1

Introduction

Network calculus is a theory dealing with queuing systems found in computer networks. Specifically, network calculus is a theory for delay and other service guarantee analysis of computer networks. Its essential idea is to use alternate algebras, particularly the min-plus algebra and max-plus algebra, [6] to transform complex non-linear network systems into analytically tractable linear systems. Since its introduction in the early 1990s [28][29][138], network calculus has developed along two tracks—deterministic and stochastic. Deterministic network calculus has been employed in the design of computer networks to provide deterministic service guarantees for regulated flows. Excellent books summarizing results for deterministic network calculus are available [18][92]. However, service guarantees are typically required by multimedia flows in the network [42][43], which often can tolerate some amount of loss or (excess) delay. For such flows, the provision of stochastic service guarantees is more important because stochastic service guarantees can make better use of the multiplexing gain in the network. This is where stochastic network calculus makes an appearance. In addition, many networks, such as wireless networks and multi-access networks, may only provide stochastic service guarantees. In wireless networks, the capacity of a wireless channel varies over time in a random manner due to channel impairment, contention, and other causes. In multi-access networks such as CSMA (carrier sense multiple access) networks, the server capacity seen by a user is highly dependent on the traffic characteristics of other users. For the analysis and provision of service guarantees in such networks, stochastic network calculus becomes even more important.

This book is devoted to summarizing results for stochastic network calculus and organized as follows. The first chapter gives an introduction to service guarantee analysis, the basic properties required from a theory for tractable analysis of computer networks, and the mathematical background used in the book. The second chapter introduces fundamental concepts and results of deterministic network calculus. The concepts include arrival curve, service curve, and strict service curve. The results include the basic properties supported by deterministic network calculus. Starting in Chapter 3, we

Y. Jiang, Y. Liu, *Stochastic Network Calculus*,
DOI: 10.1007/978-1-84800-127-5_1,

introduce fundamental concepts and results for stochastic network calculus. Specifically, Chapter 3 introduces traffic models for stochastic network calculus and their relations with each other as well as with some well-known traffic models such as the effective bandwidth model. Chapter 4 defines server models for stochastic network calculus and introduces their relations with each other. Chapter 5 summarizes results related to the basic properties for stochastic network calculus under different combinations of traffic and server models introduced in Chapters 2 to 4. These results are presented without considering the possible independence of flows and servers. Similar to Chapter 5, Chapter 6 also presents results under various combinations of traffic and server models. The key difference between these two chapters is that Chapter 6 is devoted to independent case analysis, where flows and service processes are independent. From Chapter 7 to Chapter 9, several extensions and/or applications of stochastic network calculus are presented under different network cases. The appendix summarizes the book and discusses open research challenges in the area.

1.1 Quality of Service Guarantees

With the development and deployment of multimedia and network technologies, multimedia has become an indispensable feature on the Internet. Multimedia applications such as Internet telephony and Internet video make diverse requirements on the services provided by the network. Quality of Service (QoS) refers to the nature of the packet delivery service provided by the network and is the collective effect of service performances determining the degree of satisfaction of a user of the service.

A quality of service guarantee, or service guarantee for short, is either deterministic or stochastic.[1] A *deterministic service guarantee* guarantees that all packets of a flow arrive at the destination within its required performance measures such as throughput, delay, and loss bounds. While such deterministic service provides the highest QoS level, its most important drawback is that it must reserve network resources based on the worst-case scenario and hence leaves a significant portion of network resources unused on average. A *stochastic service guarantee* allows the QoS objectives specified by a flow to be guaranteed with a probability smaller than one. By allowing some packets to violate the required QoS measures, stochastic service guarantees can better exploit the statistical multiplexing gain at network links and hence improve network utilization.

A deterministic service guarantee may be modeled such that the experienced service must never be worse than the desired service, which may be expressed in the following form:

[1] The literature also uses *statistical service guarantee* or *probabilistic service guarantee* rather than the stochastic service guarantee in this book.

$$\Pr\{\text{Experienced service is not worse than desired service}\} = 1. \quad (1.1)$$

Many methods have been proposed in the literature to derive the worst-case bounds. The works, including [28][29][15][18][19] on deterministic QoS guarantee analysis, have been developed into an elegant theory under the name of network calculus [92], which will be referred to as *deterministic network calculus* in this book.

Similarly, a stochastic service guarantee may be expressed as

$$\Pr\{\text{Experienced service is worse than desired service}\} \leq \varepsilon, \quad (1.2)$$

where ε is the permissible probability that a packet violates the desired performance [42][43]. It can be seen that the deterministic service guarantee is a special case of the stochastic service guarantee with $\varepsilon = 0$ in (1.2). The focus of this book is on stochastic service guarantee analysis.

1.2 Basic Properties for Network Analysis

A computer network consists of data flows and network elements. Correspondingly, a theory for network analysis is typically built on two fundamental concepts: *traffic model* and *server model*. A traffic model characterizes the traffic behavior of a flow, and a server model characterizes the service behavior of a network element.

In order to easily apply a theory to network analysis, its traffic models and server models should satisfy some basic properties. The requirement of these basic properties is illustrated in the following example.

Consider a simple network domain consisting of three network nodes $S1$—$S3$ and three flows $F1$—$F3$, as shown in Figure 1.1. Assume $F1$ and $F2$ belong to the same traffic class and share the same edge-to-edge path crossing the network domain. At the second node $S2$, there is a crossing flow $F3$ that shares the server capacity of $S2$ with $F1$ and $F2$. Suppose we are interested in and want to analyze the edge-to-edge delay performance of the path on which $F1$ and $F2$ cross the network. While there can be many approaches to the analysis, the following is an intuitively simple one.

First, a certain traffic model $\mathcal{AM}$ and a certain server model $\mathcal{SM}$ should be properly chosen to represent flows and nodes, so that single-node analysis can be conducted to obtain the delay performance of a flow crossing the node. In addition, the single-node analysis should also give the characterization of the output, which can be represented with the same traffic model $\mathcal{AM}$, so that the single-node delay analysis can be repeatedly extended to a sequence of nodes. Given these, third in the simple approach, since $F1$ and $F2$ belong to the same traffic class and share the same path, an immediate idea is to use an aggregate flow $F1,2$ to represent the two flows before entering the first node $S1$. This implicitly requires that the aggregate flow $F1,2$ be represented using characteristics of both $F1$ and $F2$. With the first point in mind, the aggregate

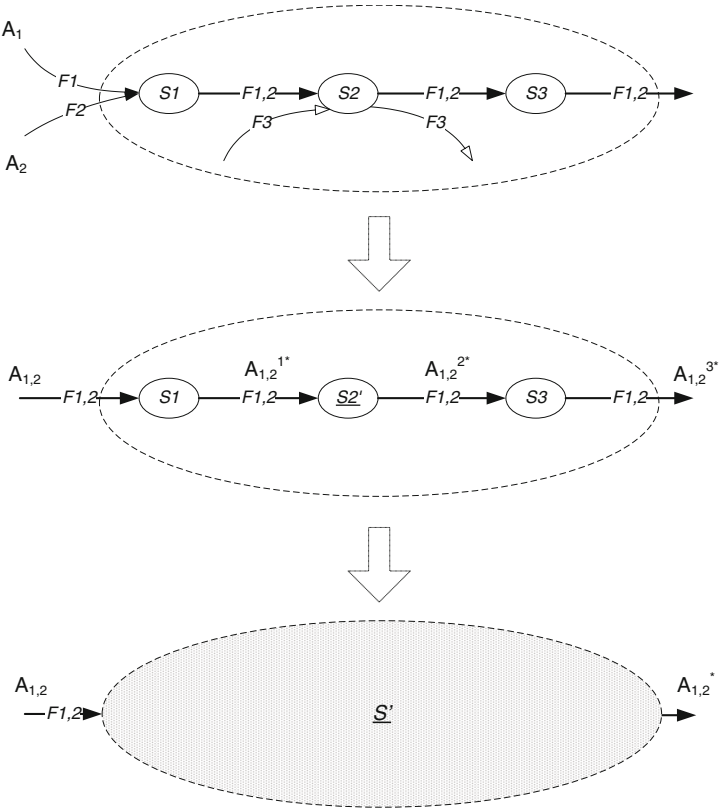

Fig. 1.1. Analysis of a simple network

flow $F1,2$ should be represented using the same traffic model $\mathcal{AM}$ as $F1$ and $F2$. Fourth, at the second node, the aggregate flow $F1,2$ is competing for service capacity with the crossing flow $F3$. A simple way for analysis is to find an equivalent server $S2'$ for the aggregate flow $F1,2$. Under this equivalent server, the aggregate flow is the only input and there is no crossing traffic. Again with the first point in mind, the equivalent server should be represented using the same server model $\mathcal{SM}$ as $S2$ and other nodes. Now, as shown by the middle figure of Figure 1.1, the network is simplified to a sequence of (possibly equivalent) servers on which the node-by-node analysis can be conducted to obtain the delay performance at each node on the path. Then, the edge-to-edge delay performance can be easily derived from the delay performance at each node. However, this is not the end since, as will be shown later (e.g., see Chapter 2), the results from node-by-node analysis can often be significantly improved if the so-called concatenation property exists. This property tells us that the concatenation of nodes can be treated as an equivalent node. As shown by the lower figure of Figure 1.1, with the concatenation property, based

on the middle figure, the network can be treated as a single black box node. By applying single-node analysis to the black-box, the desired edge-to-edge delay result is derived, which can be much better than the result obtained from the node-by-node analysis.

In summary, as discussed in the above example, a theory should ideally have the following five basic properties (P.1)—(P.5) to ease tractable network analysis.

- (P.1) – *Service Guarantees*
  Under a chosen traffic model and a chosen server model, single-node stochastic service guarantees such as backlog and delay guarantees can be derived.
- (P.2) – *Output Characterization*
  The output of a flow from a server can be represented using the same traffic model as for the input flow.
- (P.3) – *Concatenation Property*
  The concatenation of servers can be represented using the same server model.
- (P.4) – *Leftover Service*
  The service available to a flow at a server with competing flows can be represented using the same server model.
- (P.5) – *Superposition Property*:
  The superposition of flows can be represented using the same traffic model.

In traditional queuing theory, many models have been proposed to characterize arrival and service. Examples are Poisson arrival processes and service processes with negative exponentially distributed service times for $M/M/1$ systems. Also, a lot of results for single-node systems are available, which correspond to (P.1). Many results are also available from traditional queuing theory that corresponding to (P.5), particularly when sources are independent. For example, the aggregate of two Poisson arrival processes, if they are independent, results in a new Poisson arrival process. Under certain conditions, the output of an $M/M/1$ system can be considered to have a Poisson arrival process, which corresponds to (P.2). However, it is generally hard to conclude that the output can be represented using the same arrival process as the input in traditional queuing theory. In addition, very few results have so far been derived for (P.3) and (P.4) in traditional queuing theory. This partly explains why it is difficult to apply traditional queuing theory to network analysis.

Throughout the rest of this book, we will see that with the five basic properties (P.1)—(P.5), network service guarantee analysis can be conducted. Unless specially highlighted, the networks considered in this book are *feedforward networks* where there are no feedback flows.

1.3 Notation and Mathematical Background

In this book, we consider a discrete time domain with a unit discretization step. We adopt the convention that a packet is considered to be received by a network element when and only when its last bit has arrived at the network element, and a packet is considered out of a network element when and only when its last bit has been transmitted by the network element. A packet can be served only when its last bit has arrived. All queues are assumed to be empty at time 0. Packets within a flow are served in first-in-first-out (FIFO) order.

1.3.1 Notation

We use various processes to model a network that is assumed to be lossless. A process is defined to be a function of time $t(\geq 0)$. It could count the (cumulative) amount of traffic (in number of bits) arriving at some network element, the amount of traffic (in number of bits) departing from the network element, the amount of service (in number of bits) provided by the network element, or the amount of service (in number of bits) that failed to be provided by the network element due to some impairment to it. In this case, we call the process (cumulative) *arrival process,* denoted by $A(t)$, (cumulative) *departure process,* denoted by $A^*(t)$, (cumulative) *service process,* $S(t)$, or (cumulative) *impairment process,* $I(t)$, respectively. We assume all such processes are defined on $t \geq 0$ and by convention have zero value at $t = 0$. We also assume these functions are left-continuous.[2]

Wherever necessary, we use subscripts to distinguish different flows and superscripts to distinguish different network elements. Specifically, A_i^h and A_i^{h*} represent the arrival and departure processes of flow i from network element h, respectively, S_i^h the service process provided to flow i by the network element, and I^h the impairment process suffered by the network element.

For any $0 \leq s \leq t$, we denote $A(s,t) \equiv A(t) - A(s)$, $A^*(s,t) \equiv A^*(t) - A^*(s)$, $S(s,t) \equiv S(t) - S(s)$, and $I(s,t) \equiv I(t) - I(s)$.

In this book, the following function sets are often used. Specifically, we denote by $\mathcal{F}$ the set of non-negative wide-sense increasing functions, where for each function $a(\cdot)$ there holds

$$\mathcal{F} = \{a(\cdot) : \forall 0 \leq x \leq y, 0 \leq a(x) \leq a(y)\}$$

and for any function $a \in \mathcal{F}$ we set $a(x) = 0$ for $\forall x < 0$.

We denote by $\bar{\mathcal{F}}$ the set of non-negative wide-sense decreasing functions where for each function $a(\cdot)$ there holds

[2] Whether the functions are left-continuous or right-continuous does not make any difference to the results in this book (e.g., see Chapter 1.1 of [92] for the discussion).

$$\bar{\mathcal{F}} = \{a(\cdot) : \forall 0 \le x \le y, 0 \le a(y) \le a(x)\}$$

and for any function $a \in \bar{\mathcal{F}}$ we also set $a(x) = 1$ for $\forall x < 0$.

We denote by $\bar{\mathcal{G}}$ the set of functions in $\bar{\mathcal{F}}$ where for each function $a(\cdot) \in \bar{\mathcal{G}}$ its nth-fold integration, denoted by $f^{(n)}(x) \equiv (\int_x^\infty dy)^n f(y)$, is bounded for any $x \ge 0$ and still belongs to $\bar{\mathcal{G}}$ for any $n \ge 0$, or

$$\bar{\mathcal{G}} = \{a(\cdot) : \forall n \ge 0, \left(\int_x^\infty dy\right)^n a(y) \in \bar{\mathcal{G}}\}.$$

A function a is said to be additive if and only if, for all x, y $a(x+y) = a(x) + a(y)$. The function is said to be sub-additive if and only if $a(x+y) \le a(x) + a(y)$ for all x and y.

For any non-negative functions a, b, the following inequalities hold trivially:

$$\sup_{0\le y\le x}[a(y) + b(y)] \le \sup_{0\le y\le x} a(y) + \sup_{0\le y\le x} b(y), \tag{1.3}$$

$$\inf_{0\le y\le x}[a(y) - b(y)] \ge \inf_{0\le y\le x} a(y) - \sup_{0\le y\le x} b(y). \tag{1.4}$$

By definition, $A(t)$, $A^*(t)$, $S(t)$, and $I(t)$ belong to $\mathcal{F}$. In addition, it can be shown that all the exponentially decaying functions and functions exhibiting sub-exponential decay belong to $\bar{\mathcal{G}}$.

For any random variable X, its cumulative distribution function (CDF), denoted by $F_X(x) \equiv P\{X \le x\}$, belongs to $\mathcal{F}$ and its complementary cumulative distribution function (CCDF), denoted by $\bar{F}_X \equiv P\{X > x\}$, belongs to $\bar{\mathcal{F}}$.

The *conventional convolution* of two functions a, b is defined as

$$(a \circledast b)(x) = \int_{-\infty}^{\infty} a(x-y)b(y)dy,$$

and the *Stieltjes convolution* of two functions a, b is defined as

$$(a * b)(x) = \int_{-\infty}^{\infty} a(x-y)db(y).$$

We use $[\cdot]^+$ to express the maximum of 0 and a given number, or $[x]^+ \equiv \max\{x, 0\}$. We shall also use $[\cdot]_1$ to denote the minimum of 1 and the given number, i.e., $[x]_1 \equiv \min\{x, 1\}$.

For service guarantee analysis of a system, which could be a network element or a network of elements, we are mainly interested in the *backlog* and *delay,* which are defined as follows [30][18][92]:

Definition 1.1. *Let $A(t)$ and $A^*(t)$ respectively be the arrival process and departure process of a lossless system. The backlog $B(t)$ in the system at time $t \ge 0$ is defined as*

$$B(t) = A(t) - A^*(t).$$

Assuming first-in-first-out (FIFO) ordering, the delay $D(t)$ at time $t \ge 0$ is defined as

$$D(t) = \inf\{d \ge 0 : A(t) \le A^*(t+d)\}.$$

1.3.2 Min-Plus Algebra Basics

In conventional algebra, addition $+$ and multiplication $\times$ are the two most common operations on elements of $\mathcal{R} = (-\infty, +\infty)$. These two operations have a number of properties, such as the closure property, associativity, commutativity, and distributivity, which make the algebraic structure $(\mathcal{R}, +, \times)$ a commutative field.

In min-plus algebra, an algebra structure of interest is $(\mathcal{R} \cup \{+\infty\}, \wedge, +)$. Here, the "addition" operation is $\wedge$ and the "multiplication" operation is $+$, where $\wedge$ denotes the *infimum* or, when it exists, the *minimum*. It can be verified that $(\mathcal{R} \cup \{+\infty\}, \wedge, +)$ has the following properties, and it is called a commutative dioid with zero element $\bar{\epsilon} = +\infty$ and identity element $\mathbf{e} = 0$:

- Closure property: $\forall a, b \in (\mathcal{R} \cup \{+\infty\})$, $a \wedge b \in (\mathcal{R} \cup \{+\infty\})$; $a + b \in (\mathcal{R} \cup \{+\infty\})$.
- Associativity: $\forall a, b \in (\mathcal{R} \cup \{+\infty\})$, $(a \wedge b) \wedge c = a \wedge (b \wedge c)$; $(a + b) + c = a + (b + c)$.
- Commutativity: $\forall a, b \in (\mathcal{R} \cup \{+\infty\})$, $a \wedge b = b \wedge a$; $a + b = b + a$.
- Distributivity: $\forall a, b, c \in (\mathcal{R} \cup \{+\infty\})$, $(a \wedge b) + c = (a + b) \wedge (b + c)$.
- Zero element: $\forall a \in (\mathcal{R} \cup \{+\infty\})$, $a \wedge \bar{\epsilon} = a$.
- Absorbing zero element: $\forall a \in (\mathcal{R} \cup \{+\infty\})$, $a + \bar{\epsilon} = \bar{\epsilon} + a = \bar{\epsilon}$.
- Identity element: $\forall a \in (\mathcal{R} \cup \{+\infty\})$, $a + \mathbf{e} = \mathbf{e} + a = a$.
- Idempotency of addition: $\forall a \in (\mathcal{R} \cup \{+\infty\})$, $a \wedge a = a$.

For functions in min-plus algebra, the following operations are often used.

The *pointwise infimum*, or *pointwise minimum* if it exists, of functions a and b is

$$(a \wedge b)(x) = \inf[a(x), b(x)].$$

The *pointwise supremum*, or *pointwise maximum* if it exists, of functions a and b is

$$(a \vee b)(x) = \sup[a(x), b(x)].$$

The *min-plus convolution* of functions a and b is

$$(a \otimes b)(x) = \inf_{0 \le y \le x} [a(y) + b(x - y)],$$

where, when it applies, "infimum" should be interpreted as "minimum".

The *min-plus deconvolution* of functions a and b is

$$(a \oslash b)(x) = \sup_{y \ge 0} [a(x + y) - b(y)],$$

where, when it applies, "supremum" should be interpreted as "maximum".

It can be verified that $(\mathcal{F}, \wedge, \otimes)$ also has the following properties and is a commutative dioid with zero element $\bar{\epsilon}$ and identity element $\mathbf{e}$, where $\bar{\epsilon}(x) = +\infty$ for all $x \ge 0$ and $\mathbf{e}(x) = 0$ if $x = 0$ and otherwise $+\infty$ [6][19][92]:

- Closure property: $\forall a, b \in \mathcal{F}$, $a \wedge b \in \mathcal{F}$; $a \otimes b \in \mathcal{F}$.
- Associativity: $\forall a, b \in \mathcal{F}$, $(a \wedge b) \wedge c = a \wedge (b \wedge c)$; $(a \otimes b) \otimes c = a \otimes (b \otimes c)$.
- Commutativity: $\forall a, b \in \mathcal{F}$, $a \wedge b = b \wedge a$; $a \otimes b = b \otimes a$.
- Distributivity: $\forall a, b, c \in \mathcal{F}$, $(a \wedge b) \otimes c = (a \otimes b) \wedge (b \otimes c)$.
- Zero element: $\forall a \in \mathcal{F}$, $a \wedge \bar{\epsilon} = a$.
- Absorbing zero element: $\forall a \in \mathcal{F}$, $a \otimes \bar{\epsilon} = \bar{\epsilon} \otimes a = \bar{\epsilon}$.
- Identity element: $\forall a \in \mathcal{F}$, $a \otimes \mathbf{e} = \mathbf{e} \otimes a = a$.
- Idempotency of addition: $\forall a \in \mathcal{F}$, $a \wedge a = a$.

The following properties also hold for $(\mathcal{F}, \wedge, \otimes)$:

- Comparison: For $\forall a_1, a_2, b_1, b_2 \in \mathcal{F}$, $a_1 \otimes a_2 \leq a_1 \wedge a_2 \leq a_1 \vee a_2$.
- Monotonicity: For $\forall a_1, a_2, b_1, b_2 \in \mathcal{F}$, if $a_1 \leq b_1$ and $a_2 \leq b_2$, then $a_1 \otimes a_2 \leq b_1 \otimes b_2$; $a_1 \wedge a_2 \leq b_1 \wedge b_2$; $a_1 \vee a_2 \leq b_1 \vee b_2$.

Similarly, it can be shown that $(\bar{\mathcal{F}}, \wedge, \otimes)$ is a commutative dioid, but with zero element $\bar{\epsilon}$ and identity element $\bar{e}$, where $\bar{\epsilon}(x) = +\infty$ for all $x \geq 0$ and $\bar{\mathbf{e}}(x) = 0$ for all $x \geq 0$. Specifically, $(\bar{\mathcal{F}}, \wedge, \otimes)$ has the following properties:

- Closure property: $\forall a, b \in \bar{\mathcal{F}}$, $a \wedge b \in \bar{\mathcal{F}}$; $a \otimes b \in \bar{\mathcal{F}}$.
- Associativity: $\forall a, b, c \in \bar{\mathcal{F}}$, $(a \wedge b) \wedge c = a \wedge (b \wedge c)$; $(a \otimes b) \otimes c = a \otimes (b \otimes c)$.
- Commutativity: $\forall a, b \in \bar{\mathcal{F}}$, $a \wedge b = b \wedge a$; $a \otimes b = b \otimes a$.
- Distributivity: $\forall a, b, c \in \bar{\mathcal{F}}$, $(a \wedge b) \otimes c = (a \otimes b) \wedge (b \otimes c)$.
- Zero element: $\forall a \in \bar{\mathcal{F}}$, $a \wedge \bar{\epsilon} = a$.
- Absorbing zero element: $\forall a \in \bar{\mathcal{F}}$, $a \otimes \bar{\epsilon} = \bar{\epsilon} \otimes a = \bar{\epsilon}$.
- Identity element: $\forall a \in \bar{\mathcal{F}}$, $a \otimes \bar{\mathbf{e}} = \bar{\mathbf{e}} \otimes a = a$.
- Idempotency of addition: $\forall a \in \bar{\mathcal{F}}$, $a \wedge a = a$.
- Comparison: $a_1 \wedge a_2 \leq a_1 \vee a_2 \leq a_1 \otimes a_2$.
- Monotonicity: If $a_1 \leq b_1$ and $a_2 \leq b_2$, then $a_1 \otimes a_2 \leq b_1 \otimes b_2$; $a_1 \wedge a_2 \leq b_1 \wedge b_2$; $a_1 \vee a_2 \leq b_1 \vee b_2$.

Besides the various properties summarized above, the min-plus convolution $\otimes$ implies the following [92].

Lemma 1.2. *If a is left-continuous and b is continuous, then for any t there exists some t_0 such that*

$$a \otimes b\,(t) \equiv a\,(t - t_0) + b\,(t_0)\,. \tag{1.5}$$

Additionally, it can be verified that for any functions α and β, there holds

$$(\alpha + c) \otimes \beta = \alpha \otimes \beta + c, \tag{1.6}$$

where c is any constant.

If α and β are sub-additive and $\alpha(0) = \beta(0) = 0$, there hold

$$\alpha \otimes \alpha = \alpha, \tag{1.7}$$

$$\alpha \otimes \beta = \alpha \text{ if } \alpha \leq \beta, \tag{1.8}$$

where c is any constant.

Furthermore, if functions α and β are sub-additive, so are $\alpha \otimes \beta$, $\alpha \wedge \beta$, and $\alpha \oslash \beta$.

1.3.3 Maximum Horizontal Distance and Maximum Vertical Distance

For ease of exposition of results, we adopt the following definitions [31] [92], which will be used throughout the rest of the book.

Definition 1.3. *Consider two functions $\alpha(t)$ and $\beta(t)$. The maximum horizontal distance between them, denoted by $h(\alpha, \beta)$, is defined as*

$$h(\alpha, \beta) = \sup_{s \geq 0} \{\inf \{\tau \geq 0 : \alpha(s) \leq \beta(s + \tau)\}\},$$

and the maximum vertical distance between them, denoted by $v(\alpha, \beta)$, is defined as

$$v(\alpha, \beta) = \sup_{s \geq 0} \{\alpha(s) - \beta(s)\} \equiv \alpha \oslash \beta(0).$$

Figure 1.2 illustrates these two concepts using functions $\alpha(t)$ and $\beta(t)$.

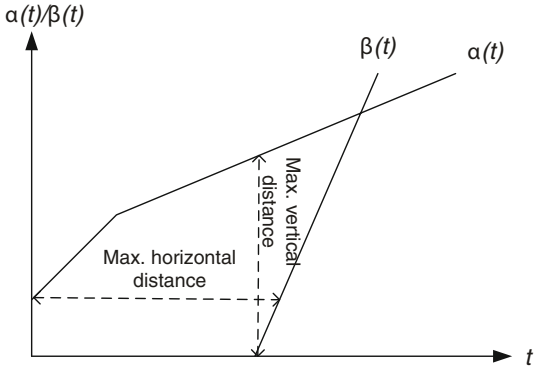

Fig. 1.2. Maximum horizontal and vertical distances between two functions

1.4 Random Variable and Stochastic Process Basics

1.4.1 Random Variables

A *random variable* X is characterized by its *cumulative distribution function* (CDF) $F_X(x)$, defined as

$$F_X(x) = P\{X \leq x\}, -\infty < x < \infty.$$

$F_X(x)$ is a non-negative, never decreasing function of x and belongs to $\mathcal{F}$. In addition, $F(-\infty) = 0$ and $F(\infty) = 1$. The *complementary cumulative*

distribution function (CCDF) of the random variable X, denoted by $\bar{F}_X$, is defined as

$$\bar{F}_X = P\{X > x\}, -\infty < x < \infty.$$

It is trivial that $F_X(x) + \bar{F}_X(x) = 1$ for any x. In addition, $\bar{F}_X$ is non-negative and non-increasing, belonging to $\bar{\mathcal{F}}$. Furthermore, $\bar{F}(-\infty) = 1$ and $\bar{F}(\infty) = 0$.

The *probability density function* (pdf) of a random variable X, denoted by $f_X(x)$, is defined as

$$f_X(x) \equiv \frac{dF_X(x)}{dx}.$$

Given the probability density function of a random variable X, its cumulative distribution function is found as

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy.$$

The *moment generating function* (MGF) of a random variable X, denoted by $M_X(\theta)$, is defined as

$$\begin{aligned} M_X(\theta) &\equiv E[e^{\theta X}] \\ &= \int_{-\infty}^{\infty} e^{\theta x} f_X(x)dx, \end{aligned}$$

where θ is a real variable.

The following inequality, known as the Chernoff bound, gives an upper bound on the CCDF of a random variable X:

$$P\{X \geq x\} \leq e^{-\theta x} E[e^{\theta X}]$$

for all $\theta \geq 0$.

Lemma 1.4. *Consider a random variable X. For any $x \geq 0$, $P\{(X)^+ > x\} = P\{X > x\}$.*

In this book, we are often concerned about the sum of a collection of random variables $\{X_i\}$, namely

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

For $Z\,(= \sum_{i=1}^{n} X_i)$, if $X_1, X_2, \ldots, X_n$ are independent, it is known that

$$f_Z(z) = f_{X_1} \circledast f_{X_2} \circledast \cdots \circledast f_{X_n}(z), \tag{1.9}$$

where the convolution is commutative. In addition,

$$M_Z(\theta) = M_{X_1}(\theta) \cdot M_{X_2}(\theta) \cdot \cdots \cdot M_{X_n}(\theta). \tag{1.10}$$

In this book, corresponding to (1.9), if $X_1, X_2, \ldots, X_n$ are independent, we often use Stieltjes convolution for F_Z,

$$F_Z(z) = F_{X_1} * F_{X_2} * \cdots * F_{X_n}(z), \tag{1.11}$$

where the Stieltjes convolution is also commutative.

In addition, if $X_1, X_2, \ldots, X_n$ are possibly dependent, the following result is important.

Lemma 1.5. *For the sum of a collection of random variables $Z = \sum_{i=1}^{n} X_i$, no matter whether they are independent or not, there holds for the CCDF of Z*

$$\bar{F}_Z(z) \leq \bar{F}_{X_1} \otimes \cdots \otimes \bar{F}_{X_n}(z). \tag{1.12}$$

Proof. We only prove for the sum of two random variables X_1 and X_2, and the proof can be easily extended to $n > 2$.

For any $z \geq x \geq 0$, $\{X_1 + X_2 > z\} \cap \{X_1 \leq x\} \cap \{X_2 \leq z - x\} = \phi$, where ϕ denotes the null set. We then have

$$\{X_1 + X_2 > z\} \subset \{X_1 > x\} \cup \{X_2 > z - x\}$$

and hence

$$P\{X_1 + X_2 > z\} \leq P\{X_1 > x\} + P\{X_2 > z - x\}.$$

Since the above inequality holds for all x, $(0 \leq x \leq z)$, we get

$$P\{X_1 + X_2 > z\} \leq \inf_{0 \leq x \leq z} [P\{X_1 > x\} + P\{X_2 > z - x\}]$$

or

$$\bar{F}_{X_1+X_2}(z) \leq \bar{F}_{X_1} \otimes \bar{F}_{X_2}(z).$$

While we shall mainly use forms similar to (1.12) to ease expressing results related to the sum of random variables, there are other inequalities that can be used to find upper bounds on the CCDF of Z. These inequalities can indeed be applied to all corresponding results in this book concerning the sum of multiple random variables. One of these inequalities is as follows.

Lemma 1.6. *For the sum of a collection of random variables $Z = \sum_{i=1}^{n} X_i$, no matter whether they are independent or not, there holds for the CCDF of Z*

$$\bar{F}_Z(z) \leq \inf_{p_1 + \cdots + p_n = 1} \{\bar{F}_{X_1}(p_1 z) + \cdots + \bar{F}_{X_n}(p_n z)\} \tag{1.13}$$

for any $1 > p_i > 0$, $i = 1, \ldots, n$, satisfying $\sum_{i=1}^{n} p_i = 1$.

1.4.2 Stochastic Processes

A *stochastic process* $X(t)$ is a collection of random variables $\{X(t), t \in T\}$ defined for each t in the index set T. The stochastic process is similarly characterized by its *cumulative distribution function* (CDF) $F_X(x,t)$, defined as for any (allowed) t,

$$F_X(x,t) = P\{X(t) \leq x\}, -\infty < x < \infty.$$

For any t, $F_X(x,t)$ is also non-negative, never decreasing on x, and belongs to $\mathcal{F}$. In addition, $F(-\infty,t) = 0$ and $F(\infty,t) = 1$.

The *complementary cumulative distribution function* (CCDF) of the stochastic process $X(t)$ is defined as

$$\bar{F}_X(X,t) = P\{X(t) > x\}, -\infty < x < \infty.$$

For any t, $\bar{F}_X(x,t)$ is non-negative and non-increasing and belongs to $\bar{\mathcal{F}}$. In addition, $\bar{F}(-\infty,t) = 1$ and $\bar{F}(\infty,t) = 0$. Furthermore, $F_X(x,t)+\bar{F}_X(x,t) = 1$ for all x and any t.

The *probability density function* (pdf) and the *moment generating function* of a stochastic process $X(t)$ are respectively defined as

$$f_X(x,t) \equiv \frac{dF_X(x,t)}{dx}$$

and

$$\begin{aligned} M_X(\theta(t),t) &\equiv E[e^{\theta(t)X(t)}] \\ &= \int_{-\infty}^{\infty} e^{\theta(t)x} f_X(x,t)dx, \end{aligned}$$

where $\theta(t)$ is a real variable possibly dependent on t.

A stochastic process $X(t)$ is said to be *stationary* if $F_X(x,t)$ remains unchanged when t shifts, that is, for any given constant τ, there holds

$$F_X(x,t+\tau) = F_X(x,t).$$

In the stationary case, for ease of expression, we often simply use $F_X(x)$ and $\bar{F}_X(x)$ to represent the CDF and CCDF, respectively.

1.4.3 Stochastic Ordering

For any two random variables X and Y, if $\bar{F}_X(x) \leq \bar{F}_Y(x)$ for all x, or in other words,

$$P\{X > x\} \leq P\{Y > x\} \text{ for all } x,$$

we then say X is stochastically smaller than Y [130][118], written as

$$X \leq_{st} Y.$$

The same notation applies when X and Y are random vectors.

Similarly, for any two stochastic processes $X(t)$ and $Y(t)$, we say $X(t)$ is stochastically smaller than $Y(t)$, written $X(t) \leq_{st} Y(t)$, if, for any t and all x, $P\{X(t) > x\} \leq P\{Y(t) > x\}$.

For two random variables X and Y, the following result holds [130]:

Lemma 1.7. *If $X \leq_{st} Y$, then $f(X) \leq_{st} f(Y)$ for any increasing function f.*

For the same mapping function of random variables, if these random variables are independent, the following result holds (see, e.g., Theorem 2.2.3 in [130]).

Lemma 1.8. *Let $X_1, \ldots, X_n$ be independent and $Y_1, \ldots, Y_n$ be independent. If $X_i \leq_{st} Y_i$, then for any wide-sense increasing function $\Phi(z_1, \ldots, z_n)$ on z_i $(i = 1, \ldots, n)$, there holds*

$$\Phi(X_1, \ldots, X_n) \leq_{st} \Phi(Y_1, \ldots, Y_n).$$

Example 1.9. As an example of Lemma 1.8, letting $\Phi(z_1, \ldots, z_n) = \sum_{i=1}^{n} z_i$, if $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are independent and $X_i \leq_{st} Y_i$ for all $i = 1, \ldots, n$, we get $\sum_{i=1}^{n} X_i \leq_{st} \sum_{i=1}^{n} Y_i$.

Example 1.10. Let $\Phi(z_1, \ldots, z_n) = \max\{z_1, \ldots, z_n\}$, which can be verified to be wide-sense increasing on z_i. Then, from Lemma 1.8, if $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are independent and $X_i \leq_{st} Y_i$ for all $i = 1, \ldots, n$, we can conclude that $\max\{X_1, \ldots, X_n\} \leq_{st} \max\{Y_1, \ldots, Y_n\}$. The conclusion holds also for mapping $\Phi(z_1, \ldots, z_n) = \min\{z_1, \ldots, z_n\}$.

For the same mapping function of random variables that are unknown if they are independent, if certain conditions hold, we have the following result (see, e.g., Theorem 2.2.4 in [130] or Theorem 4.3.3 in [108]).

Lemma 1.11. *Suppose that for random variables $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$, there holds $\{X_1, \ldots, X_n\} \leq_{st} \{Y_1, \ldots, Y_n\}$. Then, for the mapping $Z(t) = \Phi(z_1, \ldots, z_n)$, if it is nondecreasing in $\{z_1, \ldots, z_n\}$, one has $Z'(t) \leq_{st} Z''(t)$, where $Z'(t) = \Phi(X_1, \ldots, X_n)$ and $Z'' = \Phi(Y_1, \ldots, Y_n)$.*

Example 1.12. For the mappings in Examples 1.9 and 1.10, if $\{X_1, \ldots, X_n\} \leq_{st} \{Y_1, \ldots, Y_n\}$, then, based on Lemma 1.11, the same stochastic ordering conclusions hold: i.e., $\sum_{i=1}^{n} X_i \leq_{st} \sum_{i=1}^{n} Y_i$, $\max\{X_1, \ldots, X_n\} \leq_{st} \max\{X_1, \ldots, X_n\}$, and $\min\{X_1, \ldots, X_n\} \leq_{st} \min\{X_1, \ldots, X_n\}$.

1.5 Min-Plus Linearity of Queuing Systems

Consider a lossless network queuing system with arrival process $A(t)$, service process $S(t)$, and departure process $A^*(t)$. In this system, the input is $A(t)$ and the output is $A^*(t)$.

By the definition of backlog in the system, the following relationship holds:

$$A^*(t) = A(t) - B(t). \tag{1.14}$$

The Lindley equation can be used to derive $B(t)$, which is

$$B(t) = \max\{0, B(t-1) + A(t-1,t) - S(t-1,t)\}. \tag{1.15}$$

Equation (1.15) is intuitively clear and says that the amount of traffic backlogged in the system at time t equals the amount of traffic backlogged at time $t-1$ plus the amount of traffic that arrived between $t-1$ and t minus the amount of traffic serviced between $t-1$ and t. By applying (1.15) iteratively to its right-hand side, it becomes

$$B(t) = \sup_{0\le s\le t}\{A(s,t) - S(s,t)\}. \tag{1.16}$$

Applying (1.16) to (1.14) results in

$$A^*(t) = \inf_{0\le s\le t}\{A(s) + S(s,t)\} = A \otimes S(t). \tag{1.17}$$

Equation (1.17) establishes the relationship between the output and the input of the queuing system considered.

Relationship (1.17) is very similar to a relationship commonly found for conventional linear communication systems where there holds

$$A^*(t) = A \circledast S(t) \tag{1.18}$$

with $S(t)$ being the impulse response of the system. For such a system, suppose $A(t) = a_1 \times A_1(t) + a_2 \times A_2(t)$ and denote by $A_i^*(t)$ the output of the system when there is only $A_i(t)$ as the input, $i = 1, 2$. The following linearity property holds: For any non-negative constants a_1 and a_2,

$$\begin{aligned} A^*(t) &= [a_1 \times A_1(t) + a_2 \times A_2(t)] \circledast S(t) \\ &= a_1 \times A_1 \circledast S(t) + a_2 \times A_2 \circledast S(t) \\ &= a_1 \times A_1^*(t) + a_2 \times A_2^*(t). \end{aligned} \tag{1.19}$$

Relationship (1.17), however, implies that the queuing system considered is non-linear in the conventional sense with the algebra structure $(\mathcal{R}, +, \times)$.

Suppose now that the input process is the min-plus addition of two processes $A_1(t)$ and $A_2(t)$ in the form

$$A(t) = (a_1 + A_1(t)) \wedge (a_2 + A_2(t)), \tag{1.20}$$

where a_1 and a_2 are any two non-negative constants. Similarly, let us denote by $A_i^*(t)$ the output of the system when there is only $A_i(t)$ as the input, $i = 1, 2$.

Then, from (1.17) and the properties of $\wedge$, $+$ and $\otimes$, we obtain the output from the system as

$$\begin{aligned} A^*(t) &= [(a_1 + A_1(t)) \wedge (a_2 + A_2(t))] \otimes S(t) \\ &= [(a_1 + A_1(t)) \otimes S(t)] \wedge [(a_2 + A_2(t)) \otimes S(t)] \\ &= [a_1 + A_1 \otimes S(t)] \wedge [a_2 + A_2 \otimes S(t)] \\ &= [a_1 + A_1^*(t)] \wedge [a_2 + A_2^*(t)]. \end{aligned} \tag{1.21}$$

Relationship (1.21) implies that the queuing system considered is linear with the min-plus algebra structure $(\mathcal{R} \cup \{+\infty\}, \wedge, +)$.

1.6 Summary and Bibliographic Comments

This chapter gives a brief introduction to network service guarantee analysis. The five basic properties needed by a theory for systematic network analysis are discussed. To help understand the analysis in the subsequent chapters, some useful notations and mathematical background are introduced that include min-plus algebra, random variable, and stochastic process basics.

The need for the five basic properties has been extensively discussed in the literature. For deterministic network calculus, a complete study of these properties can be found in [18] [92]. For stochastic network calculus, they have also been studied in the literature, although in most cases separately. For example, the superposition property, the output characterization property, and the service guarantee property were studied in [87] [138] [15]. The leftover service property was addressed in [99] [115]. The need for the concatenation property was independently discussed in [73] [24]. The initial effort of addressing the five basic properties together was made by Jiang and Emstad [73]. Jiang [69] proved for the first time all the five basic properties for both the general case and independent case under some specific traffic and server models to be introduced in Chapters 2 and 3.

The notation $(\int_x^\infty dy)^n f(y)$ and the special function set $\bar{\mathcal{G}}$ were initially introduced by Starobinski and Sidi [128] to stochastic service guarantee analysis. The requirement that functions belong to $\bar{\mathcal{G}}$ comes from relations that will be shown in Chapter 5, between the output and the input of a network element, between the delay and backlog performances and the input and the service of the network element, as well as between the service of a concatenation of network elements and the service of each network element. In this book, we often require for any order n that the multiple integral $(\int_x^\infty dy)^n f(y)$ be bounded. However, if the size of the network is known *a priori*, this requirement can be correspondingly relaxed for n.

Problems

1.1. Prove the properties of $(\mathcal{F}, \wedge, \otimes)$.

1.2. Prove the properties of $(\bar{\mathcal{F}}, \wedge, \otimes)$.

1.3. Prove the commutativity property of the conventional convolution operation $\circledast$.

1.4. Prove the commutativity property of the Stieltjes convolution operation $*$ when the two functions are cumulative distribution functions.

1.5. Prove Lemma 1.2.

1.6. Show the maximum horizontal and vertical distances for $\alpha(t)$ and $\beta(t)$ as shown in Figure 1.2.

1.7. Let $\alpha(t) = \min\{M + C \cdot t, \rho t + \sigma\}$ and $\beta(t) = r \cdot t + \theta$ with $r \geq \rho > 0$ and $C > \rho > 0$. Show the maximum horizontal and vertical distances for $\alpha(t)$ and $\beta(t)$.

1.8. Prove Lemma 1.4.

1.9. Prove Lemma 1.6.

1.10. Prove Lemma 1.7.

1.11. Prove Lemma 1.8.