Introduction to Queueing Theory and Stochastic Teletraffic Models

by Moshe Zukerman

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, mathematics, probability and stochastic processes normally taught in an electrical engineering course. For students who have some but not sufficiently strong background in probability and stochastic processes, we provide, in the first few chapters, a revision of the relevant concepts in these areas.

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" [13]; 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.

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.

As mentioned above, the first two chapters provide a revision of probability and stochastic processes topics relevant to the queueing and teletraffic models of this book. The content of these chapters is mainly based on [13, 23, 62, 67, 68, 69]. These chapters are intended for students who have some background in these topics. Students with no background in probability and stochastic processes are encouraged to study the original textbooks that include far more explanations, illustrations, discussions, examples and homework assignments. For students with background, we provide here 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. 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 – 13 provide analyses of a wide range of queueing and teletraffic models that fall under the category of continuous-time Markov-chain processes. Chapter 14 provides an example of a discrete-time queue that is modelled as a discrete-time Markov-chain. In Chapter 15, various aspects of a single server queue with Poisson arrivals and general service times are studied, mainly focussing on mean value results as in [12]. Then, in Chapter 16, some selected results of a single server queue with a general arrival process and general service times are provided. Next, in Chapter 17, we extend our discussion to queueing networks. Finally, in Chapter 18, 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 1.18 describes how properties of Gaussian distribution can be applied to link dimensioning. Section 6.4 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/∞ queueing model to a multiple access performance problem [12] is discussed in Section 7.5. In Sections 8.6 and 9.4, discussions on dimensioning and related utilization issues of a multi-channel system are presented. Section 17.3 guides the reader to simulate a mobile cellular network. Section 18.6 describes a traffic model applicable to the Internet.

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

Contents

1 Revision of 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. We aim to provide in this chapter sufficient coverage for readers that have some probability background. 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 without prior probability background may be aided by additional probability texts such as [13] and [68].

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 than $Y = g(X)$ 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, than $Y = \sum_{i=1}^{n} X_i$ is also a random variable.

1.2 Probability, Conditional Probability and Independence

Consider a sample space S. Let A be a set in S, the probability of A is the function on S, 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 are represented by the theoretical concept 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,

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

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

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

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

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

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,

$$A = \bigcup_{i=1}^{n} (A \cap B_i) \tag{4}$$

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

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

Thus, by Eq. (1),

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

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. (6) and (1) (again), we obtain the following formula for conditional probability between two events:

$$P(B_1 \mid A) = \frac{P(A \mid B_1)P(B_1)}{\sum_{i=1}^{n} P(A \mid B_i) \times P(B_i)}. \tag{7}$$

The latter is known as Bayes' formula.

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

$$F_X(x) = P(X \le x). \tag{8}$$

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

$$\bar{F}_X(x) = P(X > x). \tag{9}$$

Consequently, for any random variable, for every $x \in R$, $F(x) + \bar{F}(x) = 1$.

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:

$$F_{X_1,\ X_2,\ ...,X_n}(x_1,\ x_2,\ ...,x_n) = P(X_1 \le x_1,\ X_2 \le x_2,\ \ldots,\ X_n \le x_n). \tag{10}$$

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

$$F_{X_1}(x_1) = F_{X_1,\ X_2,\ ...,\ X_n}(x_1,\ \infty,\ \ldots,\ \infty). \tag{11}$$

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

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

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

$$P_{X_1}(x_1) = \sum_{x_2} \cdots \sum_{x_n} P_{X_1,\ X_2,\ ...,\ X_n}(x_1,\ x_2,\ \ldots,\ x_n). \tag{13}$$

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. In this section and in sections 1.5, 1.6, 1.7, when we mention random variables or their probability and distribution function, 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.

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

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}. \tag{14}$$

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 (14)

$$P_{X|Y}(x \mid y) = \frac{P_{X,Y}(x, y)}{P_Y(y)}. \tag{15}$$

Noticing that

$$P_Y(y) = \sum_x P_{X,Y}(x, y), \tag{16}$$

we obtain by (15)

$$\sum_x P_{X|Y}(x \mid y) = 1. \tag{17}$$

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 probability and distribution functions there. By (15)

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

and by symmetry

$$P_{X,Y}(x, y) = P_{Y|X}(y \mid x)P_X(x) \tag{19}$$

so the latter and (16) gives

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

which is another version of the law of total probability (6).

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

$$P_{U,V}(u, v) = P_U(u)P_V(v) \quad \text{for all } u, v. \tag{21}$$

Notice that by (19) and (21), we obtain an equivalent definition of independent random variables U and V which is

$$P_{U|V}(u \mid v) = P_U(u) \tag{22}$$

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

$$P_Y(y) = \sum_{x_2,\ x_3,\ \ldots,\ x_k:\ x_2+x_3+\ \ldots\ +x_k \leq y} \left(P_{X_1}(y - \Sigma_{i=2}^{k} x_i) \prod_{i=2}^{k} P_{X_i}(x_i) \right). \tag{23}$$

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 $P_{X_1}(x)$.

1.8 Selected Discrete Random Variables

We present here several discrete random variables and their corresponding probability distributions. Although our cover here is not exhaustive, we do consider all the discrete random variables mentioned later in this book.

1.8.1 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:

$$\begin{aligned} P(X = 1) &= p \\ P(X = 0) &= 1 - p. \end{aligned} \tag{24}$$

1.8.2 Geometric

The geometric random variable X 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

$$P(X=i)=(1-p)^{i-1}p \quad \text{for} \quad i=1,2,3, \ \ldots \ . \tag{25}$$

1.8.3 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 1.1

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, namely, $0 \geq N_1 - N_2 \geq -1$. Assume that each 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 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 $\lceil x \rceil$ is the smallest integer greater or equal to x. □

1.8.4 Poisson

A Poisson random variable with parameter λ has the following probability function:

$$P(X = i) = e^{-\lambda}\frac{\lambda^i}{i!} \qquad i = 0,\ 1,\ 2,\ 3,\ \ldots\ . \tag{26}$$

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. It will be shown rigorously using Z-transform in Subsection 1.14.1 that if n increases and np stays constant, a binomial random variable with parameters n and p approaches Poisson with parameter $\lambda = np$. 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 packets) n is large. The probability p of a customer (or packet) making a call within a given short period of time is small, and the calls are typically independent. 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.

Homework 1.2

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 $P_X(x)$. □

1.8.5 Pascal

The Pascal random variable X with parameters k (integer and ≥ 1) and p (real within (0,1]), represents a sum of k geometric random variables each with parameter p. For X to be equal to i, we must have a successful Bernoulli trial at the ith trial because this is the successful trial associated with the kth 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 ith trial is equal to the probability of a Bernoulli random variable with parameter p equal to 1, which is 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 can multiply their probabilities to obtain

$$P(X = 1) = \binom{i-1}{k-1} p^{k}(1-p)^{i-k} \qquad i = k,\ k+1,\ k+2,\ \ldots\ . \tag{27}$$

1.9 Continuous Random Variables and their Probability Functions

Continuous random variable 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 $A \subset R$,

$$P(X \in A) = \int_A f(x)dx. \tag{28}$$

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,

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \tag{29}$$

If we consider Eq. (28), letting $A = [a, b]$, we obtain,

$$P(a \leq x \leq b) = \int_a^b f(x)dx. \tag{30}$$

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. (30), we obtain

$$P(x = a) = \int_a^a f(x)dx = 0. \tag{31}$$

As a result, the distribution function $F(x)$ 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. (30), we obtain

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(s)ds. \tag{32}$$

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. (12), 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

$$P(\{X, Y\} \in A) = \iint_{\{X,Y\}\in A} f_{X,Y}(x, y)dxdy. \tag{33}$$

The continuous equivalence of the first equality in (20) is:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx. \tag{34}$$

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

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}. \tag{35}$$

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

$$\int_{-\infty}^{\infty} f_{X|Y}(x \mid y)dx = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)dx}{f_Y(y)} = \frac{f_Y(y)}{f_Y(y)} = 1. \tag{36}$$

Notice the equivalence between the conditional probability (1) and the conditional density (35). By (35)

$$f_{X,Y}(x,y) = f_Y(y)f_{X|Y}(x \mid y) \tag{37}$$

so

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\infty}^{\infty} f_Y(y)f_{X|Y}(x \mid y)dy. \tag{38}$$

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 $X \in A$. Thus,

$$P(A) = P(X \in A) = \int_A f_X(x)dx = \int_A \int_{-\infty}^{\infty} f_Y(y)f_{X|Y}(x \mid y)dydx. \tag{39}$$

Hence,

$$P(A) = \int_{-\infty}^{\infty} f_Y(y) \int_A f_{X|Y}(x \mid y)dxdy \tag{40}$$

and therefore

$$P(A) = \int_{-\infty}^{\infty} f_Y(y)P(A \mid Y = y)dy \tag{41}$$

which is the continuous equivalence of the Law of Total Probability (6).

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.

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

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

$$f_Y(y) = \iint_{x_2,\ x_3,\ \ldots,\ x_k:\ x_2,\ x_3,\ \ldots,\ x_k \leq y} \left(f_{X_1}(y - \Sigma_{i=2}^{k} x_i) \prod_{i=2}^{k} f_{X_i}(x_i) \right). \tag{42}$$

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}.

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

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise.} \end{cases} \tag{43}$$

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

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{44}$$

The uniform (0,1) random variable is very important in simulations. Almost all computers programs have a function which generates 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,

$$U_1(0, 1) = F(x_1) \tag{45}$$

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

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, $P(Y \leq x) = F(x)$. $\square$

Homework 1.3

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

$$P(X > t) = \begin{cases} 0 & \text{for } t \leq 0 \\ (1 - \frac{t}{s})^k & \text{for } 0 < t < 1 \\ 1 & \text{otherwise.} \end{cases} \tag{46}$$

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

1.10.2 Exponential

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

$$f(x) = \begin{cases} \mu e^{-\mu x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{47}$$

Its distribution function is given by

$$F(x) = \int_0^x \mu e^{-\mu s} ds = 1 - e^{-\mu x} \qquad x \geq 0. \tag{48}$$

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

$$\bar{F}(x) = e^{-\mu x} \qquad x \geq 0. \tag{49}$$

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. This property is called the *memoryless* property of a random variable. In particular, a random variable is called memoryless if

$$P(X > s + t \mid X > t) = P(X > s). \tag{50}$$

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. To show that exponential random variable is memoryless we show that Eq. (50) 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 1.4

Write a computer program that generates a sequence of 100 random deviates from an exponential distribution with $\mu = 1$. □

Let X_1 and X_2 be exponentially distributed random variables with parameters λ_1 and λ_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

$$P(X > t) = P(\min[X_1, X_2] > t) = P(X_1 > t, X_2 > t) = e^{-\lambda_1 t} e^{-\lambda_2 t} = e^{-(\lambda_1+\lambda_2)t}. \tag{51}$$

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 (41) as follows:

$$P(X_1 < X_2) = \int_0^\infty (1 - e^{-\lambda_1 t})\lambda_2 e^{-\lambda_2 t} dt = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \tag{52}$$

In the following table we explain how to obtain the latter from the continuous version of the law of total probability (41) by pointing out the equivalence between the corresponding terms in the two equations.

term in (41)	**equivalent term in (52)**
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,

$$P(X_1 > X_2) = \frac{\lambda_2}{\lambda_1 + \lambda_2}. \tag{53}$$

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 Hyper-Exponential

Let X_i for $i = 1, 2, 3, \ldots, k$ be k independent exponential random variables with parameters λ_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

$$f_X(x) = \sum_{i=1}^k p_i f_{X_i}(x). \tag{54}$$

1.10.4 Erlang

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

$$f_X(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}. \tag{55}$$

Homework 1.5

Let X_i, $i = 1,\ 2,\ \ldots,\ k$ be k independent exponentially distributed random variables each with parameter λ, prove by induction that the random variable X defined by the sum $X = \sum_{i=1}^{k} X_i$ has Erlang distribution with parameter k and λ. In other words, $f_X(x)$ of (55) is a k-fold convolution of $\lambda e^{-\lambda x}$. □

1.10.5 Hypo-Exponential

Let X_i, $i = 1,\ 2,\ \ldots,\ k$ be k independent exponentially distributed random variables each with parameters λ_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 al the k random variables are identically distributed.

1.10.6 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 σ^2 if its density is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/2\sigma^2} \qquad -\infty < x < \infty. \tag{56}$$

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 (not necessarily of the same distribution, but each has a finite variance) 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 1.17 we will further discuss the central limit theorem and demonstrate its applicability to the telecommunication link dimensioning problem in Section 1.18.

1.10.7 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 γ and δ by its complementary distribution function which is given by

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

Homework 1.6

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

1.11 Moments

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

$$E[X] = \sum_{\{n:P(n)>0\}} nP_X(n). \tag{57}$$

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

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)dx. \tag{58}$$

A very 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:

$$E[Z] = \int_0^{\infty} P(Z > z)dz = \int_0^{\infty} [1 - F_Z(z)]dz. \tag{59}$$

The discrete equivalence of the latter is:

$$E[Z] = \sum_{n=0}^{\infty} P(Z > n) = \sum_{n=0}^{\infty} [1 - F_Z(n)]. \tag{60}$$

Homework 1.7

Use geometrical arguments to show (59) and (60). □

Homework 1.8

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 (46) and (59). □

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

$$E[g(X)] = \sum_{\{k:P_X(k)>0\}} g(k)P_X(k) \tag{61}$$

for a discrete random variable and

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx \tag{62}$$

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

$$E[aX] = aE[X], \tag{63}$$

$$E[X - b] = E[X] - b, \tag{64}$$

and

$$E[aX - b] = aE[X] - b. \tag{65}$$

The n**th moment** of the random variable X is defined by $E[X^n]$. Substituting $g(X) = X^n$ in (61) and in (62), the nth moment of X is given by:

$$E[X^n] = \sum_{\{k:P_X(k)>0\}} k^n P_X(k) \tag{66}$$

for a discrete random variable and

$$E[X^n] = \int_{-\infty}^{\infty} x^n f_X(x)dx \tag{67}$$

for a continuous random variable. Similarly, the nth central moment of random variable X is defined by $E[(X - E[X])^n]$. Substituting $g(X) = (X - E[X])^n$ in (61) and in (62), the n**th central moment** of X is given by:

$$E[(X - E[X])^n] = \sum_{\{k:P(k)>0\}} (k - E[X])^n P_X(k) \tag{68}$$

for a discrete random variable and

$$E[(X - E[X])^n] = \int_{-\infty}^{\infty} (x - E[X])^n f_X(x)dx \tag{69}$$

for a continuous random variable. By definition the first moment is the mean. The second central moment is called the **variance**. It is defined as

$$var[X] = E[(X - E[X])^2]. \tag{70}$$

The variance of a random variable X is given by

$$var[X] = \sum_{\{k:P(k)>0\}} (k - E[X])^2 P_X(k) \tag{71}$$

if X is discrete, and by

$$var[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x)dx \tag{72}$$

if it is continuous.

By (70) we obtain

$$var[X] = E[(X - E[X])^2] = E[X^2 - 2XE[X] + (E[X])^2] = E[X^2] - (E[X])^2. \tag{73}$$

In the following table we provide the mean and the variance of some of the above described random variables.

random variable	parameters	mean	variance
Bernoulli	$0 \leq p \leq 1$	p	$p(1-p)$
binomial	n and $0 \leq p \leq 1$	np	$np(1-p)$
Poisson	$\lambda > 0$	λ	λ
uniform	a and b	$(a+b)/2$	$(b-a)^2/12$
exponential	$\mu > 0$	$1/\mu$	$1/\mu^2$
Gaussian	m and σ	m	σ^2
Pareto	$\delta > 0$ and $1 < \gamma \leq 2$	$\frac{\delta\gamma}{(\gamma-1)}$	∞

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

While the mean provides the average, or the average of possible values a random variable can take weighted according to its probability function or density, 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 σ_X, or simply σ, and defined by

$$\sigma_X = \sqrt{var[X]}. \tag{74}$$

Hence the variance is often denoted by σ^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 mean of sum of random variables is always the sum of their means, namely,

$$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] \tag{75}$$

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

$$var[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} var[X_i]. \tag{76}$$

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

$$E[\Pi_{i=1}^{n} X_i] = \Pi_{i=1}^{n} E[X_i]. \tag{77}$$

In many application it is useful to use the concept of **Conditional Expectation (or Mean)** to derive moments of unknown distributions. It is defined by:

$$E[X \mid Y] = E_X[P(X \mid Y)], \tag{78}$$

where the subscript X indicates that the mean is over X. For example, the conditional expectation of two discrete random variables is defined by

$$E[X \mid Y = j] = \sum_i iP(X = i \mid Y = j). \tag{79}$$

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

$$E[X \mid Y = y] = \int_{x=-\infty}^{\infty} x f_{X|Y}(x \mid y)dx. \tag{80}$$

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 Y is discrete) we obtain

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

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

$$E[X] = E_Y[E[X \mid Y]]. \tag{81}$$

The latter also applies to continuous random variables. In this case we have:

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

Homework 1.9

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:

$$P_X(x) = E_Y[P(X = x \mid Y = y)]. \tag{82}$$

Define the **Conditional Variance** as

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

This gives rise to the following useful formula for the variance of a random variable known as EVVE:

$$var[X] = E[var[X \mid Y]] + var[E[X \mid Y]]. \tag{84}$$

To show EVVE, we recall (73): $var[X] = E[X^2] - (E[X])^2$, and (81): $E[X] = E_Y[E[X \mid Y]]$, we obtain

$$var[X] = E[E[X^2 \mid Y]] - (E[E[X \mid Y]])^2. \tag{85}$$

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

$$var[X] = E[var[X \mid Y] + (E[X \mid Y])^2] - (E[E[X \mid Y]])^2 \tag{86}$$

or

$$var[X] = E[var[X \mid Y]] + E[E[X \mid Y]]^2 - (E[E[X \mid Y]])^2. \tag{87}$$

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

$$var[E[X \mid Y]] = E[E[X \mid Y]]^2 - (E[E[X \mid Y]])^2, \tag{88}$$

observing that the right-hand side of (88) equals to the last two terms in the right-hand side of (87), 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 ϕ, has mean ϕ_e and variance ϕ_v. The number of packets in each flow, denoted ς, has mean ς_e and variance ς_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 ϕ and ς are independent $E[W|\phi] = \phi\varsigma_e$ and by (81)

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

Note that the relationship

$$W_e = \phi_e\varsigma_e \tag{89}$$

is also obtained directly by (77). In fact, the above proves (77) 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

$$W_v = \phi_v\varsigma_v + \varsigma_v\phi_e^2 + \phi_v\varsigma_e^2. \tag{90}$$

Homework 1.10

1. Provide detailed derivations of Equations (89) and (90) using (81) and (84).
2. Derive Equations (89) and (90) in a different way, considering the independence of the number of packets in each flow and the number of flows arriving per second. □

1.12 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

$$S_m = \frac{\sum_{i=1}^{n} X(i)}{n} \tag{91}$$

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

$$S_v = \frac{\sum_{i=1}^{n} [X(i) - S_m]^2}{n-1} \tag{92}$$

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

Homework 1.11

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. □

1.13 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 ith 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_is. 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

$$cov(X_1, X_2) = E[(X_1 - E[X_1])(X_2 - E[X_2])]. \tag{93}$$

Intuitively, by Eq. (93), 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. (93),

$$cov(X_1, X_2) = E[X_1 X_2] - E[X_1]E[X_2]. \tag{94}$$

Hence, by (77), 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

$$var[X_1 + X_2] = var[X_1] + var[X_2] + 2cov(X_1, X_2). \tag{95}$$

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 1.12

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. $\square$

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

$$var(S_k) = \sum_{i=1}^{k} var[X_i] + 2\sum_{i<j} cov[X_i, X_j] \tag{96}$$

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 1.13

Prove Eq. (96).

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 (96), we obtain

$$var(S_k) = \sum_{i=1}^{k} var[X_i]. \tag{97}$$

Homework 1.14

Use Eq. (96) 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

$$cov(X_1, X_2) \leq \sqrt{var[X_1]var[X_2]}, \tag{98}$$

the correlation of two random variables X and Y denoted by $corr(X, Y)$ is defined by

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}, \tag{99}$$

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

Homework 1.15

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. □

1.14 Transforms

Transforms are very useful in analysis of probability models and queueing systems. We will first consider the following general definition [13] for a transform function Γ of a random variable X:

$$\Gamma_X(\omega) = E[e^{\omega X}] \tag{100}$$

where ω 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,

$$\Gamma_X(\omega) = \int_{-\infty}^{\infty} e^{\omega x} f_X(x) dx. \tag{101}$$

Taking derivative with respect to ω leads to

$$\Gamma_X'(\omega) = \int_{-\infty}^{\infty} x e^{\omega x} f_X(x) dx. \tag{102}$$

Letting $\omega \to 0$, we obtain

$$\lim_{\omega \to 0} \Gamma_X'(\omega) = E[X], \tag{103}$$

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

$$\lim_{\omega \to 0} \Gamma_X^{(n)}(\omega) = E[X^n]. \tag{104}$$

Homework 1.16

Derive Eq. (104) using (101) – (103) completing all the missing steps. □

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

$$\Gamma_X(\omega) = E[e^{\omega X}] = \lambda \int_{x=0}^{\infty} e^{\omega x} e^{-\lambda x} dx, \tag{105}$$

which gives after some derivations

$$\Gamma_X(\omega) = \frac{\lambda}{\lambda - \omega}. \tag{106}$$

Homework 1.17

Derive Eq. (106) from (105) □

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

$$\Gamma_Y(\omega) = E[e^{\omega Y}] = E[e^{\omega(aX+b)}] = e^{\omega b}E[e^{\omega aX}] = e^{\omega b}\Gamma_X(\omega a). \tag{107}$$

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

$$\Gamma_Y(\omega) = E[e^{\omega Y}] = E[e^{\omega(X_1+X_2)}] = E[e^{\omega X_1}]E[e^{\omega X_2}] = \Gamma_{X_1}(\omega)\Gamma_{X_2}(\omega). \tag{108}$$

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_is are n independent and identically distributed (IID) random variables, then

$$\Gamma_Y(\omega) = E[e^{\omega Y}] = [\Gamma_{X_1}(\omega)]^n. \tag{109}$$

Let us now consider a Gaussian random variable X with parameters m and σ and density

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-m)^2/2\sigma^2} \qquad -\infty < x < \infty. \tag{110}$$

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 σ. Taking the first derivative and putting $\omega = 0$, we obtain

$$E[X] = \Gamma_X^{'}(0) = m. \tag{111}$$

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

$$E[X^2] = \Gamma_X^{(2)}(0) = \sigma^2 + m^2. \tag{112}$$

Thus,

$$var[X] = E[X^2] - E[X^2] = \sigma^2 + m^2 - m^2 = \sigma^2. \tag{113}$$

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 σ, then the random variable X defined as

$$X = \frac{Y-m}{\sigma} \tag{114}$$

is standard Gaussian.

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

$$\Gamma_X(\omega) = e^{(\omega^2/2)} \tag{115}$$

is the transform of a standard Gaussian random variable.

Homework 1.18

Show the consistency between the results obtained for transform of a Gaussian random variable, (107), (114) and (115). $\square$

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,

$$P(Y = y) = \sum_{i=1}^{N} p_i P(X_i = y) \tag{116}$$

or for continuous densities

$$f_Y(y) = \sum_{i=1}^{n} p_i f_{X_i}(y). \tag{117}$$

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 (100) 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 (100) 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 θ is real. This is a special case of (100) by setting $\omega = i\theta$.

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

1.14.1 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

$$\Pi_X(z) = E[z^X] = \sum_{i=0}^{\infty} p_i z^i, \tag{118}$$

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 1.19

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 ith derivative of $\Pi_X(z)$.
3. $E[X] = \lim_{z \to 1^-} \Pi_X^{(1)}(z)$. $\square$

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,

$$\Pi_X^{(i)}(1) = E[X(X-1)\ldots(X-i+1)]. \tag{119}$$

Homework 1.20

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

$$var[X] = \Pi_X^{(2)}(1) + \Pi_X^{(1)}(1) - (\Pi_X^{(1)}(1))^2. \quad \square \tag{120}$$

Homework 1.21

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

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

$$\Pi_Y(z) = z^b \Pi_X(za). \tag{121}$$

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

$$\Pi_Y(z) = \Pi_{X_1}(z)\Pi_{X_2}(z)\Pi_{X_3}(z) \ \ldots \ \Pi_{X_n}(z). \tag{122}$$

If $X_1, X_2,$ ..., X_n are also identically distributed, then

$$\Pi_Y(z) = [\Pi_{X_1}(z)]^n. \tag{123}$$

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

$$\Pi_X(z) = (1-p)z^0 + pz^1 = 1 - p + pz. \tag{124}$$

Its mean is $E[X] = \Pi_X^{(1)}(1) = p$ and by (120) its variance is $p(1-p)$.

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

$$\Pi_X(z) = p\sum_{i=1}^{\infty}(1-p)^{i-1}z^i = \frac{pz}{1-(1-p)z}. \tag{125}$$

Its mean is $E[X] = \Pi_X^{(1)}(1) = 1/p$ and by (120) 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

$$\Pi_X(z) = (1-p+pz)^n = [1+(z-1)p]^n. \tag{126}$$

Homework 1.22

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 (120) its variance is $np(1-p)$.

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

$$\Pi_X(z) = \sum_{i=0}^{\infty} p_i z^i = e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i z^i}{i!} = e^{(z-1)\lambda}. \tag{127}$$

Its mean is $E[X] = \Pi_X^{(1)}(1) = \lambda$ and by (120) its variance is also equal to λ.

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

$$\lim_{n\to\infty}(1-p+pz)^n = \lim_{n\to\infty}[1+(z-1)\lambda/n]^n = e^{(z-1)\lambda} \tag{128}$$

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.

1.14.2 Laplace Transform

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

$$\mathcal{L}_X(s) = E[e^{-sX}] = \int_0^\infty e^{-sx} f_X(x) dx. \quad (129)$$

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

$$\mathcal{L}_{X_1+X_2+\ \ldots\ +X_n}(s) = \mathcal{L}_{X_1}(s)\mathcal{L}_{X_2}(s)\ \ldots\ \mathcal{L}_{X_n}(s). \quad (130)$$

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

$$\mathcal{L}_Y(s) = e^{-sb}\mathcal{L}_X(sa). \quad (131)$$

The nth moment of random variable X is given by

$$E[X^n] = (-1)^n \mathcal{L}_X^{(n)}(0) \quad (132)$$

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

$$var[X] = E[X^2] - (E[X])^2 = (-1)^2\mathcal{L}_X^{(2)}(0) - ((-1)\mathcal{L}_X^{(1)}(0))^2 = \mathcal{L}_X^{(2)}(0) - (\mathcal{L}_X^{(1)}(0))^2. \quad (133)$$

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

$$\mathcal{L}_X(s) = \frac{\lambda}{\lambda + s}. \quad (134)$$

Homework 1.23

Derive (130)–(134) 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

$$\mathcal{L}_Y(s) = E[e^{-sY}] = E_N[E[e^{-s(X_1+X_2+\ \ldots\ +X_N|N)}]]. \quad (135)$$

Therefore, by independence of the X_i,

$$\mathcal{L}_Y(s) = E_N[E[e^{-sX_1} + E[e^{-sX_2} +\ \ldots\ + E[e^{-sX_N}]] = E_N[(\mathcal{L}_X(s))^N]. \quad (136)$$

Therefore

$$\mathcal{L}_Y(s) = \Pi_N[(\mathcal{L}_X(s))]. \quad (137)$$

An interesting example of (137) is the case where the X_i are IID exponentially distributed each with parameter λ, 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

$$\mathcal{L}_Y(s) = \Pi_N\left(\frac{\lambda}{\lambda + s}\right). \tag{138}$$

We also know that N is geometrically distributed, so $\Pi_N(z) = pz/[1 - (1 - p)z]$. Therefore, from (138), we obtain,

$$\mathcal{L}_Y(s) = \frac{\frac{p\lambda}{\lambda+s}}{1 - \frac{(1-p)\lambda}{\lambda+s}} \tag{139}$$

and after some algebra we obtain

$$\mathcal{L}_Y(s) = \frac{p\lambda}{s + p\lambda}. \tag{140}$$

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

Homework 1.24

Let X_1 and X_2 be exponential random variables with parameters λ_1 and λ_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 λ_1.

Hint

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

Homework 1.25

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 μ_1 and μ_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 μ_1 and μ_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_is are IID exponentially distributed random variables with parameter μ.

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. $\square$

1.15 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

$$\Gamma_X(\omega_1, \omega_2, \ \ldots, \ \omega_k) = E[s^{\omega_1 X_1, \omega_2 X_2, \ \ldots, \ \omega_k X_k}]. \tag{141}$$

1.16 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**:

$$P(X > C) \leq \frac{E[X]}{C}. \tag{142}$$

Homework 1.26

Prove Eq. (142).

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. (142) follows. $\square$

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

$$P(\mid X - E[X] \mid > C) \leq \frac{var[X]}{C^2}. \tag{143}$$

Homework 1.27

Prove Eq. (143).

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

$$P(\mid X-E[X] \mid > C^*\sigma) \leq \frac{1}{(C^*)^2} \tag{144}$$

for $C^* > 0$.

Homework 1.28

Prove and provide interpretation to Eq. (144).

Guide

Observe that the right-hand side of (144) is equal to $\frac{var[X]}{var[X](C^*)^2}$. □

Homework 1.29

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 $\epsilon > 0$,

$$P(\mid S_k - E[S_k] \mid < \theta\sigma(S_k) \ \text{ for all } \ k = 1, 2, \ldots, n) \geq 1 - \frac{1}{\theta^2}. \tag{145}$$

The interested reader may consult Feller [23] 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 ith 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.

1.17 Limit Theorems

Let $X_1, X_2, X_3, \ \ldots, \ X_k$ be a sequence of IID random variables with mean λ and variance σ^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_is are independent, we obtain

$$var[\bar{S}_k] = \frac{\sigma^2}{k}. \tag{146}$$

Homework 1.30

Prove Eq. (146). $\square$

Applying Chebyshev's inequality, we obtain

$$P(\mid \bar{S}_k - \lambda \mid \geq \varepsilon) \leq \frac{\sigma^2}{k\varepsilon^2} \ \text{ for all } \ \varepsilon > 0. \tag{147}$$

Noticing that as k approaches infinity, the right-hand side of (147) 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 λ. Then

$$P\left(\left|\frac{X_1 + X_2 + X_3 + \ \ldots \ + X_k}{k} - \lambda\right| \geq \varepsilon)\right) \to 0 \ \text{ as } \ k \to \infty \ \text{ for all } \ \varepsilon > 0. \tag{148}$$

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 1.10.6. Let $X_1, X_2, X_3, \ \ldots, \ X_k$ be k IID random variables with common mean λ and variance σ^2. Let random variable Y_k be defined as

$$Y_k = \frac{X_1 + X_2 + X_3 + \ \ldots \ + X_k - k\lambda}{\sigma\sqrt{k}}. \tag{149}$$

Then,

$$\lim_{k \to \infty} P(Y_k \leq y) = \Phi(y) \tag{150}$$

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 1.31

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

As we mentioned in Section 1.10.6, 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 λ and variance σ^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 λ_n and σ_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".

1.18 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".

1.18.1 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 no higher than α.

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 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

$$\sum_{i=C+1}^{N} \binom{N}{i} p^i (1-p)^{N-i} \leq \alpha. \tag{151}$$

Since the left-hand side of (151) 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 (151) for C values of $N-1$, $N-2$, ... until we find the first C value for which the inequality (151) 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 α.

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 σ.

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

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

$$C_G = Np + 3\sqrt{Np(1-p)}. \tag{152}$$

Note that α 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 α 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:

$$C_{opt} = \min\left[N, Np + 3\sqrt{Np(1-p)}\right]. \tag{153}$$

1.18.2 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 ith 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

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

has a Gaussian distribution with mean

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

and variance

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

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:

$$C_{opt} = \min\left[\sum_{i=1}^{N} R_i, E[\Sigma_R] + 3\sqrt{var[\Sigma_R]}\right]. \tag{154}$$

Homework 1.32

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}. $\square$

1.18.3 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

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

has a Gaussian distribution with mean

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

and variance

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

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

$$C_{opt} = \min\left[\sum_{i=1}^{N} R_{J(i)}(i), E[\Sigma_R] + 3\sqrt{var[\Sigma_R]}\right]. \tag{155}$$

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, 15, 52]. 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 $\{X_n,\ n = 1,\ 2,\ 3,\ \ldots\ \}$, where X_n is a random variable representing, for example, the number of calls arriving within the nth 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 $\{X_n,\ n = 1,\ 2,\ 3,\ \ldots\ \}$ is *strictly stationary* if for any subset of $\{X_n\}$, say, $\{X_{n(1)},\ X_{n(2)},\ X_{n(3)},\ \ldots,\ X_{n(k)}\}$, for any integer m the joint probability function $P(X_{n(1)},\ X_{n(2)},\ X_{n(3)},\ \ldots,\ X_{n(k)})$, is equal to the joint probability function

$P(X_{n(1)+m},\ X_{n(2)+m},\ X_{n(3)+m},\ \ldots,\ X_{n(k)+m})$. In other words, $P(X_{n(1)+m},\ X_{n(2)+m},\ X_{n(3)+m},\ \ldots,\ X_{n(k)+m})$ 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 moments of order higher than two are zero, and therefore, 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,

$$E[X] = E[X_n]. \tag{156}$$

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,

$$U(k) = cov(X_n, X_{n+k}). \tag{157}$$

Notice that, the case of $k = 0$ in Eq. (157), namely,

$$U(0) = cov(X_n, X_n) = var[X_n] \tag{158}$$

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

$$U(-k) = U(k) \tag{159}$$

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:

$$S = \sum_{i=1}^{\infty} U(i) \tag{160}$$

and

$$v = \sum_{i=-\infty}^{\infty} U(i). \tag{161}$$

Notice that

$$v = 2S + var[X_n]. \tag{162}$$

Another important definition of the AVR which justifies its name is

$$v = \lim_{n\to\infty} \frac{var[S_n]}{n}. \tag{163}$$

We will further discuss these concepts in Section 18.1.

Homework 2.1

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

$$\lim_{n\to\infty} \frac{var[S_n]}{n} = 2S + var[X_n] \tag{164}$$

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 (96), we obtain

$$var(S_k^*) = k^* var[X_n] + 2k^* S$$

which leads to (164). □

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:

$$C(k) = \frac{U(k)}{var[X_n]}. \tag{165}$$

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

$$\hat{E}[A] = \frac{1}{m}\sum_{n=1}^{m}\hat{A}_n, \tag{166}$$

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

$$\hat{U}(k) = \frac{1}{m-k}\sum_{n=k+1}^{m}(\hat{A}_{n-k} - E[A])(\hat{A}_n - E[A]). \tag{167}$$

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 ith 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 interarrival times Δ_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:

$$\lim_{\Delta t \to 0} P(X(t+\Delta t) - X(t) > 1 \mid X(t+\Delta t) - X(t) \geq 1) = 0. \tag{168}$$

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. At each time-slot we conduct a Bernoulli experiment. Then for each time-slot i, we conduct a 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

$$P[N(n) = i] = \binom{n}{i} p^i (1-p)^{n-i} \qquad i = 0,\ 1,\ 2,\ \ldots,\ n. \tag{169}$$

The concept of an interarrival 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 interarrival 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 interarrival times are IID, so we can drop the index i of Δ_i, designating the i interarrival time, and consider the probability function of the random variable Δ representing any interarrival time. Because Δ represents a number of Bernoulli trials until a success, it is geometrically distributed, and its probability function is given by

$$P(\Delta = i) = p(1-p)^{i-1} \qquad i = 1,\ 2,\ \ldots\ . \tag{170}$$

Another important statistical measure is the time it takes n until the ith arrival. This time is a sum of i interarrival 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

$$P[\text{the } i\text{th arrival occurs in time slot } n] = \binom{n-1}{i-1} p^i (1-p)^{n-i} \qquad i = i,\ i{+}1,\ i{+}2,\ \ldots\ . \tag{171}$$

The reader may notice that the on-off sources discussed in Section 1.18 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. $\square$

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 ith 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

$$P_a = 1 - \prod_{i=1}^{N}(1 - p_i). \tag{172}$$

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 λt.

These three conditions will be henceforth called the *Three Poisson process conditions*.

By definition, the Poisson process has what is called *stationary increments* [56, 67], that is, for any $t_2 > t_1$, the random variable $X(t_2) - X(t_1)$, and the random variable $X(t_2 + u) - X(t_1 + u)$ 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 chace of having a occurrence. Therefore, occurrences are equally likely to happen at all times. This property is also called *time-homogeneity* [13].

Another important property of the Poisson process is that the inter-arrival times of occurrences is exponentially distributed with parameter λ. 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 interarrival time is given by

$$E[T] = \frac{1}{\lambda}. \tag{173}$$

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* [67], 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 λ. 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 λ. 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. $\square$

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 λ_1 and λ_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 λ_1 and the other with parameter λ_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 (51)

$$P(T > t) = e^{-(\lambda_1+\lambda_2)t}. \tag{174}$$

Thus, the interarrival 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 $\lambda_1 + \lambda_2 +, \ldots, +\lambda_N$. $\square$

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 λ_1 and λ_2, respectively, and we are interested in the probability of $T_1 < T_2$. By (52),

$$P(T_1 < T_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \tag{175}$$

Before we introduce further properties of the Poisson process, we shall introduce the following definition: a function $g(\cdot)$ is $o(\Delta t)$ if

$$\lim_{\Delta t \to 0} \frac{g(\Delta t)}{\Delta t} = 0. \tag{176}$$

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 Δt, 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 $X(\Delta t)$ has a Poisson distribution, therefore

$$P(N(\Delta t) = 0) = e^{-\lambda\Delta t} \tag{177}$$

and developing it into a series gives,

$$P(X(\Delta t) = 0) = 1 - \lambda\Delta t + o(\Delta t). \tag{178}$$

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. □

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 follows a Poisson process with parameter λ. 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 $N(t)$. $\square$

It may be interesting to notice that the interarrival 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 interarrival times are Erlang distributed with parameter λ and 2.

Homework 2.7

1. Prove the last statement.
2. Derive the mean and the variance of the interarrival 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 interarrival 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. $\square$

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

$$E[X(t)] = \frac{\int_0^T X(t)dt}{T} \tag{179}$$

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 λ. 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 λT. $\square$

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 λ_i for a period of time that is exponentially distributed with parameter δ_i. Then it moves to mode (state) j where it behaves like a Poisson process with parameter λ_j for a period of time that is exponentially distributed with parameter δ_j. The parameters are called *mode duration parameters* [77][78],[79]. 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: λ_0, λ_1, δ_0, and δ_1. The MMPP(2) behaves as a Poisson process with parameter λ_0 for a period of time that is exponentially distributed with mode duration parameter δ_0. Then moves to mode 1 where it behaves like a Poisson process with mode duration parameter λ_1 for a period of time that is exponentially distributed with parameter δ_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 Θ. 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,

$$P(X_{n+1} = j \mid X_n = i) = P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \ldots). \tag{180}$$

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

$$P_{ij} = P(X_{n+1} = j \mid X_n = i) \ \text{ for all } n. \tag{181}$$

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 (181), let us define the n-step transition probability from state i to state j as

$$P_{ij}^{(n)} = P(X_n = j \mid X_0 = i). \tag{182}$$

The following is known as the Chapman-Kolmogorov equation:

$$P_{ij}^{(n)} = \sum_{k \in \Theta} P_{ik}^{(m)} P_{kj}^{(n-m)}, \tag{183}$$

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

$$\pi_n(j) = \sum_{k \in \Theta} P_{kj} \pi_{n-1}(k), \tag{184}$$

or based on the initial distribution by

$$\pi_n(j) = \sum_{k \in \Theta} P_{kj}^{(n)} \pi_0(k), \quad j \in \Theta \tag{185}$$

or, in matrix notation

$$\pi_n(j) = \sum_{k \in \Theta} P_{kj}^{(n)} \pi_0(k), \tag{186}$$

Let the vector Π_n be defined by $\Pi_n = \{\pi_n(j),\ j = 0,\ 1,\ 2,\ 3,\ \ldots\}$. The vector Π_n can be obtained by

$$\Pi_n = \Pi_{n-1}\mathbf{P} = \Pi_{n-2}\mathbf{P}^2 = \ \ldots\ = \Pi_0 \mathbf{P}^n. \tag{187}$$

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

$$P_{ji}^{(n)} > 0. \tag{188}$$

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. $\square$

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. $\square$

The latter implies that the state space Θ 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. $\square$

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. $\square$

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

$$R_i = \min\{n : X_n = i \mid X_0 = i\}. \tag{189}$$

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 than 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.

$$\mathbf{\Pi} = \lim_{\mathbf{n}\to\infty} \mathbf{\Pi_n} = \lim_{\mathbf{n}\to\infty} \mathbf{\Pi_0 P^n} \tag{190}$$

and it satisfies

$$\mathbf{\Pi} = \textbf{row of } \lim_{\mathbf{n}\to\infty} \mathbf{P^n} \tag{191}$$

where *row of* $\lim_{n\to\infty} \mathbf{P}^n$ is any row of the matrix $\mathbf{P}^n$ as n approaches ∞. 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 π_j, $j = 0, 1, 2, \ldots$, the components of the vector $\mathbf{\Pi}$. That is, π_j is the steady-state probability of the Markov-chain to be at state j. Namely,

$$\pi_j = \lim_{n\to\infty} \pi_n(j) \quad \text{for all } j. \tag{192}$$

By equation (184), we obtain

$$\pi_n(j) = \sum_{i=0}^{\infty} P_{ij}\pi_{n-1}(i), \tag{193}$$

then by the latter and (192), we obtain

$$\pi_j = \sum_{i=0}^{\infty} P_{ij}\pi_i. \tag{194}$$

Therefore, recalling that π 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:

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij} \quad \text{for all } j, \tag{195}$$

$$\sum_{j=0}^{\infty} \pi_j = 1 \tag{196}$$

and

$$\pi_j \geq 0 \qquad \text{for all } j. \tag{197}$$

In this case:

$$\pi_j = \frac{1}{E[R_j]}. \tag{198}$$

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 ith 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 Θ is finite, one of the equations in (195) is redundant and replaced by (196).

In matrix notation equation (195) is written as: $\mathbf{\Pi} = \mathbf{\Pi}\mathbf{P}$.

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 (195) exists. However, in this case, the jth component of $\mathbf{\Pi}$, namely π_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 (195), we obtain the following steady-state equations:

$$\pi_j \sum_{i=0}^{\infty} P_{ji} = \sum_{i=0}^{\infty} \pi_i P_{ij} \quad \text{for } j = 0, 1, 2 \cdots . \tag{199}$$

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 (198).

6. Consider a binary communication channel between a transmitter and a receiver where B_n is the value of the nth 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 (195), we obtain,

$$p_0 P_{01} = p_1 P_{10}.$$

Then substituting the latter in the second equation of (195), we obtain

$$p_1 P_{12} = p_2 P_{21}.$$

Continuing in the same way, we obtain

$$p_i P_{i,i+1} = p_{i+1} P_{i+1,i}, \quad i = 0, 1, 2, \cdots . \tag{200}$$

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 (195).

Homework 2.18

Solve the local balance equations together with the normalizing equations for the p_i, $i = 0, 1, 2, \cdots$.

Guide

Recursively, write all p_i, $i = 0, 1, 2, 3, \cdots$ 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

$$Q_{ij} = P(X_n = j \mid X_{n+1} = i). \tag{201}$$

By the definition of conditional probability, we obtain,

$$Q_{ij} = \frac{P(X_n = j \bigcap X_{n+1} = i)}{P(X_{n+1} = i)} \tag{202}$$

or

$$Q_{ij} = \frac{P(X_n = j)P(X_{n+1} = i \mid X_n = j)}{P(X_{n+1} = i)} \tag{203}$$

and if π_j denotes the steady-state probability of the Markov-chain $\{X_n\}$ to be in state j, and let $n \to \infty$, we obtain

$$Q_{ij} = \frac{\pi_j P_{ji}}{\pi_i}. \tag{204}$$

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 (204), we obtain,

$$\pi_i P_{ij} = \pi_j P_{ji} \text{ forall } i, j. \tag{205}$$

The set of equations (205) is also a necessary and sufficient condition for time reversibility. This set of equations is called the *detailed balance conditions*. Note that the condition of (205) is equivalent to the same condition if we only consider adjacent pairs i, j, namely, pair such that P_{ij} and/or P_{ji} are non-zero. If they are both equal to zero then (205) is satisfied by default.

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 4×4 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 nth bit is received correctly, and let $Y_n = 0$ if the nth 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 $k \geq 0$,

$$P(X_{t+s} = j \mid X_t = i, X_v = k_v, v \leq t) = P(X_{t+s} = j \mid X_t = i). \tag{206}$$

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 δ_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:

$$\begin{array}{ll} (1) & P_{ii} = 0 \;\text{ for all } i \\ (2) & \quad \sum_j P_{ij} = 1. \end{array}$$

An example of a continuous-time Markov-chain is a Poisson process with rate λ. 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 λ.

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 λ for each of these exponential intervals, this parameter depends of the state of the process and it is denoted δ_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 δ_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 δ_i, and at any point of occurrence, the process increases by one (from its previous value i to $i + 1$) with probability υ_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

$$U_0 = \frac{1}{b_0}. \tag{207}$$

and

$$U_i = \frac{1}{\delta_i} + \vartheta_i[U_{i-1} + U_i]. \tag{208}$$

Homework 2.23

Explain equations (207) and (208).

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 (208) 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,

$$U_i = \frac{1}{b_i + d_i} + \frac{d_i}{b_i + d_i}[U_{i-1} + U_i] \tag{209}$$

or

$$U_i = \frac{1}{b_i} + \frac{d_i}{b_i} U_{i-1}. \tag{210}$$

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,

$$P_{ij}(t) = P[X(t_0 + t) = j \mid X(t_0) = i]. \tag{211}$$

The continuous time version of the Chapman-Kolmogorov equations are

$$P_{ij}(t + \tau) = \sum_{n=0}^{\infty} P_{in}(t) P_{nj}(\tau) \;\; \text{for all } t \ge 0, \tau \ge 0. \tag{212}$$

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:

$$P'_{ij}(t) = \sum_{n \ne i} \delta_i P_{in} P_{nj}(t) - \delta_i P_{ij}(t) \text{ for all } i, j \text{ and } t \ge 0. \tag{213}$$

For a birth-and-death process the latter become

$$P'_{0j}(t) = b_0\{P_{1j}(t) - P_{0j}(t)\}. \tag{214}$$

and

$$P'_{ij}(t) = b_i P_{i+1,j}(t) + d_i P i - 1, j(t) - (b_i + d_i) P_{ij}(t) \text{ for all} i > 0. \tag{215}$$

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 the process starts from any state, the random variable that represents the time it returns to that state has finite mean. As in the case of discrete-time Markov-chain a Markov-chain is said to be *stable* if all its states are positive recurrent.

Henceforth we only consider a 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

$$\pi_j = \lim_{t\to\infty} P_{ij}(t). \tag{216}$$

The π_j values are called steady-state probabilities or stationary probabilities of the continuous-time Markov-chain. In particular, π_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 π_js 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

$$Q_{ij} = \delta_i P_{ij} \text{ for } i \neq j \tag{217}$$

and

$$Q_{ii} = -\sum_{j\neq i} Q_{ij}. \tag{218}$$

Remarks:

- The state-space can be finite or infinite and hence the matrices $\mathbf{P}$ and $\mathbf{Q}$ can also be finite or infinite.
- In Eq. (217), 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 π_js, we solve the following set of steady-state equations:

$$0 = \sum_i \pi_i Q_{ij} \quad \text{for all } j \tag{219}$$

and

$$\sum_j \pi_j = 1. \tag{220}$$

Denoting $\mathbf{\Pi} = [\pi_{\mathbf{0}}, \pi_{\mathbf{1}}, \pi_{\mathbf{2}}, \ \ldots]$, Eq. (219) can be written as

$$0 = \mathbf{\Pi Q}. \tag{221}$$

To explain Eqs. (219), notice that, by (217) and (218), for a particular j, the equation

$$0 = \sum_i \pi_i Q_{ij} \tag{222}$$

is equivalent to the equations

$$\pi_j \sum_{i \neq j} Q_{ji} = \sum_{i \neq j} \pi_i Q_{ij} \tag{223}$$

or

$$\pi_j \sum_{i \neq j} \delta_j P_{ji} = \sum_{i \neq j} \pi_i \delta_i P_{ij} \tag{224}$$

which the following global balance equations if we consider all j.

$$\pi_j \sum_{i \neq j} \delta_j P_{ji} = \sum_{i \neq j} \pi_i \delta_i P_{ij} \quad \text{for all } j, \tag{225}$$

or using the Q_{ij} notation,

$$\pi_j \sum_{i \neq j} Q_{ji} = \sum_{i \neq j} \pi_i Q_{ij} \quad \text{for all } j. \tag{226}$$

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. (222) 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 return to all states infinitely many times, during 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. (224) 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 (219) and (220) is dependent and one of the equations in (219) is redundant in the finite state space case.

Due to the fact that the $\mathbf{Q}$ matrix may be too large, it may not be possible to solve the steady-state equations (219). Actually, the case of a large state-space (or large $\mathbf{Q}$ matrix) is common in practice. Consider for example a 49 cell GSM mobile network, and assume that every cell has 23 voice channels. Assuming Poisson arrivals and exponential holding 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.

For the case of continuous-time birth-and-death process, $Q_{ij} = 0$ for $|i - j| > 0$. As in the discrete-time case, under this special condition, the global balance equations (225) can be simplified to the local balance equations. We start with the first equation of (225) and using the condition $Q_{ij} = 0$ for $|i - j| > 0$, we obtain

$$\pi_0 Q_{01} = \pi_1 Q_{10} \tag{227}$$

The second equation is

$$\pi_1 [Q_{10} + Q_{12}] = \pi_0 Q_{01} + \pi_2 Q_{21}. \tag{228}$$

Then Eq. (228) can be simplified using (227) and we obtain

$$\pi_1 Q_{12} = \pi_2 Q_{21}. \tag{229}$$

In a similar way, by repeating the process, we obtain the following local balance equations.

$$\pi_i Q_{i,i+1} = \pi_{i+1} Q_{i+1,i} \quad i = 0, 1, 2, \ \cdots \ . \tag{230}$$

2.5.6 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 π_is in a continuous-time Markov-chains), 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 we do not need a separate Poisson inspector, but we could inspect the process at occurrences of any given independent Poisson process which is part of the continuous Markov-chain. 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.

2.5.7 Reversibility

We have discussed the **time reversibility** concept in the context of discrete-time Markov-chain. In the case of 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 continuous-time Markov-chain that has a unique steady-state solution and that its $[P_{ij}]$ matrix would give rise to 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 as the original process, the reverse process spends an exponentially distributed amount of time with parameter δ_i when in state i.

$$P\{X(t) = i,\ \text{for } t \in [u-v, u] \mid X(u) = i\} \quad = \quad \frac{P\{X(t) = i,\ \text{for } t \in [u-v, u] \mid X(u) = i\}}{P[X(u) = i]}$$

$$= \frac{P[X(u-v)=i]e^{-\delta_i v}}{P[X(u)=i]} = e^{-\delta_i v}.$$

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 continuous-time Markov-chain is that its embedded chain is time reversible. As we have learned when we discussed reversibility of discrete-time Markov-chain, 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:

$$\hat{\pi}_i P_{ij} = \hat{\pi}_j P_{ji} \text{ for all adjacent } i,j. \tag{231}$$

The equivalent condition in the case of a continuous-time Markov-chain is the existence of positive π_i for all states i that sum up to unity that satisfy the detailed balance equations of a continuous-time Markov-chain, defined as:

$$\pi_i Q_{ij} = \pi_j Q_{ji} \text{ for all adjacent } i,j. \tag{232}$$

Homework 2.25

Derive (232) from (231). □

It is important to notice that a birth-and-death process that 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,

$$\frac{T_{i,i+1}(L)}{L} = \frac{T_{i+1,i}(L)}{L}. \tag{233}$$

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

$$\pi_i Q_{i,i+1} = \frac{T_{i,i+1}(L)}{L} = \frac{T_{i+1,i}(L)}{L} = \pi_{i+1} Q_{i+1,i}, \tag{234}$$

so our birth-and-death process is time reversible. This is an important result for the present context because many of our queueing models are special cases of the birth-and-death process.

2.5.8 Multi-Dimensional Continuous Time Markov-chains

The extension discussed earlier to multi-dimensional discrete-time Markov-chain applies also to the case of continuous-time Markov-chain. 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 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.

3 General Queueing Concepts

In general, queueing systems may be characterized by complex input process, service time distribution 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. This modelling simplification is commonly made in the context of packet switching networks like the Internet that are based on the store and forward principle. Typically, packets on their ways to their destinations arrive at a router where they are stored and further forwarded according to addresses in their headers. One of the most fundamental elements in this process is the single-server queue. 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 [41], 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 / waiting room

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), G (General), GI (General and independent), and Geom (Geometric). The fourth position is used for the number of buffer places in addition to the number of servers and it is usually not used if the waiting room is unlimited.

For example, M/M/1 denotes a single-server queue with Poisson arrival process and exponential service time with infinite buffer. M/G/k/k denotes a k-server queue with no additional waiting room accept at the servers with the arrival process being Poisson.

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 collected. 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/∞ 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, a single-server queue 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., μ denotes the service rate. Further, let λ be the mean 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: λL. The average number of customers (amount of work) that has been served during time period L is equal to $\mu\hat{U}L$. Since L is large and the queue is stable, these two values are 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, jobs or packets) in the system. Let p_n be the probability that there are n in the system. Having the utilization, we can readily obtain p_0 the probability that the system is empty. For the G/G/1 queue we have,

$$p_0 = 1 - \hat{U} = 1 - \lambda/\mu. \tag{235}$$

If we have a multi-server queue, e.g. G/G/k/$k+n$, then 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 utilization.

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* [48, 71, 72]. It has two forms. The first form is:

$$E[Q] = \lambda E[D] \tag{236}$$

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 complete, respectively. In 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:

$$E[N_Q] = \lambda E[W_Q] \tag{237}$$

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. (236) 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 $\lambda E[D]$.

To obtain (237) from (236), notice that

$$E[Q] = E[N_Q] + \hat{U} = E[N_Q] + \lambda/\mu \tag{238}$$

and

$$E[D] = E[W_Q] + 1/\mu. \tag{239}$$

Substituting (238) and (239) in (236), (237) follows.

For a graphical proof of Little's Formula for the case of G/G/1 queue see [12]. 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

$$E[Q] = \frac{1}{L}\int_0^L Q(t)dt. \tag{240}$$

Also notice that

$$\int_0^L Q(t)dt = \sum_{i=1}^{A(L)} W_i \tag{241}$$

where W_i is the time spent in the system by the ith 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,

$$\frac{1}{L}\int_0^L Q(t)dt = \frac{1}{L}\sum_{i=1}^{A(L)} W_i \tag{242}$$

and realizing that

$$\lambda = A(L)/L, \tag{243}$$

and

$$E[D] = \frac{1}{A(L)}\sum_{i=1}^{A(L)} W_i, \tag{244}$$

we obtain

$$E[Q] = \frac{1}{L}\int_0^L Q(t)dt = \frac{A(L)}{L}\frac{1}{A(L)}\sum_{i=1}^{A(L)} W_i = \lambda E[D]. \tag{245}$$

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, we must remember that we deal here only with systems that are in steady-state. The system must be in steady-state for Little's formula to apply.

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 so this is the mean time spent in the system that includes just the server. The mean arrival rate into that system must be equal to λ 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 π_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

$$\hat{U} = \lambda/\mu.$$

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 [33, 61]. 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:

$$\hat{U} = (1 - P_b)\lambda\mu^{-1}, \tag{246}$$

where μ^{-1} is the mean service time. Let $\rho = \lambda/\mu$, we obtain

$$P_b = 1 - \frac{\hat{U}}{\rho}. \tag{247}$$

3.4 Work Conservation

Another important concept in queuing theory is the concept of *work conservation*. A queuing system is said to be work conservative if no server is idle if there is 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.

3.5 PASTA

Many of the queueing models we consider in this book involve Poisson arrival processes. The Poisson Arrivals See Time Averages (PASTA) property discussed in the previous section is important for analysis and simulations of such queueing models. Let us further explain and prove this important property.

The PASTA property means 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. 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. Of course, the arrival process of this person is not a Poisson process as there are no arrivals during the time the person is in the house.

In addition to the Poisson arrival assumption, 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 (SSQ) 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.

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' formula and the condition that arrivals after time t are independent of $Q(t)$, we obtain that

$$A_k(t) = P[Q(t) = k]. \tag{248}$$

Then, by taking the limit of both sides of (248), 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 3.1

Prove Eq. (248). □

3.6 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 interarrival 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 σ_a^2 the observed variance of these n observations. Their values are given by

$$\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{249}$$

and

$$\sigma_a^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})^2. \tag{250}$$

Then the confidence interval of $\bar{a}$, with confidence α, $0 \leq \alpha \leq 1$, is given by $(\bar{a} - U_r, \bar{a} + U_r)$, where

$$U_r = \{t_{(1-\alpha)/2,(n-1)}\}\frac{\sigma_a}{\sqrt{n}} \tag{251}$$

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 $t_{0.025,10} = 2.23$.

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* [26]. 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 interarrival 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.

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. □

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. □

There will be many homework assignments in this book that require simulations and in some cases a guide will be provided.

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 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 λ being the arrival rate and μ 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 interarrival 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 λ and μ satisfy the stability condition $\lambda < k\mu$, there must exist an integer $\hat{n}$, $1 \leq \hat{n} \leq k$ such that

$$(\hat{n} - 1)\mu < \lambda \leq \hat{n}\mu, \tag{252}$$

or equivalently

$$\frac{\hat{n} - 1}{\lambda} < \frac{1}{\mu} \leq \frac{\hat{n}}{\lambda}. \tag{253}$$

Homework 5.1

Show that

$$\hat{n} = \left\lceil \frac{\lambda}{\mu} \right\rceil \tag{254}$$

satisfies $1 \leq \hat{n} \leq k$ and (253). 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

$$E[Q] = \frac{\lambda}{\mu}. \tag{255}$$

Homework 5.2

Perform the algebraic operations that lead to (255). □.

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 λ, 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

$$\hat{U} = \frac{\lambda}{k\mu}. \tag{256}$$

Notice that Equations (255) and (256) applies also to D/D/∞ for finite λ and μ. Eq. (255) gives the mean queue-size of D/D/∞ (by Little's formula, or by following the arguments that led to Eq. (255)) and for D/D/∞, we have that $\hat{U} = 0$ by (256). Also notice that in D/D/∞ 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 kth 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

$$P_b = \frac{\left\lceil \frac{\lambda}{\mu} \right\rceil - k}{\left\lceil \frac{\lambda}{\mu} \right\rceil}. \tag{257}$$

Let $A = \lambda/\mu$, the mean-queue size is obtained using Little's formula to be given by

$$E[Q] = \frac{\lambda}{\mu}(1 - P_b) = \frac{kA}{\lceil A \rceil}. \tag{258}$$

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

$$\hat{U} = \frac{E[Q]}{k} = \frac{A}{\lceil A \rceil}. \tag{259}$$

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,

$$P(Q = k - 1) = \frac{\frac{k(\lceil A \rceil - A)}{\lambda}}{\frac{\lceil A \rceil}{\lambda}}. \tag{260}$$

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

$$P(Q = k) = \frac{\frac{\lceil A \rceil}{\lambda} - \frac{k(\lceil A \rceil - A)}{\lambda}}{\frac{\lceil A \rceil}{\lambda}}. \tag{261}$$

Homework 5.4

1. Show that the results for the queue-size probabilities $P(Q = k - 1)$ and $P(Q = k)$ in (260) and (261) are consistent with the result for the mean queue-size in (258). 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 (258) and the results for $P(Q = k - 1)$ and $P(Q = k)$ in (260) and (261). Use it also to confirm the results you obtained for the D/D/3/3 queue.

4. Consider a D/D/1/k queue. 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	$\lambda < \mu$	λ/μ	λ/μ
D/D/1	$\lambda > \mu$	∞	1
D/D/k	$\lambda < k\mu$	$A = \lambda/\mu$	A/k
D/D/k	$\lambda > k\mu$	∞	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 λ and μ 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 interarrival 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 λ and service times are assumed to be IID and exponentially distributed with parameter μ, 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 λ 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 λ, and the other is the time until the next departure - exponentially distributed with parameter μ. 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. (217) and (218), 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. (219) and performing some simple algebraic operations, we obtain the following steady-state equations for the M/M/1 queue.
$\pi_0\lambda = \pi_1\mu$
$\pi_1\lambda = \pi_2\mu$
. . .
and in general:

$$\pi_i\lambda = \pi_{i+1}\mu, \text{ for } i = 0,\ 1,\ 2,\ \ldots\ . \tag{262}$$

To explain (262) intuitively, Let L be a very long time. During L, the total time that the process stays in state i is equal to $\pi_i L$. Since the arrival process is Poisson, the mean number of transitions from state i to $i + 1$ is equal to $\lambda\pi_i L$. For $i = 0$, $\lambda\pi_0 L$ is also the mean number of events that occur during L (because there are no departures at state 0). However, for $i \geq 1$, 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. Therefore, the mean number of departures that occur in state i (transitions from i to $i-1$) during L is equal to the product of the mean number of events $(\lambda + \mu)\pi_i L$ and the probability that an event is a departure which is $\mu/(\lambda + \mu)$, namely

$$(\lambda + \mu)\pi_i L \frac{\mu}{\lambda + \mu} = \mu \pi_i L.$$

In a similar way, the mean number of arrivals that occur in state i during L is

$$(\lambda + \mu)\pi_i L \frac{\lambda}{\lambda + \mu} = \lambda \pi_i L.$$

Since L is very long, we must have that the number of transitions from state i to $i+1$ is equal to the number of transitions from state $i+1$ to i. Therefore, $\pi_i \lambda L = \pi_{i+1}\mu L$, and dividing both sides by L, we obtain (262).

Of course, the sum of the steady-state probabilities must be equal to one, so we have the additional equation

$$\sum_{j=0}^{\infty} \pi_j = 1. \tag{263}$$

Let $\rho = \lambda/\mu$, 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:

$$\pi_i = \rho^i \pi_0 \text{ for } i = 0,\ 1,\ 2,\ \ldots\ . \tag{264}$$

As M/M/1 is a special case of G/G/1, we can use Eq. (235) to obtain $\pi_0 = 1 - \rho$, so

$$\pi_i = \rho^i (1 - \rho) \text{ for } i = 0,\ 1,\ 2,\ \ldots\ . \tag{265}$$

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:

$$E[Q] = \frac{\rho}{1 - \rho}. \tag{266}$$

Homework 6.1

Perform the algebraic operations that lead to (266). □

6.2 Delay Statistics

By (265), 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 μ, 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. (140) in Section 1.14.2). Therefore the total delay of any arriving customer in an M/M/1 system must be exponentially distributed. This can also be intuitively explained. Because both geometric and exponential distributed random variables are memoryless, a geometrically distributed sum of IID exponential random variables is also memoryless. And since the exponential is the only continuous memoryless distribution, the total delay of any arriving customer 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 (266) invoking Little's formula. Another way to obtain the mean delay is by noticing from (265) 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 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,

$$E[D] = \frac{1}{(1-\rho)\mu} = \frac{1}{\mu-\lambda}. \tag{267}$$

Homework 6.3

Verify that (266) and (267) 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

$$\delta_D(x) = \begin{cases} (\mu-\lambda)e^{(\lambda-\mu)x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{268}$$

6.3 Using Z-Transform

The Z-transform defined in Section 1.14, 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 nth equation of (262) by z^n. Summing up both sides will give

$$\frac{\Psi(z)-\pi_0}{z} = \rho\Psi(z) \tag{269}$$

where $\Psi(z) = \sum_{i=0}^{\infty} \pi_i z^i$. Letting z approach 1 (from below) gives

$$\pi_0 = 1-\rho \tag{270}$$

which is consistent with what we know already. Substituting it back in (269) gives after simple algebraic manipulation:

$$\Psi(z) = \frac{1-\rho}{1-\rho z}. \tag{271}$$

Taking derivative and substituting $z = 1$, after some algebra we obtain

$$E[Q] = \Psi^{(1)}(1) = \frac{\rho}{1-\rho} \tag{272}$$

which is again consistent with what we know about M/M/1 queue.

Homework 6.4

1. Derive equations (269) – (272).
2. Derive the variance of the M/M/1 queue-size using Z-transform. □

6.4 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 (265)–(268) is that while the queue-size statistics are dependent only on ρ (the ratio of the arrival rate and 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 λ to $N\lambda$ and we increase the service rate from μ to $N\mu$ (maintaining the same ρ), 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 λ and service rate μ. 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 μ) 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 λ there is a service rate μ such that our delay-related QoS measure is just met, then if the arrival rate increases from λ to $N\lambda$, and we aim to find the service rate μ^* such that the delay-related QoS measure is just met, we will need to make sure that the spare capacity is maintained, that is

$$\mu - \lambda = \mu^* - N\lambda \tag{273}$$

or

$$\mu^* = \mu + (N-1)\lambda \tag{274}$$

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

$$M_{mg} = \frac{N\mu - \mu^*}{N\mu} \tag{275}$$

so by (274), we obtain

$$M_{mg} = \frac{N-1}{N}(1-\rho). \tag{276}$$

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. (276) 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 λ and μ. 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. (276).

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 ρ 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 ρ 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

$$E[Q_{TDMA}] = \frac{\hat{\rho}}{1-\hat{\rho}}, \tag{277}$$

and the mean delay is

$$E[D_{TDMA}] = \frac{1}{\hat{\mu}-\hat{\lambda}}. \tag{278}$$

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

$$E[Q_{FMUX}] = \frac{\hat{\rho}}{1-\hat{\rho}} = E[Q_{TDMA}]. \tag{279}$$

However, we can observe an N-fold reduction in the mean delay:

$$E[D_{FMUX}] = \frac{1}{N\hat{\mu}-N\hat{\lambda}} = \frac{E[D_{TDMA}]}{N}. \tag{280}$$

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 6.5

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 (bitrate) 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. (278) and substitute $E[D_{TDMA}] = 0.05$ and $\hat{\lambda} = 1000$, to compute $\hat{\mu}$. This gives $\hat{\mu} = 1020$ [packets/s] or bitrate of 1.02 Mb/s per each user. For 100 users the required rate is 102,000 packets/s or bitrate 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. □

6.5 The Departure Process

According to the so-called Burke theorem [16], in steady-state, the departure process of a stable M/M/1, where $\rho < 1$, is a Poisson process with parameter λ and is independent of the number in the queue after the departures occur. To see why this is so all we need is to realize that the queue size process in an M/M/1 queue with $\rho < 1$ is an irreducible, aperiodic and stable Markov-chain. As such it must be reversible. Therefore, the points in time of arrivals in the forward process correspond to points in time that the Markov-chain is increased by one. These points represent the arrival process and therefore follow a Poisson process. By reversibility, in steady-state, the arrival process of the reversed process must also follow Poisson process with parameter λ and this process is the departure process of the forward process. Therefore the departure process is Poisson and 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 λ, we also know that, in steady-state, the inter-departure times are also exponentially distributed with parameter λ. 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 ρ. If the queue is non-empty, the time until the next departure is exponentially distributed with parameter μ – 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 λ 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 λ and the other with parameter μ. Let U_1 and U_2 be two independent exponential random variables with parameters λ and μ, 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

$$f_D(t) = \rho\mu e^{-\mu t} + (1-\rho) f_U(t). \tag{281}$$

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 (281), we obtain

$$f_D(t) = \rho\mu e^{-\mu t} + (1-\rho)\frac{\lambda\mu}{\mu-\lambda}\left(e^{-\lambda t} - e^{-\mu t}\right) \tag{282}$$

which after some algebra gives

$$f_D(t) = \lambda e^{-\lambda t}. \tag{283}$$

This result is consistent with Burke theorem.

Homework 6.6

Complete all the algebraic details in the derivation of equations (281) – (283). □

Another way to show consistency with Burke theorem is the following. Consider a stable ($\rho < 1$) M/M/1 queue. Let d_ϵ be the unconditional number of departures, in steady state, that leave the M/M/1 queue during a small interval of time of size ϵ, 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 ϵ 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 λ in steady-state.

Homework 6.7

So far we have discussed the behaviour of the M/M/1 departure process in steady-state. You are now asked demonstrate that the M/M/1 departure process may not be Poisson with parameter λ if we do not assume steady-state condition. Consider an M/M/1 system with arrival rate λ and service rate μ, 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 λt 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 λt 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) $D(t) = X_p(\lambda t) - 1$ (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. (214) and (215). Then notice that

$$D(t) = P_{00}(t)X_p(\lambda t) + [1 - P_{00}(t)][X_p(\lambda t) + 1]. \quad \square$$

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

$$D_k(t) = P[Q(t) = k]. \tag{284}$$

Then, by taking the limit of both sides of (284), we show that the queue size seen by a leaving customer is statistically identical to the queue size seen by an independent observer. □

Homework 6.8

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 ρ values). Plot all the results in a graph including the confidence intervals (bars). □

6.6 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 λ. 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 ρ. Considering also that $E[T_I] = 1/\lambda$, we obtain

$$\frac{E[T_B]}{E[T_B] + \frac{1}{\lambda}} = \rho. \tag{285}$$

Therefore,

$$E[T_B] = \frac{1}{\mu - \lambda}. \tag{286}$$

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 λ and μ, 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.

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 6.9

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 6.10

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. □

6.7 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; μ = service rate; λ = 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. (52), 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 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 λ and the other (departure) with parameter μ. 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 6.11

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/∞

The next queueing system we consider is the M/M/∞ 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 λ and each server renders service which is exponentially distributed with parameters μ. 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 λ/μ. 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/∞ queue. Notice that we use the notation A here for the ratio λ/μ while we used the notation ρ for this ratio in the M/M/1 case. Clearly, both represent the offered traffic by definition. Also, both ρ and A represent the mean number of busy servers in the M/M/1 and M/M/∞ 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/∞ queue. According to Little's formula, the mean number of customers in the system is equal to the arrival rate (λ) times the mean time a customer spends in the system which is equal to $1/\mu$ in the case of M/M/∞. Because there are no losses in M/M/∞ so that all the arriving traffic enters the service-system we obtain

$$E[Q] = \lambda(1/\mu) = A. \tag{287}$$

In M/M/1, we must have that ρ cannot exceed unity for stability. In M/M/1 ρ also represents the the server utilization which cannot exceeds unity. However, in M/M/∞, A can take any non-negative value and we often have $A > 1$. M/M/∞ is stable for any $A \geq 0$. Notice that in M/M/∞ 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 $j\mu > \lambda$).

Offered traffic is measured in *erlangs* named after the Danish mathematician A. K. Erlang who was the originator of queueing theory and teletraffic. One erlang represents traffic load of one arrival, on average, per mean service time. This means that, on average, traffic load of one erlang will keep one server busy continuously, or two servers each of which 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/∞, 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 offered traffic = carried traffic occurs because there are no losses in M/M/∞.

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 section.

7.2 Steady-State Equations

Like M/M/1, the M/M/∞ 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 in M/M/1, the arrival rate is independent of changes in the queue-size. However, unlike M/M/1, in M/M/∞, 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 μ and one with parameter λ.

Considering a birth-and-death process that represents the queue evolution of a M/M/∞ queueing system, we obtain the following steady-state equations for the steady-state probabilities π_i (for $i = 0, 1, 2, \ldots$) of having i customers in the system.
$\pi_0\lambda = \pi_1\mu$
$\pi_1\lambda = \pi_2 2\mu$
...
and in general:

$$\pi_n\lambda = \pi_{n+1}(n + 1)\mu, \text{ for } n = 0,\ 1,\ 2,\ \ldots\ . \tag{288}$$

Of course, the sum of the steady-state probabilities must be equal to one, so we again have the additional equation

$$\sum_{j=0}^{\infty} \pi_j = 1. \tag{289}$$

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:

$$\pi_n = \frac{A^n\pi_0}{n!} \text{ for } n = 0,\ 1,\ 2,\ \ldots\ . \tag{290}$$

To obtain π_0, we sum up both sides of Eq. (290), and because the sum of the π_ns equals one, we obtain

$$1 = \sum_{n=0}^{\infty} \frac{A^n\pi_0}{n!}. \tag{291}$$

By the definition of Poisson random variable, see Eq. (26), we obtain

$$1 = \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!}. \tag{292}$$

Thus,

$$e^{\lambda} = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$$

which is also the well-known Maclaurin series expansion of e^{λ}. Therefore, Eq. (291) reduces to

$$1 = \pi_0 e^{A}, \tag{293}$$

or

$$\pi_0 = e^{-A}. \tag{294}$$

Substituting the latter in Eq. (290), we obtain

$$\pi_n = \frac{e^{-A} A^n}{n!} \text{ for } n = 0,\ 1,\ 2,\ \ldots\ . \tag{295}$$

By Eq. (295) we observe that the distribution of the number of busy channels (simultaneous calls or customers) in an M/M/∞ system is Poisson with parameter A.

7.4 Insensitivity

The above results for π_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, 55, 59, 70], all we need is the mean of the distribution and the results are insensitive to higher moments of the distribution. 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 [30, 69, 70].

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 λL 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) [55] - for example, a large population of customers making phone calls. These customers are independent of each other. In M/G/∞, 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 is occupies a server at that point in time. Therefore, the server occupancy events are also a result of 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/∞, in steady state a Poisson distributed number of servers with parameter A which is independent of the shape of the service time distribution.

7.5 Applications

7.5.1 A multi-access model

An interesting application of the M/M/∞ system is the following multi-access problem (see Problem 3.8 in [12]). Consider a stream of packets that their arrival times follow a Poisson process with parameter λ. 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 μ. 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/∞ 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.

7.5.2 Birth rate evaluation

Another application of the M/M/∞ system (or M/G/∞ system) is to the following problem. Consider a city with population 3,000,000, and assume that the birth rate λ is constant. Average life time of people in this city is 78 years. How to compute the birth rate? Using the M/M/∞ model (or actually the M/G/∞ 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.

8 Erlang B Formula

The next queueing system we consider is the M/M/k/k queueing system where the number of servers is k. We assume that the arrival process is Poisson with parameter λ and each server renders service which is exponentially distributed with parameters μ. 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 Offered, Carried and Overflow Traffic

The offered traffic under is the same as under M/M/∞ 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

$$E(Q) = \frac{(1 - P_b)\lambda}{\mu} = (1 - P_b)A. \quad (296)$$

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/∞ 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

$$A - (1 - P_b)A = P_bA.$$

8.2 Solving the Steady-State Equations

The steady-state equations for M/M/k/k are the same as the first k steady-state equations for M/M/∞. Accordingly, we obtain for M/M/k/k:

$$\pi_n = \frac{A^n \pi_0}{n!} \text{ for } n = 0,\ 1,\ 2,\ \ldots, k. \quad (297)$$

To obtain π_0, we again sum up both sides of the latter. This leads to

$$\pi_0 = \frac{1}{\sum_{n=0}^{k} \frac{A^n}{n!}}. \quad (298)$$

Substituting Eq. (298) in Eq. (297), we obtain

$$\pi_n = \frac{\frac{A^n}{n!}}{\sum_{n=0}^{k} \frac{A^n}{n!}} \text{ for } n = 0,\ 1,\ 2,\ \ldots, k. \tag{299}$$

The relationship between (299) and (290) is interesting. Notice that M/M/∞ behaves like M/M/k/k as long as the process does not visit states above k. The transition rates of M/M/∞ and M/M/k/k are exactly the same between states $0, 1, 2, \cdots, k$. Observing (299) 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 (290). Since (290) is merely the Poisson distribution, (299) is the truncated Poisson distribution. Accordingly, to obtain (299), we can simply consider (290), and firstly set $\pi_j = 0$ for all π_j with $j > k$. Then for $0 \leq j \leq k$ we set the π_j for the M/M/k/k values by dividing the π_j values of (290) by the sum $\sum_{j=0}^{k} \pi_j$ of the π_j values in the M/M/∞ model. This is equivalent to considering the M/M/∞ model and deriving the conditional probability of the process being in state j for $j = 0, 1, 2, \cdots, k$, conditioning on the process being within the states $j = 0, 1, 2, \cdots, k$. This conditional probability is exactly the steady-state probabilities π_j of the M/M/k/k model.

The most important quantity out of the values obtained by Eq. (299) is π_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 π_k for an M/M/k/k system loaded by offered traffic A is usually denoted by $E_k(A)$ and is given by:

$$E_k(A) = \frac{\frac{A^k}{k!}}{\sum_{n=0}^{k} \frac{A^n}{n!}}. \tag{300}$$

Eq. (300) is known as Erlang's loss Formula, or Erlang B Formula, published first by A. K. Erlang in 1917 [22].

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 λL. 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. (300) has many applications for telecommunications network design. Given its importance, it is necessary to be able to compute Eq. (300) quickly and exactly for large values of k. This will enable us to answer a dimensioning question such as "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. (300), 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. (300), we obtain

$$\frac{E_m(A)}{E_{m-1}(A)} = \frac{\frac{\frac{A^m}{m!}}{\sum_{j=0}^{m}\frac{A^j}{j!}}}{\frac{\frac{A^{m-1}}{(m-1)!}}{\sum_{j=0}^{m-1}\frac{A^j}{j!}}} = \frac{\frac{\frac{A^m}{m!}}{\sum_{j=0}^{m}\frac{A^j}{j!}}}{\frac{\frac{A^{m-1}}{(m-1)!}}{\sum_{j=0}^{m}\frac{A^j}{j!}-\frac{A^m}{m!}}} = \frac{A}{m}(1-E_m(A)). \tag{301}$$

Isolating $E_m(A)$, this leads to

$$E_m(A) = \frac{AE_{m-1}(A)}{m + AE_{m-1}(A)} \text{ for } m = \ 1,\ 2,\ \ldots, k. \tag{302}$$

Homework 8.1

Complete all the details in the derivation of Eq. (302). □

When $m = 0$, there are no servers (circuits) available, and therefore all customers (calls) are blocked, namely,

$$E_0(A) = 1. \tag{303}$$

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 [37]

$$I_m(A) = \frac{1}{E_m(A)} \tag{304}$$

and the recursion

$$I_m(A) = 1 + \frac{m}{A}I_{m-1}(A) \text{ for } m = \ 1,\ 2,\ \ldots, k. \tag{305}$$

with the initial condition $I_0(A) = 1$.

A useful formula for $I_m(A)$ due to Jagerman [38] is:

$$I_m(A) = A\int_0^{\infty} e^{-Ay}(1+y)^m dy. \tag{306}$$

8.4 The Special Case: M/M/1/1

Homework 8.2

Derive a formula for the blocking probability of M/M/1/1 in four ways: (1) by Erlang B Formula (300), (2) by the recursion (302), (3) by the recursion (305), and (4) by Jagerman Formula (306). □

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

$$\frac{\lambda(1 - E[Q])}{\mu} = E[Q]. \tag{307}$$

Isolating $E[Q]$, the blocking probability is given by

$$\pi_1 = E[Q] = \frac{A}{1 + A}. \tag{308}$$

8.5 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 8.3

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 [73])

By (306) 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,

$$\begin{aligned}\lim_{k\to\infty} \frac{1}{\mathbf{E(A,k)}} &= \lim_{k\to\infty} \textstyle\int_0^\infty e^{-t} \left(1 + \frac{t}{A}\right)^k dt \\ &= \textstyle\int_0^\infty e^{-t} \cdot e^{\frac{t}{A/k}} dt \\ &= \textstyle\int_0^\infty e^{-(1-\frac{1}{A/k})t} dt.\end{aligned}$$

Then observe that

$$\lim_{k\to\infty} \frac{1}{\mathbf{E(A,k)}} = \begin{cases} \infty & \text{if } A/k \leq 1 \\ \frac{1}{1-\frac{1}{A/k}} & \text{if } A/k > 1. \end{cases} \tag{309}$$

And the desired result follows. $\square$

Homework 8.4 (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 8.5

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 λ 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 λ. With negligible variability, M/D/k/k behaves like D/D/k/k and the results follow. □

8.6 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 8.6

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 (296) using Little's formula. Substituting π_k for P_b in (296), 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

$$\hat{U} = \frac{(1 - \pi_k)A}{k}. \tag{310}$$

Homework 8.7

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 increased with the traffic.

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 8.8

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 that 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 $k = A$.

8.7 Critical Loading

A system where the offered traffic load is equal to the system capacity is called *critically loaded* [34]. 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

$$\lim_{k\to\infty} E_k(A) = \frac{\tilde{C}}{\sqrt{k}}. \tag{311}$$

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.

8.8 Insensitivity and Many Classes of Customers

We have discussed in Section 7.4, 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/∞ and M/G/k/k. For M/G/k/k, also the blocking probability is insensitive to the shape of the service time distribution [30, 69, 70].

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 π_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

$$\hat{U} = (1 - \pi_k)A = [1 - A/(1+A)]A = A/(1+A). \tag{312}$$

This comparison is summarized in the following table:

	M/M/1/1	D/D/1/1
π_0	$1/(1+A)$	$1-A$
π_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 (297) 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 arrives at rate of λ_i each of which requires exponentially distributed service with parameter μ_i, for $i = 1, 2$. There are k servers without waiting room (without additional buffer). The aim is to derive the blocking probability.

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

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

the service time of an arbitrary customer has hyperexponential distribution because with probability p it is exponentially distributed with parameter μ_1, and with probability $1 - p$, it is exponentially distributed with parameter μ_2.

The mean service time (holding time) is therefore given by

$$E(S) = \frac{p}{\mu_1} + \frac{1-p}{\mu_2}$$

so $A = \lambda E(S)$, and Erlang B applies.

Homework 8.9 [12]

Extend the results obtained for two classes of customers to the case of n classes of customers.
$\square$

Homework 8.10

This assignment applies to both M/M/∞ and M/M/k/k models.

1. Consider an M/M/∞ 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 [12].

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 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 busy) 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 8.11

Consider an M/M/k/k queue with a given arrival rate λ 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 8.12

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

   $$E[Loss] = 1P(X = k + 1) + 2P(X = k + 2) + 3P(X = k + 3) + \ldots$$

   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

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt. \tag{313}$$

The lower incomplete Gamma function is defined by

$$\gamma(a, x) = \int_0^x t^{a-1}e^{-t}dt. \tag{314}$$

The upper incomplete Gamma function is defined by

$$\Gamma(a, x) = \int_x^\infty t^{a-1}e^{-t}dt. \tag{315}$$

Accordingly,

$$\gamma(a, x) + \Gamma(a, x) = \Gamma(a).$$

For an integer k, we have

$$\Gamma(k) = (k - 1)! \tag{316}$$

$$\Gamma(k,x) = (k-1)!e^{-x}\sum_{m=0}^{k-1}\frac{x^m}{m!}. \tag{317}$$

Therefore,

$$e^{-A}\sum_{m=0}^{k-1}\frac{A^i}{i!} = \frac{\Gamma(k,A)}{\Gamma(k)} \tag{318}$$

so

$$1-e^{-A}\sum_{m=0}^{k-1}\frac{A^i}{i!} = 1-\frac{\Gamma(k,A)}{\Gamma(k)} = \frac{\Gamma(k)-\Gamma(k,A)}{\Gamma(k)} = \frac{\gamma(k,A)}{\Gamma(k)}. \tag{319}$$

Now notice that

$$\begin{aligned}
E[Loss] &= 1P(X=k+1)+2P(X=k+2)+3P(X=k+3)+\ldots \\
&= \sum_{i=k+1}^{\infty}(i-k)A^i\frac{e^{-A}}{i!} \\
&= A\sum_{i=k+1}^{\infty}A^{i-1}\frac{e^{-A}}{(i-1)!} - k\sum_{i=k+1}^{\infty}A^i\frac{e^{-A}}{i!} \\
&= A\sum_{i=k}^{\infty}A^i\frac{e^{-A}}{i!} - k\sum_{i=k+1}^{\infty}A^i\frac{e^{-A}}{i!} \\
&= A\left[1-e^{-A}\sum_{i=0}^{k-1}\frac{A^i}{i!}\right] - k\left[1-e^{-A}\sum_{i=0}^{k}\frac{A^i}{i!}\right].
\end{aligned}$$

□

8.9 A Markov-chain Simulation of M/M/k/k

We 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:
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);
μ = service rate;
λ = 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 8.13

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 using the general simulation principles of Section 4.2. In your comparison consider accuracy (closeness to the analytical results), the length of the confidence intervals and running times. □

Homework 8.14

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 8.15

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 λ and that any packet is of exponential length with parameter μ. 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 λ 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 $10 \leq F \leq 100$. □

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 λ. The service time of each of the k servers is exponentially distributed with parameter μ. 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:

$$\pi_n = \frac{A^n\pi_0}{n!} \text{ for } n = 0,\ 1,\ 2,\ \ldots,\ k-1 \tag{320}$$

and

$$\pi_n = \frac{A^n\pi_0}{k!k^{n-k}} \text{ for } n = k,\ k+1,\ k+2,\ \ldots\ . \tag{321}$$

To obtain π_0, we sum up both sides of Eqs. (320) and (321), and because the sum of the π_ns equals one, we obtain an equation for π_0, which its solution is

$$\pi_0 = \left(\sum_{n=0}^{k-1}\frac{A^n}{n!} + \frac{A^k}{k!}\frac{k}{(k-A)}\right)^{-1}. \tag{322}$$

Substituting the latter in Eqs. (320) and (321), 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:

$$C_k(A) = \sum_{n=k}^{\infty}\pi_n = \frac{A^k}{k!}\frac{k}{(k-A)}\pi_0 = \frac{\frac{A^k}{k!}\frac{k}{(k-A)}}{\sum_{n=0}^{k-1}\frac{A^n}{n!} + \frac{A^k}{k!}\frac{k}{(k-A)}}. \tag{323}$$

Homework 9.1

Derive Eq. (323). □

By Eqs. (300) and (323) we obtain the following relationship:

$$C_k(A) = \frac{kE_k(A)}{k - A[1 - E_k(A)]}. \tag{324}$$

Homework 9.2

1. Derive Eq. (324);
2. Show that $C_k(A) \geq E_k(A)$. □

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 9.3

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.

Using the above notation, we have

$$E[Q] = E[N_Q] + E[N_s] \tag{325}$$

and

$$E[D] = E[W_Q] + E[S]. \tag{326}$$

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 λ 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

$$E[N_s] = \frac{\lambda}{\mu} = A. \tag{327}$$

To obtain $E[N_Q]$, we consider two mutually exclusive and exhaustive events: $\{Q \geq k\}$, and $\{Q < k\}$. Recalling (81), we have

$$E[N_Q] = E[N_Q \mid Q \geq k]P(Q \geq k) + E[N_Q \mid Q < k]P(Q < k). \tag{328}$$

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 λ 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 (328) that

$$E[N_Q] = C_k(A)\frac{A}{k - A}. \tag{329}$$

Homework 9.4

Derive Eq. (329) by a direct approach using $E[N_Q] = \sum_{n=k}^{\infty}(n-k)\pi_n$.

Guide

By (321),

$$E[N_Q] = \sum_{n=k}^{\infty}(n-k)\pi_n = \sum_{n=k}^{\infty}(n-k)\frac{A^n\pi_0}{k!k^{n-k}}$$

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 (329) follows. $\square$

Homework 9.5

Confirm consistence between (329) and (266). □

By (325), (327) and (329), we obtain

$$E[Q] = C_k(A)\frac{A}{k-A} + A. \tag{330}$$

Therefore, by Little's formula

$$E[W_Q] = \frac{C_k(A)\frac{A}{k-A}}{\lambda} = \frac{C_k(A)}{\mu k - \lambda}. \tag{331}$$

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,

$$E[D] = \frac{C_k(A)}{\mu k - \lambda} + \frac{1}{\mu}. \tag{332}$$

Another useful measure is the so-called *delay factor* [17]. It is defined as the ratio of the mean waiting time in the queue to the mean service time. Namely, it is given by

$$D_F = \frac{\frac{C_k(A)}{\mu k - \lambda}}{\frac{1}{\mu}} = \frac{C_k(A)}{k-A}. \tag{333}$$

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 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. (324), 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. (332). A similar procedure can be used to find the minimal k such that delay factor requirement is met.

9.5 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

$$\hat{U} = \frac{E[N_s]}{k} = \frac{A}{k}. \tag{334}$$

Homework 9.6

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 9.7

Consider the M/M/2 queue with arrival rate λ and service rate μ of each server.

1. Show that
$$\pi_0 = \frac{2-A}{2+A}.$$
2. Derive formulae for π_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. (328) – (330), and show that in both ways you obtain
$$E[Q] = \frac{4A}{4-A^2}.$$

□

10 Engset Loss Formula

The Engset loss formula applies to telephony situations where the number of customers is small 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 exponential with mean $1/\hat{\lambda}$. We also assume that there is not 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.

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/∞, 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, 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, in the present case, 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 μ 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:

$$\pi_n(M-n)\hat{\lambda} = \pi_{n+1}(n+1)\mu, \text{ for } n = 0,\ 1,\ 2,\ \ldots,\min\{M,k\}-1. \tag{335}$$

Therefore, after standard algebraic manipulations of (335), we can write π_n, for $n = 0,\ 1,\ 2,\ \ldots,\min\{M,k\}$, in terms of π_0 as follows:

$$\pi_n = \binom{M}{n} \left(\frac{\hat{\lambda}}{\mu}\right)^n \pi_0, \text{ for } n = 0,\ 1,\ 2,\ \ldots, \min\{M, k\}, \tag{336}$$

or, using the notation $\hat{\rho} = \hat{\lambda}/\mu$, we obtain

$$\pi_n = \binom{M}{n} \hat{\rho}^n \pi_0, \text{ for } n = 0,\ 1,\ 2,\ \ldots, \min\{M, k\}. \tag{337}$$

Homework 10.1

Derive Eqs. (336) and (337). □

Of course, the sum of the steady-state probabilities must be equal to one, so we again have the additional normalizing equation

$$\sum_{j=0}^{\min\{M,k\}} \pi_j = 1. \tag{338}$$

By (337) together with the normalizing Eq. (338), we obtain

$$\pi_0 = \frac{1}{\sum_{j=0}^{\min\{M,k\}} \binom{M}{j} \hat{\rho}^j}.$$

Therefore, by (337), we obtain

$$\pi_n = \frac{\binom{M}{n} \hat{\rho}^n}{\sum_{j=0}^{\min\{M,k\}} \binom{M}{j} \hat{\rho}^j}, \text{ for } n = 0,\ 1,\ 2,\ \ldots, \min\{M, k\}. \tag{339}$$

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, π_k does not give the blocking probability. Still, π_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 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

$$P_b = \frac{\hat{\lambda}(M-k)\pi_k}{\hat{\lambda}\sum_{i=0}^{k}(M-i)\pi_i}. \tag{340}$$

Substituting (336) and (337) in (340) and performing few algebraic manipulations, we obtain the Engset loss formula that gives the blocking probability for the case $M > k$ as follows.

$$P_b = \frac{\binom{M-1}{k}\hat{\rho}^k}{\sum_{i=0}^{k}\binom{M-1}{i}\hat{\rho}^i}. \tag{341}$$

Notice that $\hat{\rho}$, defined above by $\hat{\rho} = \hat{\lambda}/\mu$, is the intensity of a **free** customer. An interesting interpretation of (341) 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. (341). $\square$

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:

$$B_i = \frac{\hat{\rho}(M-i)B_{i-1}}{i+\hat{\rho}(M-i)B_{i-1}} \qquad i = 1, 2, 3, \ \ldots, \ k \tag{342}$$

with the initial condition

$$B_0 = 1. \tag{343}$$

Homework 10.3

Derive Eqs. (342) and (343).

Guide

By (341) 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 (342). 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. $\square$

10.4 Insensitivity

In his original work [21], Engset assumed that the idle time as well as the holding time are exponentially distributed. These assumptions have been relaxed in [18] and now it is known that Engset formula applies also to arbitrary idle and holding time distributions (see also [35]).

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

$$T = M\frac{\hat{\lambda}}{\hat{\lambda}+\mu} = \frac{\hat{\rho}M}{(1+\hat{\rho})}. \tag{344}$$

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

$$T_c = \sum_{i=0}^{k} i\pi_i. \tag{345}$$

The offered traffic is obtained by averaging the intensities of the free customers weighted by the corresponding probabilities of their numbers, as follows.

$$T_o = \sum_{i=0}^{k} \hat{\rho}(M-i)\pi_i. \tag{346}$$

To compute the values for T_c and T_o in terms of the blocking probability P_b, we first realize that

$$T_c = T_o(1-P_b), \tag{347}$$

and also,

$$T_o = \sum_{i=0}^{k} \hat{\rho}(M-i)\pi_i = \hat{\rho}M - \hat{\rho}\sum_{i=0}^{k} i\pi_i = \hat{\rho}(M-T_c) \tag{348}$$

and by (347) – (348) we obtain

$$T_c = \frac{\hat{\rho}M(1-P_b)}{[1+\hat{\rho}(1-P_b)]} \tag{349}$$

and

$$T_o = \frac{\hat{\rho}M}{[1+\hat{\rho}(1-P_b)]}. \tag{350}$$

Notice that when $P_b = 0$, we have

$$T_o = T = T_c, \tag{351}$$

and when $P_b > 0$, we obtain by (344), (349) and (350) that

$$T_o > T > T_c. \tag{352}$$

Homework 10.4

Using (344), (349) and (350), show (351) and (352). □

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 μ stay fixed and assume $M > k$. Considering the steady-state equations (335), 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 $\hat{\lambda} \to \infty$, we must have $\pi_1 > 1$ which leads to contradiction; thus, $\pi_0 = 0$, repeating the same argument for the steady-state equations (335) 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 (345),

$$T_c = k.$$

and by (346),

$$T_o = \hat{\rho}(M - k) \to \infty.$$

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 (344), the intended traffic load is

$$T = \frac{\hat{\rho} M}{(1 + \hat{\rho})} \to M.$$

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 $\hat{\lambda} + \mu \to \mu$, the limit of the intended load will take the form

$$\lim T = M \frac{\hat{\lambda}}{\mu} = \hat{\rho} M. \tag{353}$$

Furthermore, under this limiting condition, the terms $\hat{\rho}(M - i)$, $i = 1, 2, 3, \ldots, k$, in (342) 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 (342), equations (302) and (342) 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. (350) we obtain,

$$\hat{\rho} = \frac{T_o}{M - T_o(1 - P_b)}. \tag{354}$$

The latter can be used together with Eq. (341) or (342) 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. (354) to obtain an estimate for $\hat{\rho}$ then the value you obtain for $\hat{\rho}$ is substituted in Eq. (341), or use the recursion (342), to obtain another value for P_b which is then substituted in Eq. (354) 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. (342). Then compute T_o and assuming ρ 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 λ_i fluctuates based on the queue size i. The service rate μ_i also fluctuates based on i. That is, when there are i customers in the system, the service is exponentially distributed with parameter μ_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 μ_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 λ_i and μ_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:

$$0 = -\pi_0\lambda_0 + \pi_1\mu_1 \tag{355}$$

and

$$0 = \pi_{i-1}\lambda_{i-1} - \pi_i(\lambda_i + \mu_i) + \pi_{i+1}\mu_{i+1} \quad \text{for} \quad i = 1, 2, 3, \ldots, k-1. \tag{356}$$

There is an additional last dependent equation

$$0 = \pi_{k-1}\lambda_{k-1} - \pi_k(\mu_k) \tag{357}$$

which is redundant. The normalizing equation

$$\sum_{i=0}^{k} \pi_i = 1 \tag{358}$$

must also be satisfied.

Notice that the equation

$$0 = -\pi_0\lambda_0 + \pi_1\mu_1$$

and the first equation of the set (356), 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 (356), 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:

$$\pi_i = \rho_i\pi_{i-1} \text{ for } i = 1, 2, 3, \ldots, k \tag{359}$$

where

$$\rho_i = \frac{\lambda_{i-1}}{\mu_i} \text{ for } i = 1, 2, 3, \ldots k.$$

Defining also $\rho_0 \equiv 1$, by (359), we obtain

$$\pi_i = \rho_i\rho_{i-1}\rho_{i-2}\ldots\rho_1\pi_0 \text{ for } i = 0, 1, 2, 3, \ldots, k. \tag{360}$$

Homework 11.1

Drive π_i for $i = 0, 1, 2 \ldots k$.

Guide

Summing up equations (359) will give an equation with $1 - \pi_0$ on its left-hand side and a constant times π_0 on its right-hand side. This linear equation for π_0 can be readily solved for π_0. Having π_0, all the other π_i can be obtained by (359). □

Having obtained the π_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 π_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,

$$P_b = \frac{\lambda_k\pi_k}{\sum_{i=0}^{k}\lambda_i\pi_i}. \tag{361}$$

Notice that if the arrival rates do not depend on the state of the system, even if the service rates do, the blocking probability is equal to π_k. To see this simply set $\lambda_i = \lambda$ for all i in Eq. (361) 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 λ_i and μ_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 λ_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. We begin the chapter by considering two extreme SSQ systems with finite buffer. The first is a D/D/1/k 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/k queue. It 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 wrong and misleading. These two extreme cases will be followed by three other cases of queues with finite buffers: the M/M/1/k, the MMPP(2)/M/1/k and the M/$\mathrm{E_n}$/1/k Queues.

12.1 D/D/1/k

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 $\rho = \lambda/\mu \leq 1$, the evolution of the D/D/1/k 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 the buffer is never full so no losses occur. 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 k and $k - 1$. The case $k = 1$ was already considered in our discussion of the M/M/k/k queue, so we assume $k > 1$. 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 μ. We also know that the arrival rate is λ, therefore the loss rate is λ - μ so the blocking probability is given by

$$P_B = \frac{\lambda - \mu}{\lambda}. \tag{362}$$

12.2 SLB/D/1/k

In this case we have an arbitrarily large burst $L_B >> k$ [packets] arrives at time 0, and no further packets ever arrive. For this case the blocking probability is

$$P_B = \frac{L_B - k}{L_B}. \tag{363}$$

and since $L_B >> k$, we have that $P_B \approx 1$. Notice that in this case L_B packets arrive during a period of time T, with $T \rightarrow \infty$, 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/k

As in the M/M/1 case, the M/M/1/k 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/k birth-and-death process, the state-space is finite limited by the buffer size.

The M/M/1/k queue is a special case of the state dependent SSQ considered in the previous section. If we set $\lambda_i = \lambda$ for all $i = 0, 1, 2 \ \ldots k$ and $\mu_i = \mu$ for all $i = 1, 2 \ \ldots k$ in the model of the previous section, that model is reduced to M/M/1/k.

As k is the buffer size, the infinitesimal generator for the M/M/1/k queue-size process is given by

$Q_{i,i+1} = \lambda$ for $i = 0, 1, 2, 3, \ \ldots, \ k-1$.
$Q_{i,i-1} = \mu$ for $i = 1, 2, 3, 4, \ \ldots, \ k$.
$Q_{0,0} = -\lambda$
$Q_{i,i} = -\lambda - \mu$ for $i = 1, 2, 3, \ \ldots, \ k$.

Substituting this infinitesimal generator in Eq. (219) and performing some simple algebraic operations, we obtain the following steady-state equations for the M/M/1/k queue.
$\pi_0 \lambda = \pi_1 \mu$
$\pi_1 \lambda = \pi_2 \mu$
$\ldots$
and in general:

$$\pi_i \lambda = \pi_{i+1} \mu, \ \text{for } i = 0, \ 1, \ 2, \ \ldots, \ k-1. \tag{364}$$

The normalizing equation is:

$$\sum_{j=0}^{k} \pi_j = 1. \tag{365}$$

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:

$$\pi_i = \rho^i \pi_0 \text{ for } i = 0, \ 1, \ 2, \ \ldots k. \tag{366}$$

Summing up both sides of (366), we obtain

$$1 = \sum_{i=0}^{k} \rho^i \pi_0 = \pi_0 \frac{1 - \rho^{k+1}}{1 - \rho}. \tag{367}$$

Therefore,

$$\pi_0 = \frac{1-\rho}{1-\rho^{k+1}}. \tag{368}$$

Substituting the latter in (366), we obtain

$$\pi_i = \rho^i \frac{1-\rho}{1-\rho^{k+1}} \text{ for } i = 0,\ 1,\ 2,\ \ldots k. \tag{369}$$

Of particular interest is the blocking probability π_k given by

$$\pi_k = \rho^k \frac{1-\rho}{1-\rho^{k+1}} = \frac{\rho^k - \rho^{k+1}}{1-\rho^{k+1}} = \frac{\rho^k(1-\rho)}{1-\rho^{k+1}}. \tag{370}$$

Notice that since M/M/1/k has a finite state-space, stability is assured even if $\rho > 1$.

A numerical solution for the M/M/1/k queue steady state probabilities follows. Set an initial value for π_0 denoted $\hat{\pi}_0$ at an arbitrary value. For example, $\hat{\pi}_0 = 1$; then compute the initial value for π_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 π_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 π_n values, we normalize the $\hat{\pi}_n$ values as follows.

$$\pi_n = \frac{\hat{\pi}_n}{\sum_{i=0}^{k} \hat{\pi}_i}. \tag{371}$$

Homework 12.1

Consider an M/M/1/k queue with $k = \rho = 1000$, estimate the blocking probability. **Answer:** 0.999. □

Homework 12.2

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 [50]:

$$R_{TCP} = \frac{1.22}{RTT\sqrt{L_{TCP}}}. \tag{372}$$

Consider a model of TCP over an M/M/1/k. That is, consider many TCP connections with a given RTT all passing through a bottleneck modeled as an M/M/1/k queue. Assuming that packet sizes are exponentially distributed, estimate TCP throughput, using Equations (370) and (372) for a given RTT, mean packet size and service rate of the M/M/1/k queue. Compare your results with those obtained by ns2 simulations [51].

Guide

Use the method of iterative fixed-point solution. See [24] and [28]. □

Homework 12.3

Consider a state dependent Markovian SSQ described as follows.
$\lambda_i = \lambda$ for $i = 0, 1, 2, \ldots, k-1$
$\lambda_k = \alpha\lambda$ where $0 \leq \alpha \leq 1$
$\mu_i = \mu$ for $i = 1, 2, 3 \ldots, k$.
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/k SSQ with arrival rate of λ and service rate of μ. □

12.4 MMPP(2)/M/1/k

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/k Queue is an SSQ with buffer size k characterized by an MMPP(2) arrival process with parameters λ_0, λ_1, δ_0, and δ_1, and exponentially distributed service time with parameter μ. The service times are mutually independent and are independent of the arrival process. Unlike the Poisson arrival process, the interarrival 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 $i = 0, 1, 2 \ldots, k$ be the probability that the arrival process is in mode m and that there are i packets in the system. After we obtain the π_{im} values, the steady-state queue size probabilities can then be obtained by

$$\pi_i = \pi_{i0} + \pi_{i1} \quad \text{for} \quad i = 0, 1, 2, \ldots, k.$$

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 δ_0 and δ_1, as follows:

$$P(m = 0) = \frac{\delta_1}{\delta_0 + \delta_1} \tag{373}$$

$$P(m = 1) = \frac{\delta_0}{\delta_0 + \delta_1}. \tag{374}$$

Because the probability of the arrival process to be in mode m (for $m = 0, 1$) is equal to $\sum_{i=0}^{k} \pi_{im}$, we obtain by (373) and (374)

$$\sum_{i=0}^{k} \pi_{im} = \frac{\delta_{1-m}}{\delta_{1-m} + \delta_m} \quad \text{for} \quad m = 0, 1. \tag{375}$$

The average arrival rate, denoted λ_{av}, is given by

$$\lambda_{av} = P(m = 0)\lambda_0 + P(m = 1)\lambda_1 = \frac{\delta_1}{\delta_0 + \delta_1}\lambda_0 + \frac{\delta_0}{\delta_0 + \delta_1}\lambda_1. \tag{376}$$

Denote

$$\rho = \frac{\lambda_{av}}{\mu}.$$

The MMPP(2)/M/1/k queueing process is a stable, irreducible and aperiodic continuous-time Markov-chain with finite state-space (because the buffer size k 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 interarrival 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 ψ move towards zero will mean infrequent fluctuations of the mode process that may lead to bursty traffic (long stay in each mode) and letting ψ move towards infinity means frequent fluctuations of the mode process. The parameter ψ 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 ψ is one good way to gain insight into performance/burstiness effects.

Therefore, the π_{im} values can be obtain by solving the following finite set of steady-state equations:

$$0 = \mathbf{\Pi Q} \tag{377}$$

where $\mathbf{\Pi} = [\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}, \pi_{20}, \pi_{21}, \ldots, \pi_{k-1,0}, \pi_{k-1,1}, \pi_{k0}, \pi_{k1}]$, and the infinitesimal generator $2k \times 2k$ matrix is $\mathbf{Q} = [Q_{\mathbf{i},\mathbf{j}}]$, where $\mathbf{i}$ and $\mathbf{j}$ are two-dimensional vectors. Its non-zero entries are:

$Q_{00,00} = -\lambda_0 - \delta_0$; $\quad Q_{00,01} = \delta_0$; $\quad Q_{00,10} = \lambda_0$;

$Q_{01,00} = \delta_1$; $\quad Q_{01,01} = -\lambda_1 - \delta_1$; $\quad Q_{01,11} = \lambda_1$;

For $k > i > 0$, the non-zero entries are:

$Q_{i0,i0} = -\lambda_0 - \delta_0 - \mu$; $\quad Q_{i0,i1} = \delta_0$; $\quad Q_{i0,(i+1,0)} = \lambda_0$;

$Q_{i1,i0} = \delta_1$; $\quad Q_{01,01} = -\lambda_1 - \delta_1 - \mu$; $\quad Q_{i1,(i+1,1)} = \lambda_1$;

and

$Q_{k0,(k-1,0)} = \mu$; $\quad Q_{k0,k0} = -\delta_0 - \mu$; $\quad Q_{k0,k1} = \delta_0$;

$Q_{k1,(k-1,1)} = \mu$; $\quad Q_{k1,k1} = -\delta_1 - \mu$; $\quad Q_{k1,k0} = \delta_1$.

In addition we have the normalizing equation

$$\sum_{i=0}^{k} \sum_{m=0}^{1} \pi_{im} = 1. \tag{378}$$

An efficient way, that normally works well for solving this set of equations, is the so called *successive substitution* method (it is also known as Gauss-Seidel, successive approximation or iterations) [20]. It can be described as follows. Consider a set of equation of the form of (377). First, isolate the first element of the vector $\mathbf{\Pi}$, in this case it is the variable π_{00} in the first equation. Next, isolate the second element of the vector $\mathbf{\Pi}$, namely π_{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 $\mathbf{\Pi}$

$$\mathbf{\Pi} = \mathbf{f}(\mathbf{\Pi}). \tag{379}$$

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 (379) 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 $\mathbf{\Pi} = \mathbf{1}$ without any regards to the normalization equation (378). 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 (378). This is the desired solution.

After obtaining the solution for Eq. (377) and (378), one may verify that (375) holds.

To obtain the blocking probability P_b we again notice that $\pi_k = \pi_{k0} + \pi_{k1}$ 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,

$$P_b = \frac{\lambda_0 \pi_{k0} + \lambda_1 \pi_{k1}}{\lambda_{av}}. \tag{380}$$

As an example, we hereby provide the infinitesimal generator for $k = 2$:

	00	**01**	**10**	**11**	**20**	**21**
00	$-\lambda_0-\delta_0$	δ_0	λ_0	0	0	0
01	δ_1	$-\lambda_1-\delta_1$	0	λ_1	0	0
10	μ	0	$-\lambda_0-\delta_0-\mu$	δ_0	λ_0	0
11	0	μ	δ_1	$-\delta_1-\mu$	0	λ_1
20	0	0	μ	0	$-\lambda_0-\delta_0-\mu$	δ_0
21	0	0	0	μ	δ_1	$-\delta_1-\mu$

Homework 12.4

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 (375) 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 12.5

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 π_{im} values and verify that (375) 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 $k = 1$ and discuss the differences. □

Homework 12.6

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 12.7

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 μ 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 ψ to observe the effect of burstiness on the utilization. Confirm your results obtained by successive substitution by simulations using confidence intervals. Demonstrate that as $\psi \to \infty$ the performance (blocking probability and utilization) achieved approaches that of the M/M/1/200 equivalence. Discuss and explain all the results you obtained. □

12.5 M/E$_\text{n}$/1/k

We consider here an M/E$_\text{n}$/1/k SSQ model characterized by a Poisson arrival process with parameter λ, buffer size size of k, and service time that has Erlang distribution with n phases (E$_\text{n}$) with mean $1/(n\mu)$. 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 12.8

Derive and plot the standard deviation to mean ratio as a function of n for an E$_\text{n}$ random variable. □

Let π_0 be the probability that the queue is empty. Also, let π_{im} be the steady-state distribution of having i customers in the system $i = 1, 2, 3, \ldots, k$ and that the customer in service is in phase m, $m = 1, 2, 3, \ldots, n$.

This model is a continuous-time Markov-chain, so the steady-state probabilities π_0 and π_{im} satisfy the following steady-state equations.

$$
\begin{aligned}
0 &= -\lambda\pi_0 + k\mu\pi_{1n} \\
0 &= -(\lambda + k\mu)\pi_{11} + n\mu\pi_{2n} + \lambda\pi_0 \\
0 &= -(\lambda + k\mu)\pi_{1m} + n\mu\pi_{1,m-1} \quad \text{for} \quad m = 2, 3, \ldots, n \\
0 &= -(\lambda + k\mu)\pi_{i1} + n\mu\pi_{i+1,n} + \lambda\pi_{i-1,1} \quad \text{for} \quad i = 2, 3, \ \ldots, k \\
0 &= -(\lambda + k\mu)\pi_{im} + n\mu\pi_{i,m-1} + \lambda\pi_{i-1,m} \quad \text{for} \quad i = 2, 3, \ \ldots, k \quad \text{and} \quad m = 2, 3, \ldots, n.
\end{aligned}
$$

The first equation equates the probability flux of leaving state 0 (to state 11) with the probability flux of entering state 0 only from state $1n$ - where there is only one customer in the system who is in its last service phase. The second equation equates the probability flux of leaving state 11 with the probability flux of entering state 11 (only from states 0 and $2n$). The third equation equates the probability flux of leaving state $1m$ ($m > 1$) with the probability flux of entering state $1m$ ($m > 1$) (only from states 0 and $1, m-1$). The fourth equation equates the probability flux of leaving state $i1$ ($i > 1$) with the probability flux of entering state $i1$ ($i > 1$) (only from states $i+1, n$ and $i-1, 1$). The last equation equates the probability flux of leaving state im ($i > 1$,$m > 1$) with the probability flux of entering state im ($i > 1$,$m > 1$) (only from states $i, m-1$ and $i-1, m$).

The probability of have i in the system, denoted π_i, is obtained by

$$\pi_i = \sum_{m=1}^{n} \pi_{im}.$$

The blocking probability is the probability that the buffer is full namely π_k. The mean queue size is obtained by

$$E[Q] = \sum_{i=1}^{k} i\pi_i.$$

The mean delay is obtained by Little's formula:

$$E[D] = \frac{E[Q]}{\lambda}.$$

Homework 12.9

Consider an M/E$_n$/1/k queue. For a wide range of parameter values (varying λ, μ, n, k) 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. □

12.6 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/$k+n$ queue, and assume that either the arrival rate λ is much higher than the 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 λ packets/s during a very long period of time L, there will be λL 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

$$P_b = \frac{\lambda L - k\mu L}{\lambda L} = \frac{\lambda - k\mu}{\lambda} = \frac{A-k}{A}, \tag{381}$$

where $A = \lambda/\mu$.

Another way to see (381) 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/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 (381) follows.

Homework 12.10

Consider an M/M/$k/k+n$ 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 λ increases the blocking probability approaches the result of (381). Present your results for a wide range of parameter values (varying λ, μ, n, k). Provide interpretation of your results. □

Homework 12.11

Consider again an M/M/1/k queue with $k = \rho = 1000$ and estimate the blocking probability, but this time use the saturated queue approach. **Answer:** 0.999. □

13 Multi-service Loss Models

We have discussed in Section 8.8 a case of a Markovian multi-server loss system (k servers without additional waiting room), involving different classes of customers where customers belong to different classes 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.

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 ([37]) and ([63]) and to earlier publications on the topic [27, 36, 39, 60, 64, 65, 66].

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 λ_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 $B(i)$.

Let j_i be the number of class-i customers in the system for $i = 1, 2, \cdots, I$. Let

$$\overrightarrow{j} = (j_1, j_2, \cdots, j_I)$$

and

$$\overrightarrow{s} = (s_1, s_2, \cdots, s_I).$$

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/∞ model described in Section 7. Accordingly, the M/M/∞ 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 (299) is a truncated version of the distribution of the number of customers in an M/M/∞ 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.

13.1 Infinite number of servers

For the case $k = \infty$, consider the I independent uni-dimensional continuous-time Markov-chains, where $X_i(t), i = 1, 2, \cdots I$, represents the evolution of the number of class-i customers in the system and characterized by the birth-rate λ_i and the death-rate $j_i\mu_i$.

Let $\pi_i(j_i)$ be the steady-state probability of the process $X_i(t), i = 1, 2, \cdots I$ being in state j_i. Then $\pi_i(j_i)$ satisfy the following steady state equations:

$$\lambda_i\pi_i(0) = j_i\mu_i\pi_i(1)$$
$$\lambda_i\pi_i(j_i) = j_i\mu_i\pi_i(j_i + 1)\text{for} \quad j_i = 1, 2, 3, \cdots$$

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/∞ model. Replacing n for j_i, λ for λ_i, and μ for $j_i\mu_i$ in the above equations, we obtain the M/M/∞ 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/∞ model, we obtain:

$$\pi_i(j_i) = \frac{e^{-A_i} A_i^{j_i}}{j_i!} \text{ for } j_i = 0,\ 1,\ 2,\ \ldots\ . \tag{382}$$

Since the processes $X_i(t),\ i = 1, 2, \cdots I$, are independent, the probability $p(\overrightarrow{j}) = p(j_1, j_2, \cdots j_I)$ that in steady-state $X_1(t) = j_1$, $X_2(t) = j_2$, $\cdots$ $X_I(t) = j_I$, is given by

$$p(\overrightarrow{j}) = p(j_1, j_2, \cdots j_I) = \prod_{i=1}^{I} \frac{e^{-A_i} A_i^{j_i}}{j_i!} e^{-A_i}. \tag{383}$$

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 (383), is called a *product-form solution.* The example given by (383) is a trivial case of a product-form solution, but in the following we will see that also when k is finite, we have a product-form solution for the present problem even if the holding times are not exponentially distributed.

13.2 Finite Number of Servers

As the case is with the Erlang Loss System, an important performance measure in the more general multi-service system with finite number of servers is the blocking probability. 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.

In the multi-server case we derive the blocking probability for each class. Namely, 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

$$p(\overrightarrow{j}) = p(j_1, j_2, \cdots j_I) = \frac{1}{C} \prod_{i=1}^{I} \frac{e^{-A_i} A_i^{j_i}}{j_i!} \quad \overrightarrow{j} \in \mathbf{F}. \tag{384}$$

where

$$C = \sum_{\overrightarrow{j} \in \mathbf{F}} \prod_{i=1}^{I} \frac{e^{-A_i} A_i^{j_i}}{j_i!}.$$

Homework 13.1

Derive (384) by truncating (383).

Guide

Consider the steady-state probability distribution of $\overrightarrow{j}$ for the case $k = \infty$ give by (383). 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 (384). □

Let $\mathbf{F}(m)$ be the subset of the states in which an arriving class m customer will not be blocked.

That is

$$\mathbf{F}(m) = \{\overrightarrow{j} \in \mathbf{F} \text{ such that } \sum_{i=1}^{I} s_i j_i \leq k - s_m\}. \tag{385}$$

Then

$$B(m) = 1 - \sum_{\overrightarrow{j} \in \mathbf{F}(m)} p(\overrightarrow{j}),\ m = 1, 2, \cdots, I. \tag{386}$$

Therefore, by (384), we obtain

$$B(m) = 1 - \frac{\sum_{\overrightarrow{j} \in \mathbf{F}(m)} \prod_{i=1}^{I} \frac{A_i^{j_i}}{j_i!}}{\sum_{\overrightarrow{j} \in \mathbf{F}} \prod_{i=1}^{I} \frac{A_i^{j_i}}{j_i!}}. \tag{387}$$

Homework 13.2

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 (384).

Then

$$\mathbf{F}(1) = \{(0,0),\ (0,1),\ (1,0),\ (2,0)\}.$$

and

$$\mathbf{F}(2) = \{(0,0),\ (1,0)\}.$$

Use (387) to obtain the blocking probability. $\square$

Recursion based on the number of busy servers

Observing (387), we realize that it is not scalable as it requires a prohibitively large amount of computations as the number of servers and the number of service classes increase. As in the case of the Erlang loss formula, we will now present a recursive algorithm [27, 37, 39, 60, 63] that is far more scalable than (387).

Define

$$\mathbf{F}_\kappa = \{\overrightarrow{j} \in \mathbf{F} \text{ such that } \sum_{i=1}^{I} s_i j_i = \kappa\} \;\; \kappa = 0, 1, 2, \cdots, k. \tag{388}$$

The set $\mathbf{F}_\kappa$ represents all the states where exactly κ servers are busy.

Then the probability that κ servers are busy is given by

$$P_\kappa = \{\overrightarrow{j} \in \mathbf{F} \text{ such that } \sum_{i=1}^{I} s_i j_i = \kappa\}, \;\; \kappa = 0, 1, 2, \cdots, k. \tag{389}$$

If we have a way to compute the P_κ values in a scalable way then we can obtain the blocking probabilities $B(m)$ for $m = 1, 2, 3, \cdots, I.$ by

$$B(m) = \sum_{\kappa=k-s_\kappa+1}^{k} P_\kappa \tag{390}$$

and also the utilization by

$$U = \frac{\sum_{\kappa=0}^{k} \kappa P_\kappa}{k}. \tag{391}$$

Define $P_\kappa = 0$, for $\kappa < 0$.

Let Π_κ be a function κ that by normalizing it we can obtain P_κ.

The following recursive algorithm computes the blocking probability for each class and the utilization in a scalable way.

Step 1: Initialize $\Pi_0 = 1$ and $\Pi_\kappa = 0, \;\kappa < 0$.

Step 2: For $\kappa = 1, 2, \cdots, k$, do

$$\Pi_\kappa = \frac{1}{\kappa} \sum_{i=1}^{I} s_i \rho_i \Pi_{\kappa - s_i}.$$

Step 3: Compute the following normalization constant

$$\hat{C} = \sum_{\kappa=0}^{k} \Pi_\kappa.$$

Step 4: For $\kappa = 0, 1, 2, \cdots, k$, do

$$P_\kappa = \frac{\Pi_\kappa}{\hat{C}}.$$

Step 5: Compute the blocking probability for each class and the overall system utilization by (390) and (391).

After initialization of the un-normalized Π_κ values in Step 1, in Step 2, we recursively compute Π_κ values , where for each κ value we balance the downwards rates from the state "there are κ servers busy" to upwards rates from the state "there are $\kappa - s_i$ servers busy" for each i. Notice that the arrival rates are normalized by the service rates so we use ρ_i instead of λ_i and 1 instead of μ_i. Steps 3 and 4 are then used to normalize the Π_κ values to obtain the P_κ values. Finally, Step 5 is used to obtain the blocking probability for each service class and the utilization from the normalized P_κ values.

Insensitivity

As discussed, the insensitivity property of the M/M/k/k system extends to the multi-service system. This makes the analyzes and results of muti-service systems very relevant for real life telecommunications systems and networks. In practice, a transmission trunk or lightpath [76] 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 robustness of the model and to its applicability to practical scenarios.

Critical Loading

As discussed in Section 8.7, 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

$$\sum_{i=1}^{I} A_i = k. \tag{392}$$

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:

$$\lim_{k\to\infty} B(i) = \frac{s_i C_{MS}}{\sqrt{k}}, \quad i = 1, 2, \ \cdots, \ I \tag{393}$$

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 (311) 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 than 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 13.3

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 (384) and using Eq. (386), and (3) by using the recursive algorithm. □

Homework 13.4

Provide examples where customers that belong to different class experience the same blocking probability. Verify the equal blocking probability using (384), 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 (386) and simulations. □

Homework 13.5

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 13.6

Study and program the convolution algorithm described in [36, 37, 63]. Also write a program for the recursion algorithm and for the method based on (384). 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. □

14 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 Θ composed of all the nonnegative integers, and a Transition Probability Matrix $\mathbf{P} = [P_{ij}]$ given by

$$P_{i,i-1} = a_0 \quad \text{for} \quad i \geq 1 \tag{394}$$

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

$$P_{i,i+k} = a_{k+1} \quad \text{for} \quad i \geq 0, k \geq 1. \tag{395}$$

Defining the steady-state probability vector by $\mathbf{\Pi} = [\pi_{\mathbf{0}}, \pi_{\mathbf{1}}, \pi_{\mathbf{2}}, \ldots]$, 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 π_is, 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 (394) and (395) in the latter, we obtain

$$\pi_0 = \pi_0[a_0 + a_1] + \pi_1 a_0 \quad (396)$$

$$\pi_1 = \pi_0 a_2 + \pi_1 a_1 + \pi_2 a_0 \quad (397)$$

$$\pi_2 = \pi_0 a_3 + \pi_1 a_2 + \pi_2 a_1 + \pi_3 a_0 \quad (398)$$

and in general

$$\pi_n = \sum_{i=0}^{n+1} \pi_i a_{n+1-i} \quad \text{for} \quad n \geq 1. \quad (399)$$

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 nth equation of the set (396) – (399) by z^n, and summing up, we obtain after some algebraic operations

$$\Pi(z) = \pi_0 a_0 - \pi_0 z^{-1} a_0 + z^{-1} A(z) \Pi(z) \quad (400)$$

which leads to

$$\Pi(z) = \frac{\pi_0 a_0 (1 - z^{-1})}{1 - z^{-1} A(z)}. \quad (401)$$

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,

$$\pi_0 = \frac{1 - A'(1)}{a_0}. \quad (402)$$

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. (402)? The answer is that nothing wrong with it. What we call π_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. (402) follows.

Homework 14.1

Provide in detail all the algebraic operations and the application of L'Hopital rule to derive equations (400), (401) and (402).

Guide

Multiplying the nth equation of the set (396) – (399) by z^n and summing up, we obtain an equation for $\Pi(z)$ by focussing first on terms involving π_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 (401) 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 (401) we obtain

$$\begin{aligned}
\lim_{x\to 1} \Pi(z) &= \lim_{x\to 1} \frac{\pi_0 a_0 (1 - z^{-1})}{1 - z^{-1}A(z)} \\
&= \lim_{x\to 1} \frac{\pi_0 a_0 z^{-2}}{z^{-2}A(z) - z^{-1}A'(z)}.
\end{aligned}$$

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 (402) follows. □

Homework 14.2

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. □

15 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 σ_s. The arrival process is assumed to be Poisson with rate λ 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.

15.1 Pollaczek Khinchin Formula: Residual Service Approach [12]

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

$$E[W_Q] = E[R] + \frac{E[N_Q]}{\mu} \tag{403}$$

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 ρ, 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 15.1

Verify that Eq. (403) holds for M/M/1. □

Observe that while Equation (403) 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 (403), we obtain,

$$E[W_Q] = \frac{E[R]}{1 - \rho}. \tag{404}$$

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

$$E[R] = \frac{1}{T} \int_0^T R(t)dt. \tag{405}$$

Following [12], let $S(T)$ be the number of service completions by time T and S_i the ith service time. Notice that the function $R(t)$ takes the value zero during times that there is no customer in service and jumps to the value of S_i at the point of time the ith 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 ith triangle is S_i. Therefore, for large T, we can ignore the last possibly incomplete triangle, so we obtain

$$E[R] = \frac{1}{T} \sum_{i=1}^{S(T)} \frac{1}{2} S_i^2 = \frac{1}{2} \frac{S(T)}{T} \frac{1}{S(T)} \sum_{i=1}^{S(T)} S_i^2. \tag{406}$$

Letting T approach infinity, the latter gives

$$E[R] = \frac{1}{2} \lambda \overline{S^2} \tag{407}$$

where $\overline{S^2}$ is the second moment of the service time.

By (404) and (407), we obtain

$$E[W_Q] = \frac{\lambda\overline{S^2}}{2(1-\rho)}. \quad (408)$$

Thus, considering (239), we obtain

$$E[D] = \frac{\lambda\overline{S^2}}{2(1-\rho)} + 1/\mu. \quad (409)$$

Using Little's formula and recalling that $\sigma_s^2 = \overline{S^2} - (1/\mu)^2$, Eq. (409) leads to the well know Pollaczek Khinchin Formula for the mean number of customers in an M/G/1 system:

$$E[Q] = \rho + \frac{\rho^2 + \lambda^2\sigma_s^2}{2(1-\rho)}. \quad (410)$$

15.2 Pollaczek Khinchin Formula: by Kendall's Recursion [42]

Let us now derive (410) in a different way. Letting q_i be the number of customers in the system immediately following the departure of the ith customer, the following recursive relation, is obtained.

$$q_{i+1} = q_i + a_{i+1} - I(q_i) \quad (411)$$

where a_i is the number of arrivals during the service time of the ith 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 [42], 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 [47]. Along with Little's and Erlang B formulae, and the Pollaczek-Khinchin equation, the Kendall's and Lindley's recursions are key foundations of queueing theory.

To understand the recursion (411), 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 ith 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 (411) and taking expectations, we obtain

$$E[q_{i+1}^2] = E[q_i^2] + E[I(q_i)^2] + E[a_{i+1}^2] - 2E[q_iI(q_i)] + 2E[q_ia_{i+1}] - 2E[I(q_i)a_{i+1}] \quad (412)$$

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 $q_iI(q_i) = q_i$. Also notice that because of the independence between a_{i+1} and q_i, and because (by (81)) the mean number of arrivals during service time in M/G/1 is equal to ρ, we obtain in steady-state that $E[I(q_i)a_{i+1}] = \rho^2$ and $E[q_ia_{i+1}] = E[q_i]\rho$. Therefore, considering (412), and setting the steady-state notation $E[a] = E[a_i]$ and $E[Q] = E[q_i]$, we obtain after some algebra

$$E[Q] = \frac{\rho + E[a^2] - 2\rho^2}{2(1-\rho)}. \tag{413}$$

To obtain $E[a^2]$, we notice that by EVVE,

$$var[a] = E[var[a \mid S]] + var[E[a \mid S]] = \lambda E[S] + \lambda^2\sigma_s^2 = \rho + \lambda^2\sigma_s^2 \tag{414}$$

recalling that S is the service time and that σ_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. (414) that

$$E[a^2] = var[a] + \rho^2 = \rho + \lambda^2\sigma_s^2 + \rho^2.$$

Therefore,

$$E[Q] = \frac{2\rho + \lambda^2\sigma_s^2 - \rho^2}{2(1-\rho)} \tag{415}$$

or

$$E[Q] = \rho + \frac{\rho^2 + \lambda^2\sigma_s^2}{2(1-\rho)} \tag{416}$$

which is identical to (410) - the Pollaczek-Khinchin Formula.

Homework 15.2

Re-derive the Pollaczek-Khinchin Formula in the two ways presented above with attention to all the details (some of which are skipped in the above derivations). □

15.3 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 (410), and after some algebra, we obtain

$$E[Q] = \frac{\rho}{1-\rho} \tag{417}$$

which is consistent with (266).

Another interesting case is the M/D/1 queue where $\sigma_s^2 = 0$. Substituting the latter in (410), we obtain after some algebra

$$E[Q] = \frac{\rho}{1-\rho} \times \frac{2-\rho}{2}. \tag{418}$$

Because the second factor of (418), 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.

15.4 Busy Period

We have defined and discussed the concept of busy period in Section 6.6 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:

$$E[T_B] = \frac{1}{\mu - \lambda}. \tag{419}$$

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 15.3

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 (419) also applies to an M/G/1 queue. □

Homework 15.4

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 ζ later. As in M/G/1, customers arrive in accordance with a Poisson process with rate λ and the mean service time is $1/\mu$. Service times are mutually independent and independent of the interarrival 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

$$E[T_I] = \frac{1}{\lambda} + \frac{1}{\zeta}.$$

□

15.5 M/G/1 with Priorities

We have already considered a non-FIFO service policy. We mentioned the LIFO policy in our discussion of the M/M/1 queue. We will now discuss non FIFO service disciplines in the context of the M/G/1 [12].

Let us consider an M/G/1 queueing system with m priority classes. Let λ_j and μ_j be the arrival and service rate of customers belonging to the jth priority class for $j = 1, 2, 3, \ldots, m$. The mean service time of customers belonging to the jth priority class is therefore equal to $1/\mu_j$. The second moment of the service time of customers belonging to the jth 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 λ_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 interarrival 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*.

15.6 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 (407), we obtain:

$$E[R] = \frac{1}{2} \sum_{j=1}^{m} \lambda_j \overline{S^2(j)}. \tag{420}$$

Homework 15.5

Derive Eq. (420). □

As in Eq. (403), we have for the highest priority,

$$E[W_Q(1)] = E[R] + \frac{E[N_Q(1)]}{\mu_1} \tag{421}$$

and similar to (403) we obtain

$$E[W_Q(1)] = \frac{E[R]}{1-\rho_1}. \tag{422}$$

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

$$E[W_Q(2)] = E[R] + \frac{E[N_Q(1)]}{\mu_1} + \frac{E[N_Q(2)]}{\mu_2} + E[W_Q(2)]\rho_1. \tag{423}$$

By the latter and Little's formula for Class 2 customers, namely,

$$E[N_Q(2)] = \lambda_2 E[W_Q(2)],$$

we obtain

$$E[W_Q(2)] = \frac{E[R] + \rho_1 E[W_Q(1)]}{1 - \rho_1 - \rho_2}. \tag{424}$$

By Eqs. (424) and (422), we obtain

$$E[W_Q(2)] = \frac{E[R]}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}. \tag{425}$$

Homework 15.6

Show that for $m = 3$,

$$E[W_Q(3)] = \frac{E[R]}{(1 - \rho_1 - \rho_2)(1 - \rho_1 - \rho_2 - \rho_3)}. \tag{426}$$

and that in general

$$E[W_Q(j)] = \frac{E[R]}{(1 - \sum_{i=1}^{j-1} \rho_i)(1 - \sum_{i=1}^{j} \rho_i)}. \quad \square \tag{427}$$

The mean delay for a jth priority class customer, denoted $E(D(j))$, is given by

$$E[D(j)] = E[W_Q(j)] + \frac{1}{\mu_j} \text{ for } j = 1, 2, 3, \ldots, m. \tag{428}$$

Homework 15.7

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? □

15.7 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

$$E[D(j)] = \frac{1}{\mu_j} + \frac{R(j)}{1 - \sum_{i=1}^{j} \rho_i} + E[D(j)] \sum_{i=1}^{j-1} \rho_i \tag{429}$$

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. (429) is simply the mean service time of a jth priority customer. The second term in 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. (404) that gives the mean time of waiting in the queue in an M/G/1 queueing system where we replace ρ of (404) 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 jth 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. (429) 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 ρ_is 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. (429) leads to

$$E[D(1)] = \frac{(1/\mu_1)(1-\rho_1)+R(1)}{1-\rho_1}, \tag{430}$$

and

$$E[D(j)] = \frac{(1/\mu_j)(1-\sum_{i=1}^{j}\rho_i)+R(j)}{(1-\sum_{i=1}^{j-1}\rho_i)(1-\sum_{i=1}^{j}\rho_i)}. \tag{431}$$

Homework 15.8

Derive Eqs. (430) and (431). □

16 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 interarrival times are IID. Consider for example the blocking probability of an M/M/1/k queue as a function of $\rho = \lambda/\mu$, then the blocking probability will gradually increase with ρ and approaches one as $\rho \to \infty$. However, we recall our discussion of the SLB/D/1/k 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

$$E[Q] = \lambda E[D] \tag{432}$$

is applicable to G/G/1.

16.1 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, 19, 58].

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 interarrival 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

$$W_q(t) = \sup_{0 \leq s < t} \{W_a(t) - W_a(s) - t + s\}. \tag{433}$$

If the queue is not empty at time t, the s value that maximizes the right-hand side of (433) 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 16.1

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. □

16.2 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 $P(Q > n)$. Note that for a G/D/1 queue we have [61]

$$Q_t = \lceil W_q(t) \rceil, \tag{434}$$

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 [61]

$$P(Q_t > n) = P(W_q(t) > n), \text{ for } n = 0, 1, 2, \ldots . \tag{435}$$

16.3 Wong's Inequality

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 interarrival 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.

This gives rise to the following problem. 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 [74] 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.

$$\rho P_{loss} \leq P(Q > k) \tag{436}$$

Roberts et al. [61] 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 [74]. Let λ be the arrival rate and μ 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, as 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 an overflow period of the 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 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 (436) follows. $\square$

Homework 16.2

Show that (436) applies to an M/M/1 queue and its M/M/1/k 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/k,

$$P_{loss} = \frac{\rho^k(1-\rho)}{1-\rho^{k+1}},$$

and for M/M/1,

$$P(Q > k) = \rho^{k+1}(1-\rho) + \rho^{k+2}(1-\rho) + \rho^{k+3}(1-\rho) + \ldots = \rho^{k+1}. \quad \square$$

Homework 16.3

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 ρ^{-1} and observe and discuss the relationship between the two. □

Homework 16.4

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 16.5

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. □

17 Queueing Networks

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 computers 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.

17.1 Jackson Networks

Consider a network made of Markovian queues (M/M/1, M/M/k and M/M/∞). An issue that is very important for such Markovian queueing networks is what is the output of such queues because in queueing networks output of one queue may be the input of another. Burke's Theorem answers this question. Burke's Theorem states that, in steady-state, the output (departure) process of M/M/1, M/M/k or M/M/∞ queue is Poisson. 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/∞ queue with arrival rate of λ will have a Poisson departure process with rate λ 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 all the traffic arrives into the first queue following a Poisson process with parameter λ. The service times required by each of the arriving customers at the two queues are independent and exponentially distributed with parameter μ. 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 λ, 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 [16] 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 network of two single-server queues called Q1 and Q2. The service times in Q1 and Q2 are exponentially distributed with parameters $\mu_1 = 1$ and $\mu_2 = 1$, respectively. Assume that the arrival rate from the outside into Q1 follows Poisson process with rate $r = 10^{-8}$. Further assume that all the packets that leave Q1 enter Q2 and that every packet that leaves Q2 leaves the network with probability 10^{-3} and return to Q1 with probability $1 - 10^{-3}$. The total arrival process into Q1 will include the original Poisson stream with rate $r = 10^{-8}$ plus all the feedback from Q2. This results in a process based on very infrequent original arrivals each of which brings with it a burst of mean of some thousand feedback arrivals from Q2. Clearly this is not a Poisson process. Furthermore, the inter-arrivals of packets within a burst, most of which are feedback from Q2, are very much dependent on the service times of Q1 and Q2, so clearly we have dependence between inter-arrival times and service times.

Nevertheless, the so-called Jackson's Theorem extends the above result to networks that are not acyclic. In other words, although the queues are not M/M/1 (or M/M/k or M/M/∞), they behave in terms of their queue-size and waiting time statistics as if they are.

Consider a network of N single-server queues in steady-state. The result can easily be extended to multi-server queues such as M/M/k and M/M/∞, but let us consider single-server queues for now. For queue i, $i = 1, 2, 3, \ldots, N$, the arrival process is Poisson with rate A_i. We allow for $A_i = 0$ for some queues, but there must be at least one queue j, such that $A_j > 0$. Once a customer completes its service in queue i, it continues to queue j with probability P_{ij}, of 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 λ_j be the total arrival rate into queue j. These arrival rates can be computed by solving the following set of equations.

$$\lambda_j = A_j + \sum_{i=1}^{N} \lambda_i P_{ij}, \qquad j = 1, 2, 3, \ldots, N. \tag{437}$$

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 jth queue are assumed mutually independent and independent of the arrival process at each queue. The service times at the jth queue are assumed exponentially distributed with parameter μ_j and mutually independent and are also assumed independent of the arrival process at that queue. Let ρ_j be defined by

$$\rho_j = \frac{\lambda_j}{\mu_j} \qquad \text{for } j = 1, 2, 3, \ldots, N. \tag{438}$$

Let Q_j be the queue-size of queue j. Then according to Jackson's Theorem, in steady-state, we have that

$$P(Q_1 = k_1, Q_2 = k_2, \ldots, Q_N = k_N) = P(k_1)P(k_2)P(k_3) \cdot \ \ldots \ \cdot P(k_N) \tag{439}$$

where $P(k_i) = \rho_i^{k_i}(1 - \rho_i)$, $\qquad$ for $i = 1, 2, 3, \ldots, N$. In other words, the queues behave as M/M/1 queues despite the fact that the network may be cyclic (not acyclic) in which case the queues are not M/M/1 queues. Therefore, the mean queue-size of the jth queue is given by

$$E[Q_j] = \frac{\rho_j}{1-\rho_j}. \tag{440}$$

The mean delay of a customer in the jth queue $E[D_j]$ can be obtain by Little's formula as follows.

$$E[D_j] = \frac{E[Q_j]}{\lambda_j}. \tag{441}$$

Using Little's formula, by considering the entire queueing network as our system, we can also derive the mean delay of an arbitrary customer $E[D]$:

$$E[D] = \frac{\sum_{j=1}^{N} E[Q_j]}{\sum_{j=1}^{N} A_j}. \tag{442}$$

Let us now consider the above-mentioned two-queue network example. Using our notation, we have $A_1 = 10^{-8}$ and $A_2 = 0$; $\mu_1 = \mu_2 = 1$; further,

$$\lambda_1 = A_1 + (1 - 10^{-3})\lambda_2$$

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/∞ 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 measures such as blocking probability or carried traffic. One useful approximation is the so-called Reduced-Load Erlang Fixed-Point Approximation which is reasonably accurate and very useful for loss networks.

Homework 17.1

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

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. □

17.2 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, 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.

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 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. Now that we have means to obtain the blocking probability on each link, we will explain how they can be used to compute 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. Let $B(j)$ be the blocking probability on link j obtained by Erlang formula where the traffic load on that link may include traffic from many other traffic routes not only the one we consider. Recalling that multiplexing of Poisson processes give another Poisson process which its rate is the sum of the individual rates, we can compute the total traffic in Erlang offered to any link.

In particular, using the notation of Section 17.1, in a similar way to (437), the λ_j values (designating the total arrival rate into queue j) can be computed by iteratively solving the following set of equations.

$$\lambda_j = A_j + \sum_{i=1}^{N} \lambda_i (1 - B(i)) P_{ij}, \qquad j = 1, 2, 3, \ \ldots, N. \tag{443}$$

To solve these equations, we start with an initial vector of λ_j values, computing the corresponding values of $B(j)$ using for example the Erlang B formula and a new set of λ_j values by (443). We continue iteratively until two consecutive sets of λ_j values are close enough. This results also in a set of $B(j)$ values. The P_{ij} values are initially set based on the traffic in routes assuming no losses. After solving (443) by the fixed point iterations just described, a new set of the P_{ij} values can be obtained considering the resulting blocking probability values ($B(j)$). The process will repeat until two consecutive sets of $B(j)$ values are close enough, and the fixed point solution is achieved for a given set of routes.

Then, by the independence assumption, we obtain an evaluation of the blocking probability by

$$B_R = 1 - \prod_{j \in R} (1 - B(j)). \tag{444}$$

The above solution based on the principles of the Reduced-Load and Erlang Fixed-Point Approximations can be applied to modelling a cellular mobile networks where each cell is equivalent to one M/M/k/k (or M/G/k/k) system, so the cellular mobile network is modelled by a network of M/M/k/k queues. Generation of new calls in a cell is equivalent to arrivals into an M/M/k/k queue and handover between cells is equivalent to traffic that completes service in one M/M/k/k system and moves to another. Another application is an Optical Burst Switching (OBS) network [62] where bursts of data are moving between OBS nodes each of which is modelled as an M/M/k/k system.

Another important applications of the Erlang Fixed-Point Approximation (also called Reduced-Load approximation) are circuit switched networks where a further step is required. For circuit switching networks having the blocking probability B_R, for each route R, implies that the traffic on the entire route R will be reduced by a factor of $(1 - B_R)$. Updating all the traffic loads on all the routes will give a new set of values for $B(j)$, $j \in R$. We update these values in (444) to obtain a new value for B_R. We repeat this process until the updated value of B_R is arbitrarily close to its previous value.

17.3 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;
$\mu(i)$ = service rate in cell i;
$\lambda(i)$ = arrival rate of new calls in cell i;
$P(i,j)$ = Matrix of routing probabilities;
$\delta(i)$ = handover rate in cell i
P_B = Blocking probability.

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

$$R(01) \leq \frac{\sum_{i=1}^{N} \lambda(i)}{\sum_{i=1}^{N} \lambda(i) + \sum_{i=1}^{N} Q(i)\mu(i) + \sum_{i=1}^{N} Q(i)\delta(i)}$$

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

$$R(01) \leq \frac{\sum_{i=1}^{N} \lambda(i) + \sum_{i=1}^{N} Q(i)\mu(i)}{\sum_{i=1}^{N} \lambda(i) + \sum_{i=1}^{N} Q(i)\mu(i) + \sum_{i=1}^{N} \delta(i)}$$

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 $\sum_{j=1}^{N} S_a(j) = MAXN_a$, the simulation ends, so we compute the blocking probabilities as follows.

$$P_B = \frac{\sum_{i=1}^{N} N_b(i)}{\sum_{i=1}^{N} \lambda(i)}.$$

Else, $S_a(i) = S_a(i) + 1$ 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

$$R(01) \leq \frac{\sum_{j=1}^{i} Q(j)\mu(j)}{\sum_{j=1}^{N} Q(j)\mu(j)}.$$

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 out of which of the N cells it handovered. To find out we use the following loop.

For $i = 1$ to N, do: If

$$R(01) \leq \frac{\sum_{j=1}^{i} Q(j)\delta(j)}{\sum_{j=1}^{N} Q(j)\delta(j)}.$$

Then stop the loop. The handover occurs out of cell i, so $Q(j) = Q(j) - 1$. Note that again we do not not need to verify that $Q(j) > 0$.

18 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.

18.1 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 the AVR discussed in Section 2.1. The AVR is related to the so-called Index of Dispersion for Counts (IDC) [31] 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 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

$$\lim_{n\to\infty} \frac{var[S_n]}{\alpha n^{2H}} = 1. \tag{445}$$

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.

18.2 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 λ, and λt 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 very useful in 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 λ where λ 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.

18.3 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: λ_0, λ_1, δ_0, and δ_1.

Queueing models involving MMPP input have been analyzed in the 70s and 80s using Z-transform [75, 77, 78, 79]. Neuts developed matrix methods to analyse such queues [54]. 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 [25, 31, 46, 53].

18.4 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 nth interval.

According to the Gaussian Autoregressive model we assume that $X_n,\ n = 1, 2, = 3\ \ldots$ is the so-called kth order autoregressive process, defined by

$$X_n = a_1 X_{n-1} + a_2 X_{n-2} + \ \ldots \ + a_k X_{n-k} b \tilde{G}_n, \tag{446}$$

where $\tilde{G}_n$ is a sequence of IID Gaussian random variables each with mean η 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 η. 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

$$X_n = a X_{n-1} + b \tilde{G}_n, \tag{447}$$

where $\tilde{G}_n$ is again a sequence of IID Gaussian random variables with mean η 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 [49] as a model of a VBR traffic stream generated by a single source of video telephony.

The X_ns 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 η with measurable (estimated) parameters of the process λ, σ^2 and S, are provided based on [4]:

$$a = \frac{S}{S + \sigma^2} \tag{448}$$

$$b = \sigma^2 (1 - a^2) \tag{449}$$

$$\eta = \frac{(1 - a)\lambda}{b} \tag{450}$$

where S is the autocovariance sum given by Eq. (160).

18.5 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)) [29], defined as follows:

$$D_n = aD_{n-1} + I_n E_n, \tag{451}$$

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 λ.

The EAR(1) has many nice and useful properties. The $\{D_n\}$ process is a sequence of exponential random variables with parameter λ. 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

$$C_{EAR1}(k) = a^k. \tag{452}$$

It is very easy to simulate the $\{D_n\}$ process, so it is useful in demonstrating by simulation the relationship between correlation in the arrival process and queueing performance.

Homework 18.1

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 (81), $\mathcal{L}_{I_1E_1}(s) = P(I = 1)E[e^{-SE_1}] + P(I = 0)E[e^{-0}] = (1-a)\lambda/(\lambda+s) + a$. 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 λ, 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 $n > 1$. □

18.6 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 λ. 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/∞ queueing system. Having m sources simultaneously active is equivalent to having m servers busy in an M/G/∞ system. M/G/∞ 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/∞ 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 nth 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 λ is per time interval. That is, the total number of transmissions to start in one time interval is λ.

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,

$$E[X_n] = \lambda r/(\gamma - 1). \tag{453}$$

Also, another important relationship for this model, which is provided here without proof, is

$$\gamma = 3 - 2H, \tag{454}$$

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/∞ model.

Homework 18.2

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 16 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 16, 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 18.3

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 ρ on mean delay. Use the ranges $0 \le a \le 1$ and $0 \le \rho \le 1$. □

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 and R. Gallager, *Data networks*, Prentice Hall, Englewood Cliff, New Jersey 1992.

[13] D. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, Belmont, Massachusetts 2002.

[14] S. K. Bose, *An Introduction to queueing systems*, Kluver Academic/Plenum Publisher, New York 2002.

[15] G. E. P. Box and G. M. Jenkins, *Time series analysis forecasting and control*, Holden Day San Francisco, 1970.

[16] P. J. Burke, "The output of a queuing system," *Operations Research*, vol. 4, 1956), pp. 699–704.

[17] Cisco Inc., "Traffic Analysis", [online] available: http://www.cisco.com/en/US/docs/ios/solutions_docs/voip_solutions/TA_ISD.html #wp1038467

[18] J. W. Cohen, "The Generalized Engset Formulae", Philips Telecommunication Review, vol. 18, 1957, pp. 158-170.

[19] J. W. Cohen, *The Single Server Queue*, North-Holland Publ. Cy., rev. ed., Amsterdam, 1982.

[20] R. B. Cooper, *Introduction to Queueing Theory*, North Holland 1981.

[21] T. Engset, "Die wahrscheinlichkeitsrechnung zur bestimmung der wahleranzahl in automatischen fernsprechamtern." *Elektrotechnische zeitschrift*, vol. 39, no. 31, Aug. 1918, pp. 304–306.

[22] 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.

[23] W. Feller, *An introduction to probability theory and its applications*, Wiley, New York, 1968.

[24] 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.

[25] W. Fischer and K. Meier-Hellstern, "The Markov Modulated Poisson process (MMPP) cookbook," *Performance Evaluation*, vol. 18, pp. 149–171, 1992.

[26] G. S. Fishman, "Discrete-event simulation," Springer-Verlag, London 2001.

[27] 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.

[28] 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.

[29] 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.

[30] D. Gross, J. F. Shortle, J. M. Thompson and C. M. Harris, *Fundamentals of queueing theory*, John Wiley & Sons Inc., 2008.

[31] 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.

[32] P. G. Hoel, S. C. Port and C. J. Stone, "Introduction to Stochastic Processes," Houghton Mifflin, 1972.

[33] F. Hübner, "Analysis of a finite-capacity asynchronous multiplexer with deterministic traffic sources", *proceedings of 7th ITC Specialist Seminar*, Moristoen, October 1990.

[34] P. J. Hunt and F. P. Kelley, "On critically loaded loss networks," *Advances in Applied Probability*, vol. 21, pp. 831-841, 1989.

[35] J. Hui, *Switching and traffic theory for integrated broadband networks*, Kluwer 1990.

[36] 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.

[37] V. B. Iversen, *Teletraffic Engineering Handbook*, Technical University of Denmark, 2009, [Online]. Available: ¡http://oldwww.com.dtu.dk/teletraffic/handbook/telenookpdf.pdf¿.

[38] D. L. Jagerman, "Some properties of the Erlang loss function," *Bell Syst. Tech. J.*, vol. 53, pp. 525–511, 1974.

[39] J. S. Kauffman, "Blocking in a shared resource environment," *IEEE Transactions on Communications*, vol. COM-29, no. 10, pp. 1474-1481, 1981.

[40] F. P. Kelly, *Reversibility and stochastic networks*, John Wiley, 1979.

[41] 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.

[42] Kendall D. G., "Some problems in the theory of queues", *J. R. Statist. Soc. B*, vol. 13, pp. 151–173, 1951.

[43] L. Kleinrock, *Queueing Systems, Volume 1 - Theory*, John Wiley and Sons 1975.

[44] L. Kleinrock, *Queueing Systems, Volume 2 - Computer Applications*, John Wiley and Sons 1976.

[45] A. Leon Garcia, *Probability and random processes for electrical engineering*, Addisson Wesley, Reading, Mass., 1989.

[46] G. Latouche and V. Ramaswami, *Introduction to matrix analytic methods in stochastic modeling*, SIAM, 1999.

[47] D. V. Lindley, "The theory of queues with a single server", *Proc. Camb. Phil. Soc.*, vol. 48, pp. 277289, 1952.

[48] J. D. C. Little, "A proof of the queuing formula: L = λW", *Operations Research*, vol. 9, pp. 383-387, 1961.

[49] 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.

[50] 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.

[51] USC/ISI, Los Angeles, "The ns simulator and the documentation", http://www.isi.edu/nsnam/ns/ .

[52] C. R. Nelson, *Applied time series analysis*, Holden Day, San Francisco, 1973.

[53] M. F. Neuts, *Structured stochastic matrices of M/G/1 type and their applications*, Marcel Dekker, NY 1989.

[54] M. F. Neuts, *Matrix-geometric solutions in stochastic models - algorithmic approach*, The John Hopkin's University Press, Baltimore (1981).

[55] G. F. Newell, "The M/G/∞ queue", *SIAM Journal on Applied Mathematics*, vol. 14, no. 1, pp. 86-88, January 1966.

[56] A. Papoulis, *Probability, random variables and stochastic processes*, 2nd Edition, McGraw Hill Book Company, New York, 1984.

[57] E. Parzen, *Stochastic processes*, Holden Day, San Francisco, 1962.

[58] E. Reich, "On the integrodifferential equation of takacs. 1." *Ann. Math. Stat.*, vol. 29, pp. 563–570, 1958.

[59] J. Riordan, "Telephone traffic time averages", *Bell System technical Journal*, vol. 30, pp. 1129-1144, 1951.

[60] 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, 1891.

[61] 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.

[62] 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.

[63] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*, ISBN 3-540-19918-7, Springer-Verlag, Berlin, Heidelber, New York, 1995.

[64] K. W. Ross and D. Tsang, "Optimal circuit access policies in an ISDN environment: a Markov decision approch", *IEEE Transactions on communications*, vol. 37, pp. 934–939, 1989.

[65] K. W. Ross and D. Tsang, "The stochastic knapsack problem", *IEEE Transactions on communications*, vol. 37, pp. 740–747, 1989.

[66] K. W. Ross and D. Tsang, "Teletraffic engineering for product-form circuit-switched networks", *Advances in Applied Probability*, vol. 22, pp. 657–675, 1990.

[67] S. M. Ross, *Introduction to probability models*, Academic Press, New York, 1993.

[68] S. M. Ross, *A first course in probability*, Macmillan, New York, 1976.

[69] S. M. Ross, *Stochastic processes*, Wiley, New York, 1996.

[70] 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.

[71] S. Stidham, "L = λW: a discounted analogue and a new proof", *Operations Research*, vol. 20, pp. 1115–1126, 1972.

[72] S. Stidham, "A last word on L = λW", *Operations Research*, vol. 22, pp. 417–421, 1974.

[73] R. Syski, *Introduction to Congestion Theory in Telephone Systems*, 2nd ed. Elsevier Science Publishers B. V., 1986.

[74] 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.

[75] U. Yechiali and P. Naor, "Queuing Problems with Heterogeneous Arrivals and Service," *Operations Research*, vol. 19, pp. 722–734, 1971.

[76] K Zhu, H Zhu, B Mukherjee, *Traffic grooming in optical WDM mesh networks*, Springer, 2005.

[77] 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.

[78] 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.

[79] 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.

Exam/Midtest Formulae Sheet

$P(X=i)=e^{-\lambda}\frac{\lambda^i}{i!} \qquad i=0,\ 1,\ 2,\ 3,\ \ldots\ . \qquad f(x)=\begin{cases}\mu e^{-\mu x} & \text{if } x\geq 0\\ 0 & \text{otherwise.}\end{cases}$

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

$E[Q]=\lambda E[D] \qquad p_0 = 1-U = 1-\lambda/\mu \qquad E[Q]=\frac{\rho}{1-\rho} \qquad E[D]=\frac{1}{\mu(1-\rho)}=\frac{1}{\mu-\lambda}$

$\delta_D(x)=\begin{cases}(\mu-\lambda)e^{(\lambda-\mu)x} & \text{if } x\geq 0\\ 0 & \text{otherwise.}\end{cases} \qquad f_X(x)=\frac{\lambda^k x^{k-1}e^{-\lambda x}}{(k-1)!}.$

$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-m)^2/2\sigma^2} \qquad -\infty<x<\infty. \qquad P\left(X>x\right)=\begin{cases}\left(\frac{x}{\delta}\right)^{-\gamma}, & x\geq\delta\\ 1, & \text{otherwise.}\end{cases}$

$E[X]=\int_{-\infty}^{\infty} x f_X(x)dx. \qquad E[Z]=\int_0^\infty P(Z>z)dz=\int_0^\infty[1-F_Z(z)]dz.$

$E[Z]=\sum_{n=0}^{\infty}P(Z>n)=\sum_{n=0}^{\infty}[1-F_Z(n)]. \qquad var[X]=\sum_{\{k:P(k)>0\}}(k-E[X])^2P_X(k)$

$var[X]=\int_{-\infty}^{\infty}(x-E[X])^2 f_X(x)dx \qquad P(X>C)\leq\frac{E[X]}{C}. \qquad P(\mid X-E[X]\mid > C)\leq\frac{var[X]}{C^2}.$

random variable	parameters	mean	variance
Bernoulli	$0\leq p\leq 1$	p	$p(1-p)$
binomial	n and $0\leq p\leq 1$	np	$np(1-p)$
Poisson	$\lambda>0$	λ	λ
uniform	a and b	$(a+b)/2$	$(b-a)^2/12$
exponential	$\mu>0$	$1/\mu$	$1/\mu^2$
Gaussian	m and σ	m	σ^2
Pareto	$\delta>0$ and $1<\gamma<2$	$\frac{\delta\gamma}{(\gamma-1)}$	∞

$E_k(A)=\frac{\frac{A^k}{k!}}{\sum_{n=0}^{k}\frac{A^n}{n!}} \qquad var[X]=E[var[X\mid Y]]+var[E[X\mid Y]]$

$P_X(x)=E_Y[P(X=x\mid Y=y)] \qquad I_n(A)=A\int_0^\infty e^{-Ay}(1+y)^n dy \qquad U=\frac{(1-\pi_k)A}{k} \qquad U=\frac{A}{k}$

$\pi_0=\left(\sum_{n=0}^{k-1}\frac{A^n}{n!}+\frac{A^k}{k!}\frac{k}{k-A}\right)^{-1} \qquad C_k(A)=\sum_{n=k}^{\infty}\pi_n=\frac{A^k}{k!}\frac{k}{k-A}\pi_0=\frac{\frac{A^k}{k!}\frac{k}{k-A}}{\sum_{n=0}^{k-1}\frac{A^n}{n!}+\frac{A^k}{k!}\frac{k}{k-A}}$

$C_k(A)=\frac{kE_k(A)}{k-A[1-E_k(A)]} \qquad \pi_k=\rho^k\frac{1-\rho}{1-\rho^{k+1}}=\frac{\rho^k-\rho^{k+1}}{1-\rho^{k+1}}=\frac{\rho^k(1-\rho)}{1-\rho^{k+1}}$

$E[Q]=\rho+\frac{\rho^2+\lambda^2\sigma_s^2}{2(1-\rho)} \qquad W_q(t)=\sup_{0\leq s<t}\{W_a(t)-W_a(s)-t+s\}$

$E[D(j)]=E[W_Q(j)]+\frac{1}{\mu_j}$ for $j=1,2,3,\ldots,m \qquad E[D(j)]=\frac{(1/\mu_j)(1-\sum_{i=1}^{j}\rho_i)+R(j)}{(1-\sum_{i=1}^{j-1}\rho_i)(1-\sum_{i=1}^{j}\rho_i)}$

$E[X_n]=\lambda r/(\gamma-1) \qquad \gamma=3-2H \qquad B_R=1-\prod_{j\in R}(1-B(j)) \qquad \rho P_{loss}\leq P(Q>k)$

$\lambda_j=A_j+\sum_{i=1}^{N}\lambda_i P_{ij}, \qquad j=1,2,3,\ \ldots,N \qquad E_n(A)=\frac{AE_{n-1}(A)}{n+AE_{n-1}(A)}$

$P_b=\frac{\binom{M-1}{k}\rho^k}{\sum_{i=0}^{k}\binom{M-1}{i}\rho^i}. \qquad B_i=\frac{\rho(M-i)B_{i-1}}{i+\rho(M-i)B_{i-1}} \qquad E[Q]=C_k(A)\frac{A}{k-A}+A$