 Infinite number of servers

For the case $k = \infty$, every arrival of any class $i$ customer can always find $s_i$ free servers, therefore this case can be viewed as $I$ independent uni-dimensional continuous-time Markov-chains, where $X_i(t), i = 1, 2, \ldots I$, represents the evolution of the number of class-$i$ customers in the system and characterized by the birth-rate $\lambda_i$ and the death-rate $j_i\mu_i$.

Let $\pi_i(j_i)$ be the steady-state probability of the process $X_i(t), i = 1, 2, \ldots I$ being in state $j_i$. Then $\pi_i(j_i)$ satisfy the following steady-state equations:

$$\lambda_i \pi_i(0) = j_i \mu_i \pi_i(1)$$
$$\lambda_i \pi_i(j_i) = j_i \mu_i \pi_i(j_i + 1) \text{for} \quad j_i = 1, 2, 3, \ldots$$

and the normalizing equation

$$\sum_{j_i=0}^{\infty} \pi_i(j_i) = 1.$$

These equations are equivalent to the equations that represent the steady-state equations of the M/M/$\infty$ model. Replacing $n$ for $j_i$, $\lambda$ for $\lambda_i$, and $\mu$ for $j_i\mu_i$ in the above equations, we obtain the M/M/$\infty$ steady-state equations. This equivalence has also a physical interpretation. Simply consider a group of $s_i$ servers as a single server serving each class-$i$ customer. Following the derivations is Section 7 for the M/M/$\infty$ model, we obtain:

$$\pi_i(j_i) = \frac{e^{-A_i} A_i^{j_i}}{j_i!} \text{ for } j_i = 0,\ 1,\ 2,\ \ldots\ . \tag{426}$$

Since the processes $X_i(t),\ i = 1, 2, \ldots I$, are independent, the probability $p(\overrightarrow{j}) = p(j_1, j_2, \ldots, j_I)$ that in steady-state $X_1(t) = j_1$, $X_2(t) = j_2$, …, $X_I(t) = j_I$, is given by

$$p(\overrightarrow{j}) = p(j_1, j_2, \ldots, j_I) = \prod_{i=1}^{I} \frac{e^{-A_i} A_i^{j_i}}{j_i!} e^{-A_i}. \tag{427}$$

The solution for the steady-state joint probability distribution of a multi-dimensional process, where it is obtained as a product of steady-state distribution of the individual single-dimensional processes, such as the one given by (427), is called a *product-form solution*.

An simple example to illustrate the product-form solution is to consider a two-dimensional multi-service loss system with $k = \infty$, and to observe that to satisfy the detailed balance equations, the steady-state probability of the state $(i, j)$ $\pi_{ij}$ is the product of

$$\pi_{0j} = \pi_{00} \frac{A_2^j}{j!}$$

and

$$\frac{A_1^i}{i!}.$$

Then realizing that

$$\pi_{00} = \pi_0(1)\pi_0(2)$$

where $\pi_0(1)$ and $\pi_0(2)$ are the probabilities that the independent systems of services 1 and 2 are empty, respectively. Thus,

$$\begin{aligned}
\pi_{ij} &= \pi_{00} \frac{A_1^i}{i!} \frac{A_2^j}{j!} \\
&= \pi_0(1)\pi_0(2) \frac{A_1^i}{i!} \frac{A_2^j}{j!} \\
&= \left(\pi_0(1) \frac{A_1^i}{i!}\right) \left(\pi_0(2) \frac{A_2^j}{j!}\right)
\end{aligned}$$

and the product form has directly been obtained from the detailed balanced equations. This illustrates the relationship of reversibility and product form solution.

Next, we consider a system with a finite number of servers, observe that for such a system the detailed balance equations also gives a product form solution because the equation

$$\pi_{ij} = \pi_{00} \frac{A_1^i}{i!} \frac{A_2^j}{j!}$$

which results directly from the detailed balance equation still holds. Note that $\pi_{00}$ is not the same in the infinite and finite $k$ cases, and it is normally different for different $k$ values.

**Homework 14.3**

Provide an example where $\pi_{00}$ is the same for different $k$ values. $\square$

### 14.6.2 Finite Number of Servers

Consider a multi-service system model where the number of servers is limited to $k$. We are interested in the probability $B(m)$ that a class $m$ customer is blocked. We begin by deriving the state probability vector $p(\overrightarrow{j})$ for all $\overrightarrow{j} \in \mathbf{F}$. By the definition of conditional probability, $p(\overrightarrow{j})$ conditional on $\overrightarrow{j} \in \mathbf{F}$ is given by

$$p(\overrightarrow{j}) = p(j_1, j_2, \ldots, j_I) = \frac{1}{C} \prod_{i=1}^{I} \frac{e^{-A_i} A_i^{j_i}}{j_i!} \quad \overrightarrow{j} \in \mathbf{F} \tag{428}$$

where

$$C = \sum_{\overrightarrow{j} \in \mathbf{F}} \prod_{i=1}^{I} \frac{e^{-A_i} A_i^{j_i}}{j_i!}.$$

**Homework 14.4**

Derive (428) by truncating (427).

**Guide**

Consider the steady-state probability distribution of $\overrightarrow{j}$ for the case $k = \infty$ give by (427). Then set $p(\overrightarrow{j}) = 0$ for all $\overrightarrow{j}$ not in $\mathbf{F}$, and normalize the probabilities $\overrightarrow{j} \in \mathbf{F}$ by dividing by them by the probability that the infinite server process is in a feasible state considering that the number of servers $k$ is finite. Then cancel out the exponentials and obtain (428). $\square$

Let $\mathbf{F}(m)$ be the subset of the states in which an arriving class $m$ customer will not be blocked. That is

$$\mathbf{F}(m) = \{\overrightarrow{j} \in \mathbf{F} \text{ such that } \sum_{i=1}^{I} s_i j_i \leq k - s_m\}. \tag{429}$$

Then

$$B(m) = 1 - \sum_{\overrightarrow{j} \in \mathbf{F}(m)} p(\overrightarrow{j}),\ m = 1, 2, \ldots, I. \tag{430}$$

Therefore, by (428), we obtain

$$B(m) = 1 - \frac{\displaystyle\sum_{\overrightarrow{j} \in \mathbf{F}(m)} \prod_{i=1}^{I} \frac{A_i^{j_i}}{j_i!}}{\displaystyle\sum_{\overrightarrow{j} \in \mathbf{F}} \prod_{i=1}^{I} \frac{A_i^{j_i}}{j_i!}}. \tag{431}$$

**Homework 14.5**

Consider the case with $k = 3$, $s_1 = 1$, $s_2 = 2$, $\lambda_1 = \lambda_2 = 1$, and $\mu_1 = \mu_2 = 1$. Find the Blocking probabilities $B(1)$ and $B(2)$.

**Guide**

Let $(i, j)$ be the state in which there are $i$ class 1 and $j$ class 2 customers in the system.

The Set $\mathbf{F}$ in this example is given by

$$\mathbf{F} = \{(0,0),\ (0,1),\ (1,0),\ (1,1),\ (2,0),\ (3,0)\}.$$

Write and solve the steady-state equations for the steady-state probabilities of the states in the set $\mathbf{F}$. Alternatively, you can use (428).

Then

$$\mathbf{F}(1) = \{(0,0),\ (0,1),\ (1,0),\ (2,0)\}.$$

and

$$\mathbf{F}(2) = \{(0,0),\ (1,0)\}.$$

Use (431) to obtain the blocking probability. □

## 14.7 Critical Loading

As discussed in Section 8.9, a critically loaded system is a one where the offered traffic load is equal to the system capacity. Accordingly, in a critically loaded multi-service loss system, the following condition holds

$$\sum_{i=1}^{I} A_i = k. \tag{432}$$

Given the tremendous increase in capacity of telecommunications networks and systems and in the number of human and non-human users of the Internet, the case of large $k$ is of a special interest. As we have learnt in the case of M/M/$k$/$k$ when the total capacity of the system is very large relative to the capacity required by any individual user, critical loading is an efficient dimensioning rule. The result for the asymptotic behavior of the blocking probability under critical loading condition can be extended to the case of a multi-service loss system as follows:

$$\lim_{k\to\infty} B(i) = \frac{s_i C_{MS}}{\sqrt{k}}, \quad i = 1, 2, \ \ldots, \ I \tag{433}$$

where $C_{MS}$ is a constant independent of $i$ and $k$. Notice that if there is only one class ($I = 1$) and $s_1 = 1$, this asymptotic result reduces to (333) by setting $C_{MS} = \tilde{C}$. Notice that as in the special case of M/M/$k$/$k$, the asymptotic blocking probability decays at the rate of $1/\sqrt{k}$, and also notice that the asymptotic class $i$ blocking probability is linear with $s_i$. This means that in the limit, if each of class 1 customers requires one server and each of the class 2 customers requires two servers, then a class 2 customer will experience twice the blocking probability experienced by a class 1 customer. Recall that, in this case, a class 1 customer requires only one server to be idle for it to be able to access a server and to obtain service, while a class 2 customer requires two idle servers to obtain service otherwise, according to our multi-service loss model, it is blocked and cleared from the system.

**Homework 14.6**

Consider the case $\lambda_1 = 1$, $s_1 = 1$, $\mu_1 = 1$, $\lambda_2 = 2$, $s_1 = 2$, $\mu_2 = 2$, $k = 4$. Obtain the blocking probability of each class in two ways: (1) by a discrete event simulation, (2) by solving the steady-state equations or (428) and using Eq. (430), and (3) by using the recursive algorithm. □

**Homework 14.7**

Provide examples where customers that belong to different class experience the same blocking probability. Verify the equal blocking probability using (428), by the recursive algorithm. and by simulations.

**Guide**

One example is with $k = 6$, and two classes of customers $s_1 = 6$ and $s_2 = 5$. Provide other examples and verify the equal blocking probability using the analysis that leads to (430) and simulations. □

**Homework 14.8**

Demonstrate by simulations the robustness of the multi-service loss model to the shape of the holding time distribution.

**Guide**

Simulate various multi-service loss systems with exponential holding time versus equivalent systems where the holding times distributions are hyper-exponential (the variance is larger than exponential), deterministic (where the variance is equal to zero), and Pareto (choose cases where the valiance is finite). Demonstrate that the blocking probability for each class is the same when the mean holding time is the same regardless of the choice of the holding time distribution. □

**Homework 14.9**

Study and program the convolution algorithm described in [46, 47, 75]. Also write a program for the recursion algorithm and for the method based on (428). For a given (reasonably large) problem, compute the blocking probability for each class. Make sure it is the same for all three alternatives. Then compare for a wide range of parameter values the running times of the various algorithms and explain the differences. □

**Homework 14.10**

Provide an example of a continuous-time Markov chain that represent a queueing model that is not reversible.

**Guide**

Consider MMPP(2)/M/1/1 and show cases that the continuous-time Merkov chain is not reversible. □

# 15 Discrete-Time Queue

To complement the considerable attention we have given to continuous-time queues, we will now provide an example of a discrete-time queueing system. Discrete-time models are very popular studies of computer and telecommunications systems because in some cases, time is divided into fixed length intervals (time-slots) and packets of information called cells are of fixed length, such that exactly one cell can be transmitted during a time-slot. Examples of such cases include technologies, such as ATM and the IEEE 802.6 Metropolitan Area Network (MAN) standard.

Let the number of cells that join the queue at different time-slots be an IID random variable. Let $a_i$ be the probability of $i$ cells joining the queue at the beginning of any time-slot. Assume that at any time-slot, if there are cells in the queue, one cell is served, namely, removed from the queue. Further assume that arrivals occur at the beginning of a time-slot means that if a cell arrives during a time-slot it can be served in the same time-slot.

In this case, the queue size process follows a discrete-time Markov-chain with state-space $\Theta$ composed of all the nonnegative integers, and a Transition Probability Matrix $\mathbf{P} = [P_{ij}]$ given by

$$P_{i,i-1} = a_0 \quad \text{for} \quad i \geq 1 \tag{434}$$

and

$$P_{0,0} = a_0 + a_1$$

$$P_{i,i} = a_1 \quad \text{for} \quad i \geq 1$$

$$P_{i,i+1} = a_2 \quad \text{for} \quad i \geq 0$$

and in general

$$P_{i,i+k} = a_{k+1} \quad \text{for} \quad i \geq 0, k \geq 1. \tag{435}$$

Defining the steady-state probability vector by $\mathbf{\Pi} = [\pi_0, \pi_1, \pi_2, \ldots]$, it can be obtained by solving the steady-state equations:

$$\mathbf{\Pi} = \mathbf{\Pi}\mathbf{P}.$$

together with the normalizing equation

$$\sum_{i=0}^{\infty} \pi_i = 1.$$

To solve for the $\pi_i$s, we will begin by writing down the steady-state equations as follows

$\pi_0 = \pi_0 P_{00} + \pi_1 P_{10}$

$\pi_1 = \pi_0 P_{01} + \pi_1 P_{11} + \pi_2 P_{21}$

$\pi_2 = \pi_0 P_{02} + \pi_1 P_{12} + \pi_2 P_{22} + \pi_3 P_{32}$

$\pi_3 = \pi_0 P_{03} + \pi_1 P_{13} + \pi_2 P_{23} + \pi_3 P_{33} + \pi_4 P_{43}$

and in general

$\pi_n = \sum_{i=0}^{n+1} \pi_i P_{i,n}$ for $n \geq 0$.

Substituting (434) and (435) in the latter, we obtain

$$\pi_0 = \pi_0[a_0 + a_1] + \pi_1 a_0 \tag{436}$$

$$\pi_1 = \pi_0 a_2 + \pi_1 a_1 + \pi_2 a_0 \tag{437}$$

$$\pi_2 = \pi_0 a_3 + \pi_1 a_2 + \pi_2 a_1 + \pi_3 a_0 \tag{438}$$

and in general

$$\pi_n = \sum_{i=0}^{n+1} \pi_i a_{n+1-i} \quad \text{for} \quad n \geq 1. \tag{439}$$

Defining $\Pi(z)$ the Z-transform of the $\mathbf{\Pi}$ vector and $A(z)$ as the Z-Transform of $[a_0, a_1, a_2, \ldots]$, multiplying the $n$th equation of the set (436) – (439) by $z^n$, and summing up, we obtain after some algebraic operations

$$\Pi(z) = \pi_0 a_0 - \pi_0 z^{-1} a_0 + z^{-1} A(z) \Pi(z) \tag{440}$$

which leads to

$$\Pi(z) = \frac{\pi_0 a_0 (1 - z^{-1})}{1 - z^{-1} A(z)}. \tag{441}$$

Then deriving the limit of $\Pi(z)$ as $z \to 1$ by applying L'Hopital rule, denoting $A'(1) = \lim_{z\to 1} A'(z)$, and noticing that $\lim_{z\to 1} \Pi(z) = 1$ and $\lim_{z\to 1} A(z) = 1$, we obtain,

$$\pi_0 = \frac{1 - A'(1)}{a_0}. \tag{442}$$

This equation is somewhat puzzling. We already know that the proportion of time the server is idle must be equal to one minus the utilization. We also know that $A'(1)$ is the mean arrival rate of the number of arrivals per time-slot and since the service rate is equal to one, $A'(1)$ is also the utilization; so what is wrong with Eq. (442)? The answer is that nothing wrong with it. What we call $\pi_0$ here is not the proportion of time the server is idle. It is the probability that the queue is empty at the slot boundary. There may have been one cell served in the previous slot and there may be an arrival or more in the next slot which keep the server busy.

The proportion of time the server is idle is in fact $\pi_0 a_0$ which is the probability of empty queue at the slot boundary times the probability of no arrivals in the next slot, and the consistency of Eq. (442) follows.

**Homework 15.1**

Provide in detail all the algebraic operations and the application of L'Hopital rule to derive equations (440), (441) and (442).

**Guide**

Multiplying the $n$th equation of the set (436) – (439) by $z^n$ and summing up, we obtain an equation for $\Pi(z)$ by focussing first on terms involving $\pi_0$ then on the remaining terms. For the remaining terms, it is convenient to focus first on terms involving $a_0$ then on those involving $a_1$, etc. Notice in the following that all the remaining terms can be presented by a double summation.

$$\begin{aligned}\Pi(z) &= \pi_0 a_0 z^0 + \pi_0 \sum_{i=1}^{\infty} a_i z^{i-1} + \sum_{j=0}^{\infty}\left[a_j \sum_{i=1}^{\infty} \pi_i z^{i-(1-j)}\right] \\ &= \pi_0 a_0 + \pi_0 z^{-1}[A(z) - a_0] + z^{-1}A(z)[\Pi(z) - \pi_0] \\ &= \pi_0 a_0 - \pi_0 z^{-1} a_0 + z^{-1}A(z)\Pi(z)\end{aligned}$$

and (441) follows.

L'Hopital rule says that if functions $a(x)$ and $b(x)$ satisfy $\lim_{x\to l^*} a(x) = 0$ and $\lim_{x\to l^*} b(x) = 0$, then

$$\lim_{x\to l^*} \frac{a(x)}{b(x)} = \frac{\lim_{x\to l^*} a(x)}{\lim_{x\to l^*} b(x)}.$$

Therefore, from (441) we obtain

$$\begin{aligned}\lim_{z\to 1} \Pi(z) &= \lim_{z\to 1} \frac{\pi_0 a_0 (1 - z^{-1})}{1 - z^{-1}A(z)} \\ &= \lim_{z\to 1} \frac{\pi_0 a_0 z^{-2}}{z^{-2}A(z) - z^{-1}A'(z)}.\end{aligned}$$

Substituting $\lim_{z\to 1} \Pi(z) = 1$ and $\lim_{z\to 1} A(z) = 1$, we obtain,

$$1 = \frac{\pi_0 a_0}{1 - A'(z)}$$

and (442) follows. □

**Homework 15.2**

Derive the mean and variance of the queue size using the Z-transform method and verify your results by simulations over a wide range of parameter values using confidence intervals. □

# 16 M/G/1

The M/G/1 queue is a generalization of the M/M/1 queue where the service time is no longer exponential. We now assume that the service times are IID with mean $1/\mu$ and standard deviation $\sigma_s$. The arrival process is assumed to be Poisson with rate $\lambda$ and we will use the previously defined notation: $\rho = \lambda/\mu$. As in the case of M/M/1 we assume that the service times are independent and are independent of the arrival process. In addition to M/M/1, another commonly used special case of the M/G/1 queue is the M/D/1 queue where the service time is deterministic.

The generalization from M/M/1 to M/G/1 brings with it a significant increase in complexity. No longer can we use the Markov-chain structure that was so useful in the previous analyzes where both service and inter-arrival times are memoryless. Without the convenient Markov chain structure, we will use different methodologies as described in this section.

## 16.1 Pollaczek Khintchine Formula: Residual Service Approach [13]

The waiting time in the queue of an arriving customer to an M/G/1 queue is the remaining service time of the customer in service plus the sum of the service times of all the customers in the queue ahead of the arriving customer. Therefore, the mean waiting time in the queue is given by

$$E[W_Q] = E[R] + \frac{E[N_Q]}{\mu} \tag{443}$$

where $E[R]$ denotes the mean residual service time. Note that for M/M/1, $E[R] = \rho/\mu$, which is the probability of having one customer in service, which is equal to $\rho$, times the mean residual service time of that customer, which is equal to $1/\mu$ due to the memoryless property of the exponential distribution, plus the probability of having no customer in service (the system is empty), which is $1 - \rho$, times the mean residual service time if there is no customer in service, which is equal to zero.

#### Homework 16.1

Verify that Eq. (443) holds for M/M/1. □

Observe that while Equation (443) is based on considerations at time of arrival, Little's formula

$$E[N_Q] = \lambda E[W_Q]$$

could be explained based on considerations related to a point in time when a customer leaves the queue and enters the server. Recall the intuitive explanation of Little's formula in Section (3) which can be applied to a system composed of the queue excluding the server. Consider a customer that just left the queue leaving behind on average $E[N_Q]$ customers that have arrived during the customer's time in the system which is on average $\lambda E[W_Q]$.

By Little's formula and (443), we obtain,

$$E[W_Q] = \frac{E[R]}{1 - \rho}. \tag{444}$$

It remains to obtain $E[R]$ to obtain results for the mean values of waiting time and queue-size.

Now that as the service time is generally distributed, we encounter certain interesting effects. Let us ask ourselves the following question. If we randomly inspect an M/G/1 queue, will the mean remaining (residual) service time of the customer in service be longer or shorter than the mean service time? A hasty response may be: shorter. Well, let us consider the following example. There are two types of customers. Each of the customers of the first type requires $10^6$ service units, while each of the customers of the second type requires $10^{-6}$ service units. Assume that the proportion of the customers of the first type is $10^{-7}$, so the proportion of the customers of the second type is $1 - 10^{-7}$. Assume that the capacity of the server to render service is one service unit per time unit and that the mean arrival rate is one customer per time unit. As the mean service time is of the order of $10^{-1}$, and the arrival rate is one, although the server is idle 90% of the time, when it is busy it is much more likely to be busy serving a customer of the first type despite the fact that these are very rare, so the residual service time in this case is approximately $0.1 \times 10^6/2 = 50,000$ which is much longer than the $10^{-1}$ mean service time. Intuitively, we may conclude that the residual service time is affected significantly by the variance of the service time.

Notice that what we have computed above is the unconditional mean residual service time which is our $E[R]$. Conditioning on the event that the server is busy, the mean residual service time will be 10 times longer. We know that if the service time is exponentially distributed, the conditional residual service time of the customer in service has the same distribution as the service time due to the memoryless property of the exponential distribution. Intuitively, we may expect that if the variance of the service time is greater than its exponential equivalence (an exponential random variable with the same mean), then the mean residual service time (conditional) will be longer than the mean service time. Otherwise, it will be shorter. For example, if the service time is deterministic of length $d$, the conditional mean residual service time is $d/2$, half the size of its exponential equivalence.

To compute the (unconditional) mean residual service time $E[R]$, consider the process $\{R(t), t \geq 0\}$ where $R(t)$ is the residual service time of the customer in service at time $t$. And consider a very long time interval $[0, T]$. Then

$$E[R] = \frac{1}{T}\int_0^T R(t)dt. \tag{445}$$

Following [13], let $S(T)$ be the number of service completions by time $T$ and $S_i$ the $i$th service time. Notice that the function $R(t)$ takes the value zero during times when there is no customer in service and jumps to the value of $S_i$ at the point of time the $i$th service time commences. During a service time it linearly decreases with rate of one and reaches zero at the end of a service time. Therefore, the area under the curve $R(t)$ is equal to the sum of the areas of $S(T)$ isosceles right triangles where the side of the $i$th triangle is $S_i$. Therefore, for large $T$, we can ignore the last possibly incomplete triangle, so we obtain

$$E[R] = \frac{1}{T}\sum_{i=1}^{S(T)}\frac{1}{2}S_i^2 = \frac{1}{2}\frac{S(T)}{T}\frac{1}{S(T)}\sum_{i=1}^{S(T)}S_i^2. \tag{446}$$

Letting $T$ approach infinity, the latter gives

$$E[R] = \frac{1}{2}\lambda\overline{S^2} \tag{447}$$

where $\overline{S^2}$ is the second moment of the service time.

By (444) and (447), we obtain

$$E[W_Q] = \frac{\lambda\overline{S^2}}{2(1-\rho)}. \tag{448}$$

Thus, considering (261), we obtain

$$E[D] = \frac{\lambda\overline{S^2}}{2(1-\rho)} + 1/\mu. \tag{449}$$

Using Little's formula and recalling that $\sigma_s^2 = \overline{S^2} - (1/\mu)^2$, Eq. (449) leads to the well known Pollaczek Khintchine formula for the mean number of customers in an M/G/1 system:

$$E[Q] = \rho + \frac{\rho^2 + \lambda^2\sigma_s^2}{2(1-\rho)}. \tag{450}$$

Observe that according to the Pollaczek Khintchine formula, if we have two M/G/1 queueing systems, where they both have the same arrival and service rates rate, but for one the variance of the service rate is higher than that of the other, the one with the higher variance will experience higher mean queue size and delay.

## 16.2 Pollaczek Khintchine Formula: by Kendall's Recursion [52]

Let us now derive (450) in a different way. Letting $q_i$ be the number of customers in the system immediately following the departure of the $i$th customer, the following recursive relation, is obtained.

$$q_{i+1} = q_i + a_{i+1} - I(q_i) \tag{451}$$

where $a_i$ is the number of arrivals during the service time of the $i$th customer, and $I(x)$ is a function defined for $x \geq 0$, taking the value 1 if $x > 0$, and the value 0 if $x = 0$. This recursion was first introduced by Kendall [52], so we will call it Kendall's Recursion. Some call it a "Lindley's type Recursion" in reference to an equivalent recursion for the G/G/1 waiting time in [58]. Along with Little's and Erlang B formulae, and the Pollaczek-Khintchine equation, the Kendall's and Lindley's recursions are key foundations of queueing theory.

To understand the recursion (451), notice that there are two possibilities here: either $q_i = 0$ or $q_i > 0$.

If $q_i = 0$, then the $i + 1$th customer arrives into an empty system. In this case $I(q_i) = 0$ and the number of customers in the system when the $i + 1$th customer leaves must be equal to the number of customers that arrives during the service time of the $i + 1$th customer.

If $q_i > 0$, then the $i + 1$th customer arrives into nonempty system. It starts it service when the $i$th customer leaves. When it starts its service there are $q_i$ customers in the system. Then during its service additional $a_{i+1}$ customers arrive. And when it leaves the system there must be $q_i + a_{i+1} - 1$ (where the '-1' represents the departure of the $i + 1$th customer).

Squaring both sides of (451) and taking expectations, we obtain

$$E[q_{i+1}^2] = E[q_i^2] + E[I(q_i)^2] + E[a_{i+1}^2] - 2E[q_i I(q_i)] + 2E[q_i a_{i+1}] - 2E[I(q_i)a_{i+1}] \tag{452}$$

Notice that in steady-state $E[q_{i+1}^2] = E[q_i^2]$, $I(q_i)^2 = I(q_i)$, $E[I(q_i)^2] = E[I(q_i)] = \rho$ , and that for any $x \geq 0$, $xI(x) = x$, so $q_i I(q_i) = q_i$. Also notice that because of the independence between $a_{i+1}$ and $q_i$, and because (by (95)) the mean number of arrivals during service time in M/G/1 is equal to $\rho$, we obtain in steady-state that $E[I(q_i)a_{i+1}] = \rho^2$ and $E[q_i a_{i+1}] = E[q_i]\rho$. Therefore, considering (452), and setting the steady-state notation $E[a] = E[a_i]$ and $E[Q] = E[q_i]$, we obtain after some algebra

$$E[Q] = \frac{\rho + E[a^2] - 2\rho^2}{2(1-\rho)}. \tag{453}$$

To obtain $E[a^2]$, we notice that by EVVE,

$$Var[a] = E[Var[a \mid S]] + Var[E[a \mid S]] = \lambda E[S] + \lambda^2\sigma_s^2 = \rho + \lambda^2\sigma_s^2 \tag{454}$$

recalling that $S$ is the service time and that $\sigma_s^2$ is its variance. Also recall that $Var[a] = E[a^2] - (E[a])^2$ and since $E[a] = \rho$, we have by Eq. (454) that

$$E[a^2] = Var[a] + \rho^2 = \rho + \lambda^2\sigma_s^2 + \rho^2.$$

Therefore,

$$E[Q] = \frac{2\rho + \lambda^2\sigma_s^2 - \rho^2}{2(1-\rho)} \tag{455}$$

or

$$E[Q] = \rho + \frac{\rho^2 + \lambda^2\sigma_s^2}{2(1-\rho)} \tag{456}$$

which is identical to (450) - the Pollaczek-Khintchine Formula.

**Homework 16.2**

Re-derive the Pollaczek-Khintchine Formula in the two ways presented above with attention to all the details (some of which are skipped in the above derivations). □

## 16.3 Special Cases: M/M/1 and M/D/1

Now let us consider the special case of exponential service time. That is, the M/M/1 case. To obtain $E[Q]$ for M/M/1, we substitute $\sigma_s^2 = 1/\mu^2$ in (450), and after some algebra, we obtain

$$E[Q] = \frac{\rho}{1-\rho} \tag{457}$$

which is consistent with (283).

Another interesting case is the M/D/1 queue in which case we have: $\sigma_s^2 = 0$. Substituting the latter in (450), we obtain after some algebra

$$E[Q] = \frac{\rho}{1-\rho} \times \frac{2-\rho}{2}. \tag{458}$$

Because the second factor of (458), namely $(2-\rho)/2$, is less than one for the range $0<\rho<1$, we clearly see that the mean number of customers in an M/M/1 queue is higher than that of an M/D/1 queue with the same arrival and service rates.

Furthermore the inverse of this factor, given by

$$R(\rho)=\frac{2}{2-\rho}$$

is the ratio of the M/M/1 mean queue size to that of M/D/1 queue size assuming both have the same traffic intensity $\rho$. Noting that $R(\rho$ in monotonically increasing in $\rho$ within the relevant range of $0<\rho<1$, and that $R(0)=1$ and $R(1)=2$, then $R(\rho)$ is bounded between 1 and 2. This implies that due to the service variability in M/M/1 its mean queue size is always higher (but it is never higher by more than 100%) than that of the equivalent M/D/1.

**Homework 16.3**

Show that $E[W_Q]$ (the time spent in the queue but not in service) for M/D/1 is half that of its M/M/1 counterpart assuming that the mean service times in both systems is the same.

**Guide**

This can be done in several ways. Here is one way to show it.

First recall that

$$E[W_Q]=E[D]-1/\mu=E[Q]/\lambda-1/\mu.$$

For M/M/1, by (457),

$$E[W_Q]=\frac{\rho}{(1-\rho)\lambda}-\frac{1}{\mu}=\frac{1}{\mu-\lambda}-\frac{1}{\mu}=\frac{\rho}{\mu-\lambda}.$$

For M/D/1, by (458),

$$E[W_Q]=\frac{\rho(2-\rho)}{2(1-\rho)\lambda}-\frac{1}{\mu}=\frac{\rho}{2(\mu-\lambda)}.$$

Clearly, $E[W_Q]$ for M/D/1 is half that of its M/M/1 counterpart.

Complete all steps. □

## 16.4 Busy Period

We have defined and discussed the concept of busy period in Section 6.9 in the context of the M/M/1 queue. The same analysis applies to the case of the M/G/1 system, and we obtain:

$$E[T_B]=\frac{1}{\mu-\lambda}. \tag{459}$$

What we learn from this is that the mean busy period is insensitive to the shape of the service time distribution. In other words, the mean busy periods of M/M/1 and M/G/1 systems are the same if the mean arrival rate and service rates are the same.

**Homework 16.4**

1. Prove that
$$\frac{E[T_B]}{E[T_B] + E[T_I]}$$
is the proportion of time that the server is busy.
2. Show that Equation (459) also applies to an M/G/1 queue. □

**Homework 16.5**

Consider an M/G/1 queueing system with the following twist. When a new customer arrives at an empty system, the server is not available immediately. The customer then rings a bell and the server arrives an exponentially distributed amount of time with parameter $\zeta$ later. As in M/G/1, customers arrive in accordance with a Poisson process with rate $\lambda$ and the mean service time is $1/\mu$. Service times are mutually independent and independent of the inter-arrival times. Find the mean busy period defined as a continuous period that the server is busy.

**Guide**

Explain and solve the following two equations:
$$\frac{E[T_B]}{E[T_B] + E[T_I]} = \rho = \frac{\lambda}{\mu}$$
and
$$E[T_I] = \frac{1}{\lambda} + \frac{1}{\zeta}.$$
□

**Homework 16.6**

Consider a multiplexer with one output link that serves data at a rate of 2.4 Gb/s. The multiplexer has a very large buffer. Data messages arrive at the multiplexer following a Poisson process at a rate of 12 messages per second. The message sizes have mean 20 Mbytes and standard deviation 15 Mbytes. The arriving messages are processed and transmitted by the output link based on the first come first served principle. Consider the system to be in steady-state and answer the following questions. Choose an appropriate queueing model for this multiplexer system, and clearly justify your choice.

Based on the model you choose, answer the following.

- Find the mean number of messages in the entire multiplexer system (in service and in the queue).

- Find the mean delay of a message (including service-time).

**Solution**

The appropriate model is M/G/1.

**Justification:**

1. The very large buffer justifies the infinite buffer model.
2. The Poisson arrivals are given.
3. Service time is neither exponential nor deterministic, so the general service time assumption is justified.
4. The service policy is given as first come first served.
5. The single output link justifies the single server assumption.

Next,

2.4 Gb/s = 2,400 Mb/s.
Mean message size = 20 Mbytes = $20 \times 8 Mbits$ = 160 Mbits.
Standard deviation of message size = 15 Mbytes = 120 Mbits
Standard deviation of service time = 120 / 2,400 = 0.05 [sec.]

$$\mu = \frac{2,400}{160 \text{ Mbits}} = 15 \text{ messages/sec.} \qquad \lambda = 12 \text{ messages/sec} \qquad \rho = \frac{\lambda}{\mu} = \frac{12}{15} = 0.8$$

Then the mean number of messages in the entire multiplexer system is obtained by:

$$E[Q] = \rho + \frac{\rho^2 + \lambda^2 \sigma_s^2}{2(1-\rho)} = 0.8 + \frac{0.8^2 + 12^2 \times 0.05^2}{2(1-0.8)} = 3.3$$

and the mean delay is obtained by Little's formula as:

$$E[D] = \frac{E[Q]}{\lambda} = \frac{3.3}{12} = 0.275 \text{ [sec.]}.$$

□

**Homework 16.6**

Consider a single server queue with infinite size buffer. There are two traffic streams of messages that arrive at the queue for processing. Each of the two traffic streams follow a Poisson processes: one with arrival rate $\lambda_1 = 0.5$ message per millisecond (ms) and the other with arrival rate $\lambda_2 = 1$ messages per ms. All messages that arrive are served based on the first come first served principle. The messages of the first stream require exponentially distributed service times with mean 0.6 ms and the messages of the second stream require exponentially distributed service times with mean 0.3 ms. Find the mean number of messages (regardless of the stream they belong to) in the system (this includes messages waiting in the queue and in service).

**Solution**

The total arrival rate is obtained by $\lambda = \lambda_1 + \lambda_2 = 0.5 + 1 = 1.5$ messages/ms.

The mean service time is obtained by the Law of iterated Expectation:

$$E[S] = \frac{\lambda_1}{\lambda_1 + \lambda_2} 0.6 + \frac{\lambda_2}{\lambda_1 + \lambda_2} 0.3 = \frac{0.5}{1.5} \times 0.6 + \frac{1}{1.5} \times 0.3 = 0.4 \text{ ms.}$$

Another way is:

$$E[S] = E[0.3 + 0.3X]$$

where $X$ is a Bernoulli random variable with parameter $p = 0.5/1.5 = 1/3$.

Thus,

$$E[S] = E[0.3 + 0.3X] = 0.3 + 0.3E[X] = 0.3 + 0.3 \times \frac{1}{3} = 0.4 \text{ ms.}$$

The variance of a message can be obtained by the Law of Total Variance (or EVVE), as follows

$$\sigma_S^2 = \frac{1}{3} \times (0.6)^2 + \frac{2}{3} \times (0.3)^2 + Var[0.3 + 0.3X]$$

where $X$ is a Bernoulli random variable with parameter $p = 1/3$.

Therefore,

$$Var[0.3 + 0.3X] = (0.3)^2 Var[X] = 0.09 \times p(1-p) = 0.09 \times \frac{1}{3} \times \frac{2}{3} = 0.09 \times \frac{2}{9} = 0.02$$

so

$$\sigma_S^2 = 0.2.$$

Now we use the Pollaczek Khintchine Formula for the mean number of messages in an M/G/1 system:

$$\rho = \lambda E[S] = 1.5 \times 0.4 = 0.6.$$

$$E[Q] = \rho + \frac{\rho^2 + \lambda^2 \sigma_s^2}{2(1-\rho)} = 0.6 + \frac{0.6^2 + (1.5)^2 \times 0.2}{2(1-0.6)} = 1.6125 \text{ messages.}$$

□

# 17 M/G/1 with non-FIFO disciplines

## 17.1 M/G/1-LIFO

The M/G/1-LIFO queue posses similar properties to the M/G/1-PS queue that we discussed in Section 13.2. They are both insensitive to the shape of the service time distribution.

We have already mentioned in Section 6.9 that the queue size process of M/M/1 is the same as that of its M/M/1-LIFO equivalence. Therefore they also have the same mean queue size and delay. Due to the insensitivity of M/G/1-LIFO, the M/M/1 results for the mean queue size, mean delay and queue size distribution are applicable also to M/G/1-LIFO.

Specifically, if we are given an M/G/1-LIFO queue with arrival rate $\lambda$ and mean service rate $1/\mu$, denote $\rho = \lambda/\mu$, then the queue size distribution is given by:

$$\pi_i = \rho^i(1-\rho) \text{ for } i = 0,\ 1,\ 2,\ \ldots\ . \tag{460}$$

The mean queue size is given by

$$E[Q] = \frac{\rho}{1-\rho} \tag{461}$$

and the mean delay is given by

$$E[D] = \frac{1}{\mu - \lambda}. \tag{462}$$

To show why M/G/1-LIFO queue is insensitive and satisfies the above equations, notice that an arriving customer that upon its arrival finds $i$ customers in the system will be served only during time when the system is in state $i+1$. Furthermore, all the customers served when the system is in state $i+1$ will be customers that have arrived when the system is in state $i$. Therefore, the time the system spends in the state $i+1$ comprises exactly the service times of the customers that arrive when the system in state $i$. Now consider a long interval of time $T$. As we denote by $\pi_i$ the proportion of time that the system is in state $i$, then during a long time interval $T$, the mean number of arrivals in state $i$ is $\lambda\pi_i T$ and their total service time is equal to $\lambda\pi_i T(1/\mu) = \rho\pi_i T$. Accordingly, the proportion of time the system is in state $i+1$ is given by

$$\pi_{i+1} = \frac{\rho\pi_i T}{T} = \rho\pi_i.$$

Since the latter holds for $i = 0, 1, 2, \ldots$, then we observe that the queue-size distribution of M/G/1-LIFO obeys the steady-state equations of M/M/1 regardless of the shape of the holding time distribution.

## 17.2 M/G/1 with $m$ priority classes

Let us consider an M/G/1 queueing system with $m$ priority classes. Let $\lambda_j$ and $\mu_j$ be the arrival and service rate of customers belonging to the $j$th priority class for $j = 1, 2, 3, \ldots, m$. The mean service time of customers belonging to the $j$th priority class is therefore equal to $1/\mu_j$. The second moment of the service time of customers belonging to the $j$th priority class is denoted $\overline{S^2(j)}$. We assume that priority class $j$ has higher priority that priority class $j+1$, so Class 1 represents the highest priority class and Class $m$ the lowest. For each class $j$, the arrival process is assumed to be Poisson with parameter $\lambda_j$, and the service times are assume mutually independent and independent of any other service times of customers belonging to the other classes, and are also independent of any inter-arrival times. Let $\rho_j = \lambda_j/\mu_j$. We assume that $\sum_{j=1}^{m} \rho_j < 1$. We will consider two priority policies: *nonpreemptive* and *preemptive resume*.

## 17.3 Nonpreemptive

Under this regime, a customer in service will complete its service even if a customer of a higher priority class arrive while it is being served. Let $E[N_Q(j)]$ and $E[W_Q(j)]$ represent the mean number of class $j$ customers in the queue excluding the customer in service and the mean waiting time of a class $j$ customer in the queue (excluding its service time), respectively. Further let $R$ be the residual service time (of all customers of all priority classes). In similar way we derived (447), we obtain:

$$E[R] = \frac{1}{2}\sum_{j=1}^{m} \lambda_j \overline{S^2(j)}. \tag{463}$$

**Homework 17.1**

Derive Eq. (463). □

As in Eq. (443), we have for the highest priority,

$$E[W_Q(1)] = E[R] + \frac{E[N_Q(1)]}{\mu_1} \tag{464}$$

and similar to (443) we obtain

$$E[W_Q(1)] = \frac{E[R]}{1-\rho_1}. \tag{465}$$

Regarding the second priority, $E[W_Q(2)]$ is the sum of the mean residual service time $E[R]$, the mean time it takes to serve the Class 1 customers in the queue $E[N_Q(1)]/\mu_1$, the mean time it takes to serve the Class 2 customers in the queue $E[N_Q(2)]/\mu_2$, and the mean time it takes to serve all the Class 1 customers that arrives during the waiting time in the queue for the Class 2 customer $E[W_Q(2)]\lambda_1/\mu_1 = E[W_Q(2)]\rho_1$. Putting it together

$$E[W_Q(2)] = E[R] + \frac{E[N_Q(1)]}{\mu_1} + \frac{E[N_Q(2)]}{\mu_2} + E[W_Q(2)]\rho_1. \tag{466}$$

By the latter and Little's formula for Class 2 customers, namely,

$$E[N_Q(2)] = \lambda_2 E[W_Q(2)],$$

we obtain

$$E[W_Q(2)] = \frac{E[R] + \rho_1 E[W_Q(1)]}{1-\rho_1-\rho_2}. \tag{467}$$

By Eqs. (467) and (465), we obtain

$$E[W_Q(2)] = \frac{E[R]}{(1-\rho_1)(1-\rho_1-\rho_2)}. \tag{468}$$

**Homework 17.2**

Show that for $m = 3$,

$$E[W_Q(3)] = \frac{E[R]}{(1 - \rho_1 - \rho_2)(1 - \rho_1 - \rho_2 - \rho_3)}. \tag{469}$$

and that in general

$$E[W_Q(j)] = \frac{E[R]}{(1 - \sum_{i=1}^{j-1} \rho_i)(1 - \sum_{i=1}^{j} \rho_i)}. \quad \square \tag{470}$$

The mean delay for a $j$th priority class customer, denoted $E[D(j)]$, is given by

$$E[D(j)] = E[W_Q(j)] + \frac{1}{\mu_j} \text{ for } j = 1, 2, 3, \ldots, m. \tag{471}$$

**Homework 17.3**

Consider the case of $m = 2$, $\lambda_1 = \lambda_2 = 0.5$ with $\mu_1 = 2$ and $\mu_2 = 1$. Compute the average delay for each class and the overall average delay. Then consider the case of $m = 2$, $\lambda_1 = \lambda_2 = 0.5$ with $\mu_1 = 1$ and $\mu_2 = 2$ and compute the average delay for each class and the overall average delay. Explain the difference between the two cases and draw conclusions. Can you generalize your conclusions? $\square$

## 17.4 Preemptive Resume

In this case an arriving customer of priority $j$ never waits for a customer of a lower priority class (of Class $i$ for $i > j$) to complete its service. Therefore, when we are interested in deriving the delay of a customer of priority $j$, we can ignore all customers of class $i$ for all $i > j$. Therefore the mean delay of a priority $j$ customer satisfies the following equation

$$E[D(j)] = \frac{1}{\mu_j} + \frac{R(j)}{1 - \sum_{i=1}^{j} \rho_i} + E[D(j)] \sum_{i=1}^{j-1} \rho_i \tag{472}$$

where $R(j)$ is the mean residual time of all customers of classes $i = 1, 2, \ldots, j$ given by

$$R(j) = \frac{1}{2} \sum_{i=1}^{j} \lambda_i \overline{S^2(i)}.$$

The first term of Eq. (472) is simply the mean service time of a $j$th priority customer. The second term is the mean time it takes to clear all the customers of priority $j$ or higher that are already in the system when a customer of Class $j$ arrives. It is merely Eq. (444) that gives the mean time of waiting in the queue in an M/G/1 queueing system where we replace $\rho$ of (444) by $\sum_{i=1}^{j} \rho_i$ which is the total traffic load offered by customers of priority $j$ or higher. From the point of view of the $j$th priority customer the order of the customers ahead of it will not affect its mean delay, so we can "mix" all these customers up and consider the system as M/G/1. The first term of Eq. (472) is the mean total work introduced to the system by customers of priorities higher than $j$ that arrive during the delay time of our $j$ priority customer. Notice that we use the $\rho_i$s there because $\rho_i = \lambda_i(1/\mu_i)$ representing the product of the mean rate of customer arrivals and the mean work they bring to the system for each priority class $i$.

Eq. (472) leads to

$$E[D(1)] = \frac{(1/\mu_1)(1-\rho_1) + R(1)}{1-\rho_1}, \tag{473}$$

and

$$E[D(j)] = \frac{(1/\mu_j)\left(1 - \sum_{i=1}^{j} \rho_i\right) + R(j)}{\left(1 - \sum_{i=1}^{j-1} \rho_i\right)\left(1 - \sum_{i=1}^{j} \rho_i\right)}. \tag{474}$$

**Homework 17.4**

Derive Eqs. (473) and (474). □

**Homework 17.5**

Consider a single server queue with two classes of customers: Class 1 and Class 2, where Class 1 customers have preemptive resume priority over Class 2 customers. Class $i$ customer arrivals follow a Poisson process with parameter $\lambda_i$, and their service times are exponentially distributed with mean $1/\mu_i$, $i = 1, 2$.

1. Derive formulae for the mean delay (including service time) of each of the classes.
2. Assume $\mu = \mu_1 = \mu_2$, let $\rho_i = \lambda_i/\mu$, $i = 1, 2$, and assume $\rho_1 + \rho_2 < 1$. Maintain $\rho_1$ and $\rho_2$ fixed and let $\mu$ approach infinity, show that under these conditions, the mean delay of either traffic class approaches zero.
3. Now assume the conditions $\rho_1 < 1$, but $\rho_1 + \rho_2 > 1$, again let $\mu = \mu_1 = \mu_2$ approach infinity and show that under these conditions, the mean delay of traffic Class 1 approaches zero.

**Guide**

For exponentially distributed service times with mean $1/\mu$, we have

$$R(1) = \frac{1}{2}\left(\lambda_1 \frac{2}{\mu}\right) = \frac{\rho_1}{\mu}.$$

$$R(2) = \frac{1}{2}\left(\lambda_1 \frac{2}{\mu} + \lambda_2 \frac{2}{\mu}\right) = \frac{\rho_1 + \rho_2}{\mu}.$$

$$E[D(1)] = \frac{(1/\mu)(1-\rho_1) + \rho_1/\mu}{1-\rho_1} = \frac{1}{\mu(1-\rho_1)}$$

This is not a surprise. It is the mean delay obtained by M/M/1 if all the traffic is of class 1 customers. We can observe clearly that if $\rho_1$ stays fixed and $\mu$ approaches infinity, the mean delay approaches zero. This applies to both 2 and 3.

$$E[D(2)] = \frac{(1/\mu)(1-\rho_1-\rho_2)+R(2)}{(1-\rho_1)(1-\rho_1-\rho_2)}.$$

Substituting $R(2)$, we obtain,

$$E[D(2)] = \frac{(1/\mu)(1-\rho_1-\rho_2)+(1/\mu)(\rho_1+\rho_2)}{(1-\rho_1)(1-\rho_1-\rho_2)} = \frac{1}{\mu(1-\rho_1)(1-\rho_1-\rho_2)}.$$

Now we can observe that if $\rho_1+\rho_2<1$, as $\mu$ approaches infinity, the mean delay also of priority 2, approaches zero. □

The last homework problem solution implies the following. For the M/M/1 with priorities model, if the queues of all priorities are stable, and if the service rate is arbitrarily high, then the mean delay is arbitrarily low regardless of the utilization. Then in such a case, there is no much benefit in implementing priorities. However, if for example, $\rho_1+\rho_2>1$ but $\rho_1<1$, then priority 1 customers clearly benefit from having priority even if the service rate (and also arrival rate) is arbitrarily large. Notice that we have observed similar results for M/M/1 without priorities. Also notice that we consider here a scenario where the service rate is arbitrarily high and the utilization is fixed which means that the arrival rate is also arbitrarily high.

# 18 Queues with General Input

In many situations where there is non-zero correlation between inter-arrival times, the Poisson assumption for the arrival process which makes queueing models amenable to analysis does not apply. In this case, we consider more general single-server queues, such as G/GI/1 and G/G/1, or their finite buffer equivalent G/GI/1/$k$ and G/G/1/$k$. In fact, the performance of a queue can be very different if we no longer assume that inter-arrival times are IID. Consider for example the blocking probability of an M/M/1/$N$ queue as a function of $\rho = \lambda/\mu$, then the blocking probability will gradually increase with $\rho$ and approaches one as $\rho \to \infty$. However, we recall our discussion of the SLB/D/1/$N$ where we demonstrate that we can construct an example of a finite buffer queue where the blocking probability approaches one for an arbitrarily low value of $\rho = \lambda/\mu$.

Note that we have already covered some results applicable to G/G/1. We already know that for G/G/1, the utilization $\hat{U}$ representing the proportion of time the server is busy satisfies $\hat{U} = \lambda/\mu$. We know that G/G/1 is work conservative, and we also know that Little's formula

$$E[Q] = \lambda E[D] \tag{475}$$

is applicable to G/G/1.

## 18.1 Reich's Formula

We would like to introduce here a new and important concept the *virtual waiting time*, and a formula of wide applicability in the study of G/G/1 queues known as *Reich's formula* [11, 21, 69].

The virtual waiting time, denoted $W_q(t)$, is the time that a packet has to wait in the queue (not including its own service) if it arrives at time $t$. It is also known as *remaining workload*; meaning, the amount of work remains in the queue at time $t$ where work is measured in time it needed to be served. We assume nothing about the inter-arrival times or the service process. The latter is considered as an arbitrary sequence representing the workload that each packet brings with it to the system, namely, the time required to serve each packet. For simplicity, we assume that the system is empty at time $t = 0$. Let $W_a(t)$ be a function of time representing the total work arrived during the interval $[0, t)$. Then Reich's formula says that

$$W_q(t) = \sup_{0 \leq s < t} \{W_a(t) - W_a(s) - t + s\}. \tag{476}$$

If the queue is not empty at time $t$, the $s$ value that maximizes the right-hand side of (476) corresponds to the point in time where the current (at time $t$) busy period started. If the queue is empty at time $t$, then that $s$ value is equal to $t$.

**Homework 18.1**

Consider the arrival process and the corresponding service duration requirements in the following Table.

| Arrival time | Service duration (work requirement) | $W_q(t^+)$ | optimal $s$ |
|---|---|---|---|
| 1 | 3 | | |
| 3 | 4 | | |
| 4 | 3 | | |
| 9 | 3 | | |
| 11 | 2 | | |
| 11.5 | 1 | | |
| 17 | 4 | | |

Plot the function $W_q(t)$ for every $t$, $0 \leq t \leq 25$ and fill in the right values for $W_q(t^+)$ and the optimal $s$ for each time point in the Table. □

## 18.2 Queue Size Versus Virtual Waiting Time

Let us now consider the queue size probability function at time $t$ $P(Q_t = n)$, for $n = 0, 1, 2, \ldots$ . Its complementary distribution function is given by $P(Q_t > n)$. Note that for a G/D/1 queue we have [72]

$$Q_t = \lceil W_q(t) \rceil, \tag{477}$$

so if we consider $n$ integer, and consider the service time to be equal to one unit of work, then for a G/D/1 queue we have the following equality for the complementary distribution functions of the virtual waiting time $P(W_q(t) > n)$ and the queue size [72]

$$P(Q_t > n) = P(W_q(t) > n), \text{ for } n = 0, 1, 2, \ldots . \tag{478}$$

The steady state probability $P(Q > x) = \lim_{t\to\infty} P(Q_t > x)$ is often called the queue overflow probability.

## 18.3 Is the Queue Overflow Probability a Good QoS Measure?

In queueing theory, we often use the steady queue overflow probability as a measure of QoS which is clearly appropriate if the arrival process is Poisson. However, it is important to understand that this is not always true. Let us consider, for example, the SLB/D/1 queue and assume that a very large burst of say $10^15$ packets arrive at the SSQ at time 0 and no packet arrive for service afterwards, and further assume that it takes 1 ms to serve each packet. Clearly, many of the packets will suffer a very long delay, but $P(Q > x) = 0$ because in steady state, the queue is empty. We admit that the SLB process is not practical, but it illustrates that point that if we have a queue with very bursty traffic, the queue overflow probability may not be appropriate to represent the QoS perceived by users.

It is appropriate to note here that recently the more modern concept of Quality of Experience (QoE) [27] became popular. While QoS is based on statistics measured in the network, the QoE is based on directly asking customers about their experience. The reader is referred to [27] for comparison between the two.

## 18.4 G/GI/1 Queue and Its G/GI/1/$k$ Equivalent

Let us consider special cases of the G/G/1 and G/G/1/$k$ queues which we call them G/GI/1 and G/GI/1/$k$, respectively. The GI notation indicates that the service times are mutually independent and independent of the arrival process and the state of the queue. We consider two queueing systems: a G/GI/1 queue and a G/GI/1/$k$ queue that are statistically equal in every aspect except for the fact that the first has an infinite buffer and the second has a finite buffer. They both have the same arrival process the distribution of their service times and the relationship of service times to inter-arrival times are all statistically the same.

In queueing theory there are many cases where it is easier to obtain overflow probability estimations of the unlimited buffer queue G/GI/1, namely, the steady-state probability that the queue size $Q$ exceeds a threshold $k$, $P(Q > k)$, than to obtain the blocking probability, denoted $P_{loss}$, of its G/GI/1/$k$ equivalent. In practice, no buffer is of unlimited size, so the more important problem in applications is the blocking probability of a G/GI/1/$k$ queue.

By applying Little's formula on the system defined by the single server we can observe that the mean delay of a G/GI/1/$k$ will be bounded above by the mean delay of its G/GI/1 equivalent. Notice that if we consider only the server system for both systems, we observe that they have the same mean delay (service time) and the one associated with the G/GI/1/$k$ has somewhat lower arrival rate due to the losses. The fact that only part of the customers offered to the G/GI/1 enter the G/GI/1/$k$ equivalent also implies that the percentiles of the delay distribution of the G/GI/1/$k$ system will be lower than those of the G/GI/1 equivalent.

An interesting problem associated with these two equivalent queues is the following. Given $P(Q > k)$ for a G/GI/1 queue, what can we say about the blocking probability of the G/GI/1/$k$ equivalent. Let us begin with two examples. First, consider a discrete-time single-server queueing model where time is divided into fixed-length intervals called slots. This example is a discrete-time version of our earlier example where we demonstrated a case of a finite buffer queue with arbitrarily low traffic and large packet loss. Assume that the service time is deterministic and is equal to a single slot. Let the arrival process be described as follows: $10^9$ packets arrive at the first time-slot and no packets arrived later. Consider the case of $k = 1$. In the finite buffer case with buffer size equal to $k$, almost all the $10^9$ packets that arrived are lost because the buffer can store only one packet. Therefore, $P_{loss} \approx 1$. However, for the case of infinite buffer where we are interested in $P(W_q > k)$, ($W_q = \lim_{t\to\infty} W_q(t)$) the case is completely the opposite. After the $10^9$ time-slots that it takes to serve the initial burst the queue is empty forever, so in steady-state $P(W_q > k) = 0$.

In our second example, on the other hand, consider another discrete-time queueing model with $k = 10^9$ and a server that serves $10^9$ customers – all at once at the end of a time slot – with probability $1 - 10^{-9}$ and $10^{90}$ customers with probability $10^{-9}$. The rare high service rate ensures stability. Assume that at a beginning of every time-slot, $10^9 + 1$ customers arrive at the buffer. This implies that one out of the arriving $10^9 + 1$ customers is lost, thus $P_{loss} \approx 10^{-9}$, while $P(W_q > k) \approx 1$. We conclude that $P_{loss}$ and $P(W_q > k)$ can be very different.

Wong [92] considered this problem in the context of an ATM multiplexer fed by multiple deterministic flows (a queueing model denoted N*D/D/1 and its finite buffer equivalent) and obtained the following inequality.

$$\rho P_{loss} \leq P(Q > k) \tag{479}$$

Roberts et al. [72] argued that it can be generalized to G/D/1 and its G/D/1/$k$ equivalent. This can be further generalized. The arguments are analogous to those made in [92]. Let $\lambda$ be the arrival rate and $\mu$ the service rate in both the G/GI/1 queue and its G/GI/1/$k$ equivalent, with $\rho = \lambda/\mu$. Consider a continuous period of time, in our G/GI/1 queue, during which $Q > k$ and that just before it begins and just after it ends $Q \leq k$, and define such time period as *overflow period*. Since the queue size at the beginning is the same as at the end of the overflow period, the number of customers that joined the queue during an overflow period must be equal to the number of customers served during the overflow period, because the server is continuously busy during an overflow period.

Now consider a G/GI/1/$k$ queue that has the same realization of arrivals and their work requirements as the G/GI/1 queue. Let us argue that in the worst case, the number of lost customers in the G/GI/1/$k$ queue is maximized if all customers that arrive during overflow periods of the equivalent G/GI/1 queue are lost. If for a given G/GI/1 overflow period, not all arriving customers in the G/GI/1/$k$ queue are lost, the losses are reduced from that maximum level without increasing future losses because at the end of a G/GI/1 overflow period, the number of customers in the equivalent G/GI/1/$k$ queue can never be more than $k$.

Consider a long period of time of length $L$, the mean number of lost customers the G/GI/1/$k$ queue during this period of time of length $L$ is $\lambda L P_{loss}$. This must be lower or equal to the number of customers that arrived during the same period of time during the G/GI/1 overflow periods. This must be equal to the number of customers served during that period of time of length $L$ during the G/GI/1 overflow periods which is equal to $\mu L P(Q > k)$.

Therefore,

$$\lambda L P_{loss} \leq \mu L P(Q > k)$$

and (479) follows. □

**Homework 18.2**

Show that (479) applies to an M/M/1 queue and its M/M/1/$N$ Equivalent, and discuss how tight is the bound in this case for the complete range of parameter values.

**Guide**

Recall that for M/M/1/$N$,

$$P_{loss} = \frac{\rho^N(1-\rho)}{1-\rho^{N+1}},$$

and for M/M/1,

$$P(Q > N) = \rho^{N+1}(1-\rho) + \rho^{N+2}(1-\rho) + \rho^{N+3}(1-\rho) + \ldots = \rho^{N+1}. \quad \square$$

**Homework 18.3**

Using the UNIX command *netstat* collect a sequence of 100,000 numbers representing the number of packets arriving recorded every second for consecutive 100,000 seconds. Assume that these numbers represent the amount of work, measured in packets, which arrive at an SSQ during 100,000 consecutive seconds. Write a simulation of an SSQ fed by this arrival process, assume that all the packets are of equal length and compute the Packet Loss Ratio (PLR) for a range of buffer sizes and the overflow probabilities for a range of thresholds. PLRs are relevant in the case of a finite buffer queue and overflow probabilities represent the probability of exceeding a threshold in an infinite buffer queue. Plot the results in two curves one for the PLR and the other for the overflow probabilities times $\rho^{-1}$ and observe and discuss the relationship between the two. □

**Homework 18.4**

Consider the sequence of 100,000 numbers you have collected. Let $E[A]$ be their average. Generate a sequence of 100,000 independent random numbers governed by a Poisson distribution with mean $\lambda = E[A]$. Use your SSQ simulation, and compute the PLR for a range of buffer sizes, and the overflow probabilities for a range of thresholds. Compare your results to those obtained in the previous Assignment, and try to explain the differences. □

**Homework 18.5**

In this exercise the reader is asked to repeat the previous homework assignment for the Bernoulli process. Again, consider the sequence of 100,000 numbers you have collected. Let $E[A]$ be their average. Generate a sequence of 100,000 independent random numbers governed by the Bernoulli distribution with mean $p = E[A]$. Use your SSQ simulation from Exercise 1, and compute the PLR for a range of buffer sizes, and the overflow probabilities for a range of thresholds. Compare your results to those obtained previously, and discuss the differences. □

# 19 Network Models and Applications

So far we have considered various queueing systems, but in each case we have considered a single queueing system in isolation. Very important and interesting models involve networks of queues. One important application is the Internet itself. It may be viewed as a network of queueing systems where all network elements, such as routers and switches, are connected and where the packets are the customers served by the various network elements and are often queued there waiting for service.

Queueing network models can be classified into two groups: (1) open queueing networks, and (2) closed queueing networks. In closed queueing networks the same customers stay in the network all the time. No new customers join and no customer leaves the network. Customers that complete their service in one queueing system goes to another and then to another and so forth, and never leaves the network. In open queueing systems new customers from the outside of the network can join any queue, and when they complete their service in the network obtaining service from an arbitrary number of queueing system they may leave the network. In this section we will only consider open queueing networks.

## 19.1 Jackson Networks

Consider a network of queues. An issue that is very important for such a queueing networks is the statistical characteristics of the output of such queues because in queueing networks, output of one queue may be the input of another.

Burke's Theorem states that, in steady-state, the output (departure) process of M/M/1, M/M/$k$ or M/M/$\infty$ queue follows a Poisson process. Because no traffic is lost in such queues, the arrival rate must be equal to the departure rate, then any M/M/1, M/M/$k$, or M/M/$\infty$ queue with arrival rate of $\lambda$ will have a Poisson departure process with rate $\lambda$ in steady-state.

Having information about the output processes, we will now consider an example of a very simple queueing network made of two identical single-server queues in series, in steady-state, where all the output of the first queue is the input of the second queue and all the customers that complete service at the second queue leave the system. Let us assume that the customers that arrive into the first queue follow a Poisson process with parameter $\lambda$. The service times required by each of the arriving customers at the two queues are independent and exponentially distributed with parameter $\mu$. This means that the amount of time a customer requires in the first queue is independent of the amount of time a customer requires in the second queue and they are both independent of the arrival process into the first queue. Since the output process of the first queue is Poisson with parameter $\lambda$, and since the first queue is clearly an M/M/1 queue, we have here nothing but two identical M/M/1 queues is series. This is an example of a network of queues where Burke's theorem [17] leads immediately to a solution for queue size and waiting time statistics. A class of networks that can be easily analyzed this way is the class of the so-called acyclic networks. These networks are characterized by the fact that a customer never goes to the same queue twice for service.

If the network is not acyclic, the independence between inter arrival times and between inter arrival and service times do not hold any longer. This means that the queues are no longer Markovians. To illustrate this let us consider a single server queue with feedback described as follows. (Normally, a single node does not constitute a network, however, this simple single queue example is the simplest model to illustrate the feedback effect, and it is not too simple, as it can easily be extended to the case of two nodes with feedback.) Customers arrive into the system from the outside according to a Poisson process with parameter $\lambda$ and the service time is exponentially distributed with parameter $\mu$. Then when the customer completes the service the customer returns to the end of the queue with probability $p$, and with probability (1-p) the customer leaves the system. Now assume that $\lambda$ is very small and $\mu$ is very high. Say $p > 0.99$. This results in an arrival process which is based on very infrequent original arrivals (from the outside) each of which brings with it a burst of many feedback arrivals that are very closed to each other. Clearly this is not a Poisson process. Furthermore, the inter-arrivals of packets within a burst, most of which are feedback from the queue output, are very much dependent on the service times, so clearly we have dependence between inter-arrival times and service times.

Nevertheless, the so-called Jackson's Theorem extends the simple result applicable to an acyclic network of queues to networks that are not acyclic. In other words, although the queues are not M/M/1 (or M/M/$k$ or M/M/$\infty$), they behave in terms of their queue-size statistics as if they are.

Jackson's Theorem can be intuitively justified for the case of a single queue with feedback as follows. Let the feedback arrivals have preemptive resume priority over all other arrivals. This priority regime will not change the queue size statistics. Now we have that the service time comprises a geometric sum of exponential random variables which is also exponential. As a result, we have an M/M/1 queue with arrival rate $\lambda$ and service rate $\mu(1-p)$.

Consider a network of $N$ single-server queues with infinite buffer in steady-state. The Jackson theorem also applies to multi-server queues, but let us consider single-server queues for now. For queue $i$, $i = 1, 2, 3, \ldots, N$, the arrival process from the outside is Poisson with rate $r_i$. We allow for $r_i = 0$ for some queues, but there must be at least one queue $j$, such that $r_j > 0$. Once a customer completes its service in queue $i$, it continues to queue $j$ with probability $P_{ij}$, $i = 1, 2, 3, \ldots, N$, or leaves the system with probability $1 - \sum_{j=1}^{N} P_{ij}$. Notice that we allow for $P_{ii} > 0$ for some queues. That is, we allow for positive probability for customers to return to the same queue they just exited.

Let $\lambda_j$ be the total arrival rate into queue $j$. These arrival rates can be computed by solving the following set of equations.

$$\lambda_j = r_j + \sum_{i=1}^{N} \lambda_i P_{ij}, \qquad j = 1, 2, 3, \ldots, N. \tag{480}$$

The above set of equations can be solved uniquely, if every customer eventually leaves the network. This means that the routing probabilities $P_{ij}$ must be such that there is a sequence of positive routing probabilities and a final exit probability that create an exit path of positive probability from each node.

The service times at the $j$th queue are assumed exponentially distributed with parameter $\mu_j$. They are assumed to be mutually independent and also independent of the arrival process at that queue. Let $\rho_j$ be defined by

$$\rho_j = \frac{\lambda_j}{\mu_j} \qquad \text{for } j = 1, 2, 3, \ldots, N. \tag{481}$$

Assume that

$$\rho_j < 1 \qquad \text{for } j = 1, 2, 3, \ \ldots, N.$$

Let $Q_j$ be the queue-size of queue $j$. Then according to Jackson's Theorem, in steady-state, we have that

$$P(Q_1 = k_1, Q_2 = k_2, \ \ldots, \ Q_N = k_N) = P(k_1)P(k_2)P(k_3) \cdot \ \ldots \ \cdot P(k_N), \tag{482}$$

where $P(k_i) = \rho_i^{k_i}(1 - \rho_i), \quad$ for $i = 1, 2, 3, \ \ldots, N$.

**Comment:** Although Jackson theorem assumes that the arrival processes from the outside follow Poisson processes, it does not assume that the input into every queue follows a Poisson processes. Therefore it does not assume that the queues are independent M/M/1 (or M/M/$k$ or M/M/$\infty$) queues. However, it turns out, according to Jackson theorem that the joint probability distribution of the queue sizes of the $N$ queues is obtained as if the queues are independent M/M/1 (or M/M/$k$ or M/M/$\infty$) queues. This result applies despite the fact that the network is cyclic (not acyclic) in which case we have demonstrated that the queues do not have to be M/M/1 queues.

Therefore, the mean queue-size of the $j$th queue is given by

$$E[Q_j] = \frac{\rho_j}{1 - \rho_j}. \tag{483}$$

The mean delay of a customer in the $j$th queue $E[D_j]$ defined as the time from the moment the customer joins the queue until it completes service, can be obtain by Little's formula as follows.

$$E[D_j] = \frac{E[Q_j]}{\lambda_j}. \tag{484}$$

Using Little's formula, by considering the entire queueing network as our system, we can also derive the mean delay of an arbitrary customer $E[D]$:

$$E[D] = \frac{\sum_{j=1}^{N} E[Q_j]}{\sum_{j=1}^{N} r_j}. \tag{485}$$

Let us now consider a network of two-queue in series where all the traffic that completes service in queue 1 enters queue 2 and some of the traffic in queue 2 leaves the system while the rest enters queue 1. This example is similar to the above mentioned example of a single queue with feedback. Using our notation, let the arrivals from the outside follow Poisson processes with rates $r_1 = 10^{-8}$ and $r_2 = 0$ and let $\mu_1 = \mu_2 = 1$. Further assume that the probability that a customer that completes service in queue 2 leaves the system is $10^{-3}$, so it exters queue 1 with probability $1 - 10^{-3}$

Accordingly,

$$\lambda_1 = r_1 + (1 - 10^{-3})\lambda_2$$

and

$$\lambda_2 = \lambda_1.$$

Thus,

$$\lambda_1 = 10^{-8} + (1 - 10^{-3})\lambda_1,$$

so

$$\lambda_1 = \lambda_2 = 10^{-5}$$

and

$$\rho_1 = \rho_2 = 10^{-5},$$

so

$$E[Q_1] = E[Q_2] = \frac{10^{-5}}{1 - 10^{-5}} \approx 10^{-5}$$

and

$$E[D_1] = E[D_2] \approx \frac{10^{-5}}{10^{-5}} = 1.$$

Recalling that the mean service time is equal to one, this means that negligible queueing delay is expected. (The word 'negligible' is used instead of 'zero' because of the approximation $1 - 10^{-5} \approx 1$ made above.) This result makes sense intuitively. Although the feedbacked traffic is more bursty than Poisson we are considering here the same packet that returns over and over again and it is impossible for the same packet to wait in the queue for itself to be served.

An open network of M/M/1, M/M/$k$ or M/M/$\infty$ queues described above is called a Jackson Network. For such network an exact solution is available. However, in most practical cases, especially when we have to deal with the so-called loss networks that comprise queues, such as M/M/$k$/$k$, where traffic is lost, we have to make additional modelling assumptions and to rely on approximations to evaluate performance measures, such as blocking probability, or carried traffic. One approximation is the so-called Reduced-Load Erlang Fixed-Point Approximation which is reasonably accurate and useful for loss networks.

**Homework 19.1**

Consider a 6-node network of M/M/1 queues, the service rate of all the queues is equal to one, i.e., $\mu_i = 1$ for $i = 1, 2, 3, \ldots, 6$. The arrival rates from the outside into the different queues is given by $r_1 = 0.6$, $r_2 = 0.5$, and $r_i = 0$ for $i = 3, 4, 5, 6$. The routing matrix is as follows

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0 | 0.4 | 0.6 | 0 | 0 | 0 |
| 2 | 0 | 0.1 | 0 | 0.7 | 0.2 | 0 |
| 3 | 0 | 0 | 0 | 0.3 | 0.7 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0.6 |
| 5 | 0 | 0 | 0 | 0.3 | 0 | 0.2 |
| 6 | 0 | 0 | 0.3 | 0 | 0 | 0 |

In this routing matrix every row gives the routing probabilities from a specific queue to other queues. The sum of the probabilities of each row is less or equal to 1.

1. Find the mean delay in each of the queues.
2. Find the mean time a packet spends in the network from the moment it enters the network until it leaves the network.
3. Find the probability that the entire network is empty. □

## 19.2 Computation of Blocking Probability in Circuit Switched Networks by Erlang Fixed-Point Approximation

Let us consider a circuit switched network made of nodes (switching centers) that are connected by links. Each link has a fixed number of circuits. In order to make a call between two nodes: source and destination, a user should reserve a free circuit in each consecutive link of a path between the two nodes. Such reservation is successful if and only if there exists a free circuit on each of the links of that path.

An important characteristic of a circuit-switched network is that once a user makes a reservation for a connection between a source and a destination the capacity for this connection is exclusively available for the user of this connection and no other users can utilize this capacity for the entire duration of this connection holding time.

To evaluate the probability that a circuit reservation is blocked, we first make the following simplifying assumptions:

1. all the links are independent,
2. the arrival process of calls for each origin-destination pair is Poisson, and
3. the arrival process seen by each link is Poisson.

Having made these assumptions, we now consider each link as an independent M/M/$k$/$k$ system for which the blocking probability is readily available by the Erlang B formula. In particular, let $a_j$ be the total offered load to link $j$ from all the routes that pass through link $j$. Recall that multiplexing of Poisson processes give another Poisson process which its rate is the sum of the individual rates. Then the blocking probability on link $j$ is obtained by

$$B_j = E_k(a_j). \tag{486}$$

Now that we have means to obtain the blocking probability on each link, we can approximate the blocking probability of a call made on a given route. Let $B(R)$ be the blocking probability of a call made on route $R$. The route $R$ can be viewed as an ordered set of links, so the route blocking probability is given by

$$B(R) = 1 - \prod_{i \in L_R} (1 - B_i). \tag{487}$$

Note that in the above equation, $L_R$ represents the set of links in route $R$.

Let $A(R)$ be the offered traffic on route $R$ and let $a_j(R)$ be the total traffic offered to link $j$ from traffic that flow on route $R$. Then $a_j(R)$ can be computed by deducting from $A(R)$ the traffic lost due to congestion on links other than $j$. That is,

$$a_j(R) = A(R) \prod_{i \in L_R;\ i \neq j} (1 - B_i) \quad \text{for } j \in L_R, \tag{488}$$

and $a_j(R) = 0$ if $j$ is not in $L_R$. This consideration to the reduced load due to blocking on other links gave rise to the name *reduced load approximation* to this procedure.

Then the total offered traffic on link $j$ is obtained by

$$a_j = \sum_{R \in \mathcal{R}} a_j(R) \tag{489}$$

where $\mathcal{R}$ is the set of all routes.

These give a set of nonlinear equations that require a fixed-point solution. Notice that Erlang B is non-linear.

To solve these equations, we start with an initial vector of $B_j$ values, for example, set $B_j = 0$ for all $j$. Since the $A(R)$ values are known, we use equation (488) to obtain the $a_j(R)$ values. Then use equation (489) to obtain the $a_j$ values, which can be substituted in equation (486) to obtain a new set of values for the blocking probabilities. Then the process repeats itself iteratively until the blocking probability values obtained in one iteration is sufficiently close to those obtained in the previous iteration.

The above solution based on the principles of the Reduced-Load and Erlang Fixed-Point Approximations can be applied to many systems and networks. For example, an application is an Optical Burst Switching (OBS) network is described in [74] where bursts of data are moving between OBS nodes each of which is modelled as an M/M/$k$/$k$ system.

We have discussed an approach to evaluate blocking probability for circuit switched networks under the so-called fixed routing regime, where a call is offered to a route, and if it is rejected it is lost and cleared from the system. There are, however, various other regimes involving alternate routing where rejected calls from a given routes can overflow to other routes. A similar Erlang fixed-point approximation can be used for circuit switching with alternate routing. See [34].

## 19.3 Parameter Conversion in Optical Circuit Switched Networks

The classical use of circuit switched networks was for the public switched telephone network (PSTN) that focused on the provision of the plain old telephone service (POTS) which was the telephone voice service of pre-1970 that was based on analog transmission of signals. Nowadays, circuit switching technology is used in optical networks based on WDM technology that carry data at multi-Tb/s rate and provide a wide range of broadband multimedia services. In WDM networks connections may be set up point-to-point or point-to-multipoint, and they may be carried by a single or multiple wavelengths. However, for simplicity, we consider here the case where connections are point-to-point, namely, every connection is between a given source and destination (SD) pair, and every connection used only a single wavelength with capacity of $c$ [Gb/s]. To clarify, we may have multiple connections running simultaneously between a given SD pair, but each of them is carried on a single wavelength on a route between the source and the destination.

In some situations, optical network designer may not have information on offered traffic in [Erlangs] or arrival rates of connection requests and connection holding times. Such input is required to evaluate blocking probability, e.g. using Erlang Fixed-Point approximation. Instead network designers may have measurements on transmitted bit-rate, wavelength utilization during a connection, and the number of bits transmitted during a connection. The following will provide simple equations to obtain offered traffic, arrival rate and mean holding time from the other potentially available information.

Let us focus our discussion on a given SD pair, but the discussion is applied to any SD pair. The connection which is set up between the SD pair uses multiple hops over multiple links which connect a sequence of optical switches, where in each link the connection uses $w$ wavelengths. After a connection is set, the wavelength capacity of $c$ Gb/s is available for data transmission between the SD pair for the entire duration of the connection holding time. This does not mean that the connection fully utilizes its capacity. If the average bit-rate of the connection over its entire holding time is 20 Gb/s and if $c = 40$ Gb/s, then we say that the utilization of this connection is equal to $20/40 = 0.5$.

Let the long-run average bit-rate between the SD pair be $r$ Gb/s and the average utilization of all connections (weighted by their holding times) that are set-up between the SD pair is $U$. In addition, for this SD pair, $A$ is the offered traffic, $h$ is the mean holding time of connections, $\lambda$ is the arrival rate of new connection requests, and $b$ is the average number of Gbits per connection. Then the following relationships hold.

$$\lambda = \frac{r}{b}, \tag{490}$$

$$h = \frac{b}{cU}, \tag{491}$$

and by definition

$$A = \lambda h. \tag{492}$$

Substituting (490) and (491) in (492), we obtain

$$A = \left(\frac{r}{b}\right)\left(\frac{b}{cU}\right) = \frac{r}{cU}. \tag{493}$$

## 19.4 A Markov-chain Simulation of a Mobile Cellular Network

A mobile cellular network can be modelled as a network of M/M/$k$/$k$ systems by assuming that the number of channels in each cell is fixed and equal to $k$, that new call generations in each cell follows a Poisson process, that call holding times are exponentially distributed and that times until handover occurs in each cell are also exponentially distributed. In the following we describe how to simulate such a network.

Variables and input parameters:
$N$ = total of M/M/$k$/$k$ Systems (cells) in the network;
$Q(i)$ = number of customers (queue size) in cell $i$ ;
$B_p$ = estimation for the blocking probability;
$N_a(i)$ = number of customer arrivals counted so far in cell $i$;
$N_b(i)$ = number of blocked customers counted so far in cell $i$;
$MAXN_a$ = maximal number of customers - used as a stopping criterion;
$\mu$ = 1/(the mean call holding time)
$\lambda(i)$ = arrival rate of new calls in cell $i$;
$P(i, j)$ = Matrix of routing probabilities;

$\delta(i)$ = handover rate in cell $i$ per call = 1/(mean time a call stays in cell $i$ before it leaves the cell)
$P_B$ = Blocking probability.
$Neib(i)$ = the set of neighboring cells of cell $i$.
$|Neib(i)|$= number of neighboring cells of cell $i$.

Again, we will repeatedly consider $R(01)$ a uniform $U(0,1)$ random deviate. A new value for $R(01)$ is generated every time it is called.

To know if the next event is an arrival, we use the following **if** statement.

If

$$R(01) \leq \frac{\sum_{i=1}^{N} \lambda(i)}{\sum_{i=1}^{N} \lambda(i) + \sum_{i=1}^{N} Q(i)\mu + \sum_{i=1}^{N} Q(i)\delta(i)}$$

then the next event is an arrival. Else, to find out if it is a departure (it could also be a handover) we use the following **if** statement. If

$$R(01) \leq \frac{\sum_{i=1}^{N} \lambda(i) + \sum_{i=1}^{N} Q(i)\mu}{\sum_{i=1}^{N} \lambda(i) + \sum_{i=1}^{N} Q(i)\mu + \sum_{i=1}^{N} Q(i)\delta(i)}$$

then the next event is a departure; else, it is a handover.

If the next event is an arrival, we need to know in which of the $N$ cells it occurs. To find out, we use the following loop.

For $i = 1$ to $N$, do: If

$$R(01) \leq \frac{\sum_{j=1}^{i} \lambda(j)}{\sum_{j=1}^{N} \lambda(j)},$$

stop the loop. The arrival occurs in cell $i$, so if

$$\sum_{j=1}^{N} N_a(j) = MAXN_a,$$

the simulation ends, so we compute the blocking probabilities as follows.

$$P_B = \frac{\sum_{i=1}^{N} N_b(i)}{MAXN_a}.$$

Else, $N_a(i) = N_a(i) + 1$ and if $Q(i) < k$ then $Q(i) = Q(i) + 1$, else the number of lost calls needs to be incremented, namely, $N_b(i) = N_b(i) + 1$.

If the next event is a departure, we need to know in which of the $N$ cells it occurs. To find out we use the following loop.

For $i = 1$ to $N$, do: If

$$R(01) \leq \frac{\sum_{j=1}^{i} Q(j)\mu}{\sum_{j=1}^{N} Q(j)\mu}.$$

Then stop the loop. The departure occurs in System $i$, so $Q(j) = Q(j) - 1$. Note that we do not need to verify that $Q(j) > 0$ (why?).

If the next event is a handover, we need to know from which of the $N$ cells it handovers out of. To find it out, we use the following loop.

For $i = 1$ to $N$, do: If

$$R(01) \leq \frac{\sum_{j=1}^{i} Q(j)\delta(j)}{\sum_{j=1}^{N} Q(j)\delta(j)}.$$

Then stop the loop. The handover occurs out of cell $i$, so $Q(i) = Q(i) - 1$. Note that again we do not need to verify that $Q(i) > 0$.

Then to find out into which cell the call handover in, we use the following:

For $j = 1$ to $|Neib(i)|$, do: If

$$R(01) \leq \frac{\sum_{k=1}^{j} P(i,k)}{\sum_{k=1}^{|Neib(i)|} P(i,k)},$$

The call handovers into cell $k$.

# 20 Stochastic Processes as Traffic Models

In general, the aim of traffic modelling is to provide the network designer with relatively simple means to characterize traffic load on a network. Ideally, such means can be used to estimate performance and to enable efficient provisioning of network resources. Modelling a traffic stream emitted from a source, or a traffic stream that represents a multiplexing of many Internet traffic streams, is part of traffic modelling. It is normally reduced to finding a stochastic process that behaves like the real traffic stream from the point of view of the way it affects network performance or provides QoS to customers.

## 20.1 Parameter Fitting

One way to choose such a stochastic process is by fitting its statistical characteristics to those of the real traffic stream. Consider time to be divided into fixed-length consecutive intervals, and consider the number of packets arriving during each time interval as the real traffic stream. Then, the model of this traffic stream could be a stationary discrete-time stochastic process $\{X_n, n \geq 0\}$, with similar statistical characteristics as those of the real traffic stream. In this case, $X_n$ could be a random variable representing the number of packets that arrive in the nth interval. Let $S_n$ be a random variable representing the number of packets arriving in $n$ consecutive intervals. We may consider the following for fitting between the statistics of $\{X_n, n \geq 0\}$ and those of the real traffic stream:

- The mean $E[X_n]$.
- The variance $Var[X_n]$.
- The AVR discussed in Section 2.1. The AVR is related to the so-called Index of Dispersion for Counts (IDC) [40] as follows: the AVR is equal to $E[X_n]$ times the IDC.

A stationary stochastic process $\{X_n, n \geq 0\}$, where autocorrelation function decays slower than exponential is said to be Long Range Dependent (LRD). Notice that if the autocovariance sum $\sum_{k=1}^{\infty} Cov(X_1, X_k)$ is infinite the autocorrelation function must decay slower than exponential, so the process is LRD. In such processes the use of AVR (or IDC) may not be appropriate because it is not finite, so a time dependent version of the IDC, i.e., IDC(n) $= Var[S_n]/E[X_n]$ may be considered. Another statistical characteristic that is suitable for LRD processes is the so-called *Hurst parameter* denoted by $H$ for the range $0 \leq H < 1$ that satisfies

$$\lim_{n\to\infty} \frac{Var[S_n]}{\alpha n^{2H}} = 1. \tag{494}$$

Each of these statistical parameters have their respective continuous-time counterparts. As the concepts are equivalent, we do not present them here. We will discuss now a few examples of stochastic processes (out of many more available in the literature) that have been considered as traffic models.

## 20.2 Poisson Process

For many years the Poisson process has been used as a traffic model for the arrival process of phone calls at a telephone exchange. The Poisson process is characterized by one parameter

$\lambda$, and $\lambda t$ is the mean as well as the variance of the number of occurrences during any time interval of length $t$. Its memoryless nature makes it amenable to analysis as noticed through the analyzes of the above-mentioned queueing systems. Its ability to characterize telephone traffic well, being characterized by a single parameter, and its memoryless nature which makes it so amenable to analysis have made the Poisson process useful in the design and dimensioning of telephone networks.

By its nature, the Poisson process can accurately model events generated by a large number of independent sources each of which generating relatively sparsely spaced events. Such events could include phone calls or generation of Internet traffic flows. For example, a download of a page could be considered such a traffic flow. However, it cannot accurately model a packet traffic stream generated by a single user or a small number of users. It is important to note here that many textbooks and practitioners do consider the Poisson process as a model of a packet traffic stream (despite the inaccuracy it introduces) due to its nice analytical properties.

Normally, the Poisson process is defined as a continuous-time process. However, in many cases, it is used as a model for a discrete sequence of a traffic stream by considering time to be divided into fixed length intervals each of size one (i.e., $t = 1$), and simply to generate a sequence of independent random numbers which are governed by a Poisson distribution with mean $\lambda$ where $\lambda$ is equal to the average of the sequence we try to model. As we fit only one parameter here, namely the mean, such model will not have the same variance, and because of the independence property of the Poisson process, it will not mimic the autocorrelation function of the real process. In an assignment below, you will be asked to demonstrate that such process does not lead to a similar queueing curves as the real traffic stream.

## 20.3 Markov Modulated Poisson Process (MMPP)

Traffic models based on MMPP have been used to model bursty traffic. Due to its Markovian structure together with its versatility, the MMPP can capture bursty traffic statistics better than the Poisson process and still be amenable to queueing analysis. The simplest MMPP model is MMPP(2) with only four parameters: $\lambda_0$, $\lambda_1$, $\delta_0$, and $\delta_1$.

Queueing models involving MMPP input have been analyzed in the 70s and 80s using Z-transform [93, 95, 96, 97]. Neuts developed matrix methods to analyse such queues [65]. For applications of these matrix methods for Queueing models involving MMPP and the use of MMPP in traffic modelling and its related parameter fitting of MMPP the reader is referred to [28, 40, 57, 64].

## 20.4 Autoregressive Gaussian Process

A traffic model based on a Gaussian process can be described as a traffic process were the amount of traffic generated within any time interval has a Gaussian distribution. There are several ways to represent a Gaussian process. The Gaussian auto-regressive is one of them. Also, in many engineering applications, the Gaussian process is described as a continuous-time process. In this section, we shall define the process as a discrete time.

Let time be divided into fixed length intervals. Let $X_n$ be a continuous random variable representing the amount of work entering the system during the $n$th interval.

According to the Gaussian Autoregressive model we assume that $X_n$, $n = 1, 2, = 3 \ldots$ is the so-called $k$th order autoregressive process, defined by

$$X_n = a_1 X_{n-1} + a_2 X_{n-2} + \ \ldots \ + a_k X_{n-k} b\tilde{G}_n, \tag{495}$$

where $\tilde{G}_n$ is a sequence of IID Gaussian random variables each with mean $\eta$ and variance 1, and $a_i$ $(i = 1, 2, \ \ldots, \ k)$ and $b$ are real numbers with $|a| < 1$.

In order to characterize real traffic, we will need to find the best fit for the parameters $a_1, \ \ldots, \ a_k, \ b$, and $\eta$. On the other hand, it has been shown in [3], [4], [5] that in any Gaussian process only three parameters are sufficient to estimate queueing performance to a reasonable degree of accuracy. It is therefore sufficient to reduce the complexity involved in fitting many parameters and use only the 1st order autoregressive process, also called the AR(1) process. In this case we assume that the $X_n$ process is given by

$$X_n = aX_{n-1} + b\tilde{G}_n, \tag{496}$$

where $\tilde{G}_n$ is again a sequence of IID Gaussian random variables with mean $\eta$ and variance 1, and $a$ and $b$ are real numbers with $|a| < 1$. Let $\lambda = E[X_n]$ and $\sigma^2 = Var[X_n]$. The AR(1) process was proposed in [60] as a model of a VBR traffic stream generated by a single source of video telephony.

The $X_n$s can be negative with positive probability. This may seem to hinder the application of this model to real traffic processes. However, in modeling traffic, we are not necessarily interested in a process which is similar in every detail to the real traffic. What we are interested in is a process which has the property that when it is fed into a queue, the queueing performance is sufficiently close to that of the queue fed by the real traffic.

Fitting of the parameters $a$ $b$ and $\eta$ with measurable (estimated) parameters of the process $\lambda$, $\sigma^2$ and $S$, are provided based on [4]:

$$a = \frac{S}{S + \sigma^2} \tag{497}$$

$$b = \sigma^2(1 - a^2) \tag{498}$$

$$\eta = \frac{(1 - a)\lambda}{b} \tag{499}$$

where $S$ is the autocovariance sum given by Eq. (175).

## 20.5 Exponential Autoregressive (1) Process

In the previous section we considered an autoregressive process which is Gaussian. What made it a Gaussian process was that the so-called *innovation process*, which in the case of the previous section was the sequence $b\tilde{G}_n$, was a sequence of Gaussian random variables. Letting $D_n$ be a sequence of inter-arrival times, here we consider another AR(1) process called *Exponential Autoregressive (1)* (EAR(1)) [33], defined as follows:

$$D_n = aD_{n-1} + I_n E_n, \tag{500}$$

where $D_0 = I_0$, $\{I_n\}$ is a sequence of IID random variables in which $P(I_n = 1) = 1 - a$ and $P(I_n = 0) = a$, and $\{E_n\}$ is a sequence of IID exponential random variables with parameter $\lambda$.

The EAR(1) has many nice and useful properties. The $\{D_n\}$ process is a sequence of exponential random variables with parameter $\lambda$. These are IID only for the case $a = 0$. That is, when $a = 0$, the $\{D_n\}$ is a sequence of inter-arrival times of a Poisson process. The autocorrelation function of $\{D_n\}$ is given by

$$C_{EAR1}(k) = a^k. \tag{501}$$

It is very easy to simulate the $\{D_n\}$ process, so it is useful to demonstrate by simulation the relationship between correlation in the arrival process and queueing performance.

**Homework 20.1**

Prove that $D_n$ is exponentially distributed for all $n \geq 0$.

**Guide**

Knowing that the statement is true for $D_0$, prove that the statement is true for $D_1$. Let $\mathcal{L}_X(s)$ be the Laplace transform of random variable $X$. By definition, $\mathcal{L}_X(s) = E[e^{-sX}]$, so $\mathcal{L}_{I_1E_1}(s) = E[e^{-sI_1E_1}]$. Thus, by (95), $\mathcal{L}_{I_1E_1}(s) = P(I = 1)E[e^{-SE_1}] + P(I = 0)E[e^{-0}] = (1-a)\lambda/(\lambda+s) + a$. By definition, $\mathcal{L}_{D_1}(s) = E[e^{-s(aD_0+I_1E_1)}] = \mathcal{L}_{D_0}(as)\mathcal{L}_{I_1E_1}(s)$. Recall that $D_0$ is exponentially distributed with parameter $\lambda$, so $\mathcal{L}_{D_0}(as) = \lambda/(\lambda+as)$. Use the above to show that $\mathcal{L}_{D_1}(s) = \lambda/(\lambda+s)$. This proves that $D_1$ is exponentially distributed. Use the recursion to prove that $D_n$ is exponentially distributed for all $n > 1$. $\square$

## 20.6 Poisson Pareto Burst Process

Unlike the previous models, the Poisson Pareto Burst Process (PPBP) is Long Range Dependent (LRD). The PPBP has been proposed as a more realistic model for Internet traffic than its predecessors. According to this model, bursts of data (e.g. files) are generated in accordance with a Poisson process with parameter $\lambda$. The size of any of these bursts has a Pareto distribution, and each of them is transmitted at fixed rate $r$. At any point in time, we may have any number of sources transmitting at rate $r$ simultaneously because according to the model, new sources may start transmission while others are active. If $m$ sources are simultaneously active, the total rate equals $mr$. A further generalization of this model is the case where the burst lengths are generally distributed. In this case, the amount of work introduced by this model as a function of time is equivalent to the evolution of an M/G/$\infty$ queueing system. Having $m$ sources simultaneously active is equivalent to having $m$ servers busy in an M/G/$\infty$ system. M/G/$\infty$ which is a name of a queueing system is also often use to describe the above describe traffic model. The PPBP is sometimes called M/Pareto/$\infty$ or simply M/Pareto [2].

Again, let time be divided into fixed length intervals, and let $X_n$ be a continuous random variable representing the amount of work entering the system during the $n$th interval. For convenience, we assume that the rate $r$ is the amount transmitted by a single source within one time interval if the source was active during the entire interval. We also assume that the Poisson rate $\lambda$ is per time interval. That is, the total number of transmissions to start in one time interval is $\lambda$.

To find the mean of $X_n$ for the PPBP process, we consider the total amount of work generated in one time interval. The reader may notice that the mean of the total amount of work generated in one time interval is equal to the mean of the amount of work transmitted in one time interval. Hence,

$$E[X_n] = \lambda r/(\gamma - 1). \tag{502}$$

Also, another important relationship for this model, which is provided here without proof, is

$$\gamma = 3 - 2H, \tag{503}$$

where $H$ is the Hurst parameter.

Having the last two equations, we are able to fit the overall mean of the process ($E[X_n]$) and the Hurst parameter of the process with those measured in a real life process, and generate traffic based on the M/Pareto/$\infty$ model.

**Homework 20.2**

Use the 100,000 numbers representing the number of packets arriving recorded every second for consecutive 100,000 seconds you have collected in the assignments of Section 18 Using the UNIX command *netstat*. Again assume that these numbers represent the amount of work, measured in packets, which arrive at an SSQ during 100,000 consecutive time-intervals. Let $E[A]$ be their average. Use your SSQ simulation of the assignments of Section 18, and compute the PLR, the correlation and the variance of the amount of work arrive in large intervals (each of 1000 packet-transmission times) for the various processes you have considered and discuss the differences. □

**Homework 20.3**

Compare by simulations the effect of the correlation parameter $a$ on the performance of the queues EAR(1)/EAR(1)/1 versus their EAR(1)/M/1, M/EAR(1)/1 and M/M/1 equivalence. Demonstrate the effect of $a$ and $\rho$ on mean delay. Use the ranges $0 \leq a \leq 1$ and $0 \leq \rho \leq 1$. □

# The End of the Beginning

It is appropriate now to recall Winston Churchill's famous quote: "Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning." In this book, the reader has been introduced to certain fundamental theories, techniques and numerical methods of queueing theory and related stochastic models as well as to certain practical telecommunications applications. However, for someone who is interested to pursue a research career in this field, there is a scope for far deeper and broader study of both the theory of queues as well as the telecommunications and other applications. For the last half a century, advances in telecommunications technologies have provided queueing theorists with a wealth of interesting problems and research challenges and it is often said that the telecommunications and information technologies actually revived queueing theory. However, this is only part of the story. There are many other application areas of queueing theory. The fact is that many exceptional queueing theorists also developed expertise in various real-life systems, operations and technologies, and have made tremendous contributions to their design, operations and understanding. This dual relationship between queueing theory and its applications will likely to continue, so it is very much encouraged to develop understanding of real-life problems as well as queueing theory. And if the aim is to become expert in both, it is not the end of the beginning, but merely the beginning.

# References


[1] R. G. Addie, T. Neame and M. Zukerman, "Performance evaluation of a queue fed by a Poisson Pareto Burst process", *Computer Networks*, vol. 40, no. 3, October 2002, pp. 377–397.

[2] R. G. Addie, M. Zukerman and T. D. Neame, "Broadband traffic modeling: simple solutions to hard problems", *IEEE Communication Magazine*, August 1998, pp. 88–95.

[3] R. G. Addie and M. Zukerman, "An approximation for performance evaluation of stationary single server queues," *IEEE Transactions on Communications*, vol. 42, no. 12, December 1994, pp. 3150–3160.

[4] R. G. Addie and M. Zukerman, "Performance evaluation of a single server autoregressive queue," *Australian Telecommunication Research*, vol. 28, no. 1, 1994, pp. 25–32.

[5] R. G. Addie and M. Zukerman, "Queues with total recall - application to the B-ISDN," in J. Labetoulle and J. W. Roberts Editors, *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, Vol. 1a in the series: Teletraffic Science and Engineering*, North Holland - Elsevier Science Publishers, Amsterdam 1994, pp. 45–54.

[6] H. Akimaru and K. Kawashima, *Teletraffic - theory and application*, ISBN 3-540-19805-9, 2nd ed., Springer, London 1999.

[7] O. D. Anderson, *Time Series analysis and forecasting - The Box Jenkins approach*, Butterworths, London, 1976.

[8] F. Baccelli, S. Machiraju, D. Veitch and J. Bolot, "The Role of PASTA in Network Measurement", pp. 231-242, vol. 36, no. 4, *Computer Communication Review, Proceedings of ACM Sigcomm 2006*, Pisa, Italy, 11-15 Sep., 2006.

[9] F. Baccelli, S. Machiraju, D. Veitch and J. Bolot, "On Optimal Probing for Delay and Loss Measurement", *Proc. 2007 ACM Sigcomm Internet Measurement Conference*, pp. 291-302, San Diego, California, 23-26 October, 2007.

[10] V. E. Benes, "Fluctuations of telephone traffic", *Ibid.*, vol. 36, pp. 965-973, 1957.

[11] V. E. Benes, *General Stochastic Processes in the Theory of Queues*, Addison Wesley, 1963.

[12] D. Bertsekas, Session 1813 - Traffic Behavior and Queuing in a QoS Environment, OPNET, a PPT presentation, 2002. [Online]. Available: `https://www.google.com.hk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CBsQFjAA&url=http%3A%2F%2Fweb.mit.edu%2Fdimitrib%2Fwww%2FOPNET_Full_Presentation.ppt&ei=iuyoVMqhOoGzuAS2y4DACA&usg=AFQjCNFrb2-KlpwhyHuVjtWexFWdYkNiWg&sig2=h77dvNwgq820Cfpt1G5_fQ&bvm=bv.82001339,d.dGY`

[13] D. Bertsekas and R. Gallager, *Data networks*, Prentice Hall, Englewood Cliff, New Jersey 1992.

[14] D. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, Belmont, Massachusetts 2002.

[15] S. K. Bose, *An Introduction to queueing systems*, Kluver Academic/Plenum Publisher, New York 2002.

[16] G. E. P. Box and G. M. Jenkins, *Time series analysis forecasting and control*, Holden Day

San Francisco, 1970.

[17] P. J. Burke, "The output of a queuing system," *Operations Research*, vol. 4, 1956), pp. 699–704.

[18] S. Butterfield, "A new Rx for crowded hospitals: Math", *ACP Hospitalist*, December 2007. [Online]. Available: `http://www.acphospitalist.org/archives/2007/12/math.htm`

[19] Cisco Inc., "Traffic Analysis". [Online]. Available: `http://www.cisco.com/c/en/us/td/docs/ios/solutions_docs/voip_solutions/TA_ISD.html`

[20] J. W. Cohen, "The Generalized Engset Formulae", Philips Telecommunication Review, vol. 18, 1957, pp. 158–170.

[21] J. W. Cohen, *The Single Server Queue*, North-Holland Publ. Cy., rev. ed., Amsterdam, 1982.

[22] R. B. Cooper, *Introduction to Queueing Theory*, North Holland 1981.

[23] T. Engset, "Die wahrscheinlichkeitsrechnung zur bestimmung der wahleranzahl in automatischen fernsprechamtern." *Elektrotechnische zeitschrift*, vol. 39, no. 31, Aug. 1918, pp. 304–306.

[24] A. K. Erlang, "Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges", *The Post Office Engineer's Journal*, vol. 10, 1917, pp. 189-197.

[25] W. Feller, *An introduction to probability theory and its applications*, Wiley, New York, 1968.

[26] D. Fesehaye, "Emulating TCP (A Reliable Internet Protocol) Using a Fixed Point Algorithm", *Proceedings of the 31st IEEE Conference on Local Computer Networks*, pp. 159–165, Tampa, FL, Nov. 2006.

[27] M. Fiedler, T. Hossfeld and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network,* vol. 24, no. 2, pp. 36–41, March-April 2010.

[28] W. Fischer and K. Meier-Hellstern, "The Markov Modulated Poisson process (MMPP) cookbook," *Performance Evaluation*, vol. 18, pp. 149–171, 1992.

[29] G. S. Fishman, "Discrete-event simulation," Springer-Verlag, London 2001.

[30] R. Fortet and Ch. Grandjean, "Congestion in a loss system when some calls want several devices simultaneously," *Proceedings of the Fourth International Teletraffic Congress*, London, England, 15–21 July, 1964.

[31] A. Fredericks, "Congestion in blocking systems – a simple approximation technique", *Bell System Technical Journal*, vol. 59, no. 6, pp. 805-827, July–August 1980.

[32] M. Garetto and M. A. Marsan and M. Meo and R. L. Cigno, "On the use of fixed point approximations to study reliable protocols over congested links", *Proceeding of IEEE GLOBECOM '03*, vol. 6, pp. 3133–3137, San Francisco, December 2003.

[33] D. P. Gaver and P. A. W. Lewis, "First-Order Autoregressive Gamma Sequences and Point Processes," *Advances in Applied Probability*, vol. 12, no. 3, pp. 727–745, Sep. 1980.

[34] A. Girard, *Routing and dimensioning in circuit-switched networks,* Addison Wesley, 1990.

[35] L. V. Green and V. Nguyen, "Strategies for cutting hospital beds: the impact in patient service", *Health Services Research*, vol. 36, no. 2, pp. 421-442, 2001.

[36] L. V. Green, "Queueing Analysis in Healthcare", in R. W. Hall editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, International Series in Operations Research; Management Science, Springer US, pp. 281-307, vol. 91, 2006.

[37] D. Gross, J. F. Shortle, J. M. Thompson and C. M. Harris, *Fundamentals of queueing theory*, John Wiley & Sons Inc., 2008.

[38] D. P. Heyman, "Comments on a Queueing Inequality," *Management Science*, vol. 26, no. 9, pp. 956-959, Sep, 1980.

[39] S. Halfin and W. Whitt, "Heavy-traffic limits for queues with many exponential servers," *Operations Research*, vol. 29, no. 3, pp. 567–588, 1981.

[40] H. Heffes and D. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE Journal on Selected Areas in Communications*, vol. 4, no. 6, pp. 856–868, Sep 1986.

[41] P. G. Hoel, S. C. Port and C. J. Stone, "Introduction to Stochastic Processes," Houghton Mifflin, 1972.

[42] F. Hübner, "Analysis of a finite-capacity asynchronous multiplexer with deterministic traffic sources", *proceedings of 7th ITC Specialist Seminar*, Morristown, October 1990.

[43] P. J. Hunt and F. P. Kelly, "On critically loaded loss networks," *Advances in Applied Probability*, vol. 21, pp. 831-841, 1989.

[44] J. Hui, *Switching and traffic theory for integrated broadband networks*, Kluwer 1990.

[45] ITU-T, "ITU-T E.500: Traffic intensity measurement principles", November 1998. [Online] Available: `https://www.itu.int/rec/T-REC-E.500-198811-S/en`

[46] V. B. Iversen, "The exact evaluation of multi-service loss system with access control," Teleteknik, English ed., vol. 31, *Proc. Seventh Nordic Teletraffic Seminar*, Lund, Sweden, August 25–27, 1987.

[47] V. B. Iversen, *Teletraffic Engineering Handbook*, Technical University of Denmark. [Online]. Available: `https://www.itu.int/dms_pub/itu-d/opb/stg/D-STG-SG02.16.1-2001-PDF-E.pdf`

[48] D. L. Jagerman, "Some properties of the Erlang loss function," *Bell Syst. Tech. J.*, vol. 53, pp. 525–511, 1974.

[49] J. S. Kauffman, "Blocking in a shared resource environment," *IEEE Transactions on Communications*, vol. COM-29, no. 10, pp. 1474-1481, 1981.

[50] F. P. Kelly, *Reversibility and stochastic networks*, John Wiley, 1979.

[51] D. G. Kendall, "Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain", *The Annals of Mathematical Statistics*, vol. 24, no. 3, Sep., 1953, pp. 338–354.

[52] D. G. Kendall, "Some problems in the theory of queues", *J. R. Statist. Soc. B*, vol. 13, pp. 151–173, 1951.

[53] L. Kleinrock, "Time-shared Systems: a theoretical treatment," *Journal of the ACM*, vol.

14 , no. 2, April 1967, pp. 242 - 261.

[54] L. Kleinrock, *Queueing Systems, Volume 1 - Theory*, John Wiley and Sons 1975.

[55] L. Kleinrock, *Queueing Systems, Volume 2 - Computer Applications*, John Wiley and Sons 1976.

[56] A. Leon Garcia, *Probability and random processes for electrical engineering*, Addisson Wesley, Reading, Mass., 1989.

[57] G. Latouche and V. Ramaswami, *Introduction to matrix analytic methods in stochastic modeling*, SIAM, 1999.

[58] D. V. Lindley, "The theory of queues with a single server", *Proc. Camb. Phil. Soc.*, vol. 48, pp. 277–289, 1952.

[59] J. D. C. Little, "A proof of the queuing formula: L = $\lambda$W", *Operations Research*, vol. 9, pp. 383–387, 1961.

[60] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications", *IEEE Transactions on Communications*, vol. 36, no. 7, pp. 834–844, July 1988.

[61] M. Mathis, J. Semke, J. Madhavi, and T. Ott, "The macroscopic behavior of the TCP congestion avoidance algorithm", *ACM Computer Cmmunication Review*, vol. 27, no. 3, pp. 67–82, July 1997.

[62] USC/ISI, Los Angeles, "The ns simulator and the documentation". [Online]. Available: `http://www.isi.edu/nsnam/ns/`.

[63] C. R. Nelson, *Applied time series analysis*, Holden Day, San Francisco, 1973.

[64] M. F. Neuts, *Structured stochastic matrices of M/G/1 type and their applications*, Marcel Dekker, NY 1989.

[65] M. F. Neuts, *Matrix-geometric solutions in stochastic models - algorithmic approach*, The John Hopkin's University Press, Baltimore (1981).

[66] G. F. Newell, "The M/G/$\infty$ queue", *SIAM Journal on Applied Mathematics*, vol. 14, no. 1, pp. 86-88, January 1966.

[67] A. Papoulis, *Probability, random variables and stochastic processes*, 2nd Edition, McGraw Hill Book Company, New York, 1984.

[68] E. Parzen, *Stochastic processes*, Holden Day, San Francisco, 1962.

[69] E. Reich, "On the integrodifferential equation of Takacs. 1." *Ann. Math. Stat.*, vol. 29, pp. 563–570, 1958.

[70] J. Riordan, "Telephone traffic time averages", *Bell System Technical Journal*, vol. 30, pp. 1129-1144, 1951.

[71] J. W. Roberts, "A service system with heterogeneous user requirements", in G. Pujolle, editor, *Performance of Data Communications Systems and Their Applications*, pp. 423-431, North Holland, 1981.

[72] J. W. Roberts, U. Mocci and J. Virtamo, *Broadband Network Teletraffic*, Final Report of Action COST 242, Lecture Notes in Computer Science, vol. 1155, Springer, 1996.

[73] N. Rönnblom, "Traffic loss of a circuit group consisting of both–way circuits which is accessible for the internal and external traffic of a subscriber group", *Teleteknik* (English edition), No. 2, pp. 79–92, 1959.

[74] Z. Rosberg, H. L. Vu, M. Zukerman and J. White, "Performance analyses of optical burst switching networks", *IEEE Journal of Selected Areas in Communications*, vol. 21, no. 7, Sept. 2003, pp. 1187–1197.

[75] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*, ISBN 3-540-19918-7, Springer-Verlag, Berlin, Heidelber, New York, 1995.

[76] K. W. Ross and D. Tsang, "Optimal circuit access policies in an ISDN environment: a Markov decision approach", *IEEE Transactions on communications*, vol. 37, pp. 934–939, 1989.

[77] K. W. Ross and D. Tsang, "The stochastic knapsack problem", *IEEE Transactions on communications*, vol. 37, pp. 740–747, 1989.

[78] K. W. Ross and D. Tsang, "Teletraffic engineering for product-form circuit-switched networks", *Advances in Applied Probability*, vol. 22, pp. 657–675, 1990.

[79] S. M. Ross, *Introduction to probability models*, Academic Press, New York, 1993.

[80] S. M. Ross, *A first course in probability*, Macmillan, New York, 1976.

[81] S. M. Ross, *Stochastic processes*, Wiley, New York, 1996.

[82] B. A. Sevastyanov, "An ergodic theorem for Markov processes and its application to telephone systems with refusals", *Theor. Probab. Appl.*, vol. 2, no. 1, pp. 104–112, 1957.

[83] S. Stidham, "L = $\lambda$W: a discounted analogue and a new proof", *Operations Research*, vol. 20, pp. 1115–1126, 1972.

[84] S. Stidham, "A last word on L = $\lambda$W", *Operations Research*, vol. 22, pp. 417–421, 1974.

[85] R. Syski, *Introduction to Congestion Theory in Telephone Systems*, 2nd ed. Elsevier Science Publishers B. V., 1986.

[86] J. Sztrik, *Basic Queueing Theory.* [Online]. Available: `http://irh.inf.unideb.hu/~jsztrik/education/16/SOR_Main_Angol.pdf`.

[87] P. G. Taylor, "Insensitivity in Stochastic Models," Chapter 3, in N. van Dijk and R. Boucherie (eds.) *Queueing Networks: A Fundamental Approach*, Springer-Verlag, pp. 121-140, Berlin, New York, 2010.

[88] D. O. Weber. Queue fever. Hospitals and Health Networks (Digitial Magazine), May 2006. [Online]. Available: `http://www.ihoptimize.org/Collateral/Documents/English-US/queuefever1.pdf`.

[89] W. Whitt, "IEOR 6707: Advanced topics in queueing theory: focus on customer contact centers", Fall 2002, [Online] Available: `http://www.columbia.edu/~ww2040/ErlangBandCFormulas.pdf`.

[90] W. Whitt, "IEOR 6711: Stochastic Models I: Solutions to Homework Assignment 1", Fall 2012, [Online] Available: `http://www.columbia.edu/~ww2040/6711F12/homewk1Sols.pdf`.

[91] R. Wilkinson, "Theories of toll traffic engineering in the USA", *Bell System Technical*

*Journal*, vol. 40, pp. 421–514, 1956.

[92] A. K. Wong, "Queueing analysis for ATM switching of continuous-bit-rate traffic – a recursion computation method," *Proceedings of GLOBECOM '90*, vol. 3, San Diego, CA, Dec. 2-5, 1990, pp. 1438–1444.

[93] U. Yechiali and P. Naor, "Queuing Problems with Heterogeneous Arrivals and Service," *Operations Research*, vol. 19, pp. 722–734, 1971.

[94] K. Zhu, H. Zhu and B. Mukherjee, *Traffic grooming in optical WDM mesh networks*, Springer, 2005.

[95] M. Zukerman and I. Rubin, "Performance of flow-controlled communications systems under bursty traffic," *Proceedings of IEEE GLOBECOM '86*, vol. 3, Houston, December 1986, pp. 1266–1271.

[96] M. Zukerman and I. Rubin, "Queueing performance of demand-assigned multi access communication systems under bursty traffic conditions," *Proceedings of IEEE ICC '86*, vol. 3, no. 57.2, Toronto, Canada, June 1986, pp. 1827–1832.

[97] M. Zukerman and I. Rubin, "On multi channel queueing systems with fluctuating parameters," *Proceedings of IEEE INFOCOM '86*, Miami, Florida, April 1986, pp. 600–608.