Jing Xie

A Temporal Network Calculus for Performance Analysis of Computer Networks

Thesis for the degree of Philosophiae Doctor

Trondheim, June 2011

Norwegian University of Science and Technology
Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Telematics

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Telematics

ISBN 978-82-471-2950-0 (printed ver.)
ISBN 978-82-471-2951-7 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2011:199

Printed by NTNU-trykk

Abstract

One inevitable trend of network development is to deliver information with various traffic characteristics and diverse Quality of Service (QoS) requirements. In response to the continually growing demand for more bandwidth, network performance analysis is needed to optimize the performance of existing technologies and evaluate the efficiency of new ones. Performance analysis investigates how traffic management mechanisms deployed in the network affect the resource allocation among users and the performance which the users experience. This topic can be investigated by constructing models of traffic management mechanisms and studying how these mechanisms perform under various types of network traffic.

To this end, appropriate mathematical models are needed to characterize the traffic management mechanisms which we are interested in and represent different types of network traffic. In addition, fundamental properties which can be employed to manipulate the models should be explored.

Over the last two decades a relatively new theory, *stochastic network calculus*, has been developed to enable mathematical performance analysis of computer networks. Particularly, several related processes are mathematically modeled, including the arrival process, the waiting process and the service process. This theory can be applied to the derivation and calculation of several performance metrics such as the backlog bound and the delay bound. The most attractive contribution of stochastic network calculus is to characterize the behavior of a process based on some bound on the complementary cumulative distribution function (CCDF). The behavior of a computer network is often subject to many irregularities and stochastic fluctuations. The models based on the bound on the CCDF are not very accurate, while they are more feasible for abstracting computer network systems and

representing various types of network traffic.

This thesis is devoted to investigate the performance of networks from the temporal perspective. Specifically, the traffic arrival process characterizes the distribution of the cumulative inter-arrival time and the service process describes the distribution of the cumulative service time. Central to finding a bound on the CCDF of the cumulative inter-arrival time and the cumulative service time, several variations of the traffic characterization and the service characterization are developed. The purpose of developing several variations to characterize the same process is to facilitate the derivation and calculation of performance metrics.

In order to derive and calculate the performance metrics, four fundamental properties are explored, including the service guarantees, the output characterization, the concatenation property and the superposition property. The four properties can be combined differently when deriving the performance metrics of a single node, a series of nodes or the superposition flow.

Compared to the available literature on stochastic network calculus which mainly focuses on studying network performance in the space-domain, this work develops a generic framework for mathematically analyzing network performance in the time-domain. The potential applications of this temporal approach include the wireless networks and the multi-access networks.

Furthermore, the complete procedure of concretizing the generic traffic models and service models is presented in detail. It reveals the key of applying the developed temporal network calculus approach to network performance analysis, i.e., to derive the bounding function which is the upper bound on the tail probability of a stochastic process. Several mathematical methods are introduced, such as the martingale, the moment generating function (MGF) and a concentration theory result.

Preface

This thesis is submitted in partial fulfillment of the requirement for the degree of philosophiae doctor (PhD) at the Norwegian University of Science and Technology (NTNU). During the four years of doctoral study, I have been funded by the Department of Telematics, NTNU. Particularly, the four years consisted of three years research work and one year teaching assistant duty.

I had the chance of coming to Trondheim and pursuing my doctoral degree because Professor Yuming Jiang forwarded an important email to me in February, 2006. This email was my first impression of NTNU where I have studied for four years and two months. I have only one supervisor, Yuming Jiang. This is different from most other PhD students at Telematics who often have two supervisors. However, his prophetic guidance in conveying knowledge has been greatly beneficial to this work. Discussing with him always cleared up my confusions. I have very much to learn from his meticulous scholarship. One of the most valuable things I have learned from his supervision is to objectively evaluate peers' work and my own work. The thesis could not have been written without him who not only supervised me but also inspired and challenged me throughout my Phd study. I really appreciate all the academic training he provided. I am very grateful for all his efforts spent on commenting and revising my papers, presentations and this thesis.

For this thesis I would like to thank the evaluation committee, Professor Sabine Wittevrongel, Professor Lie-Liang Yang and Professor Peder Johannes Emstad. Particularly, I would like to express my thanks to Professor Øivind Kure for his time and effort spent on coordinating my PhD defense.

I would like to thank all former and current colleagues at Telematics for creating a friendly, fair and enjoyable work atmosphere. Espe-

cially, I want to mention Randi, Mona, Pål and Asbjørn, who helped me to solve numberless practical and trivial problems and go through the complicated administrative processes.

Thanks for all friends with whom I could share my troubles and express my happiness. I could not list all of their names here, but I remember all the time spent with them. Special thanks to Huo Peng, Shi Min, Haifeng, Shanshan, Yezi, Patcharee, Linda, Yuehong, Anne and Nor, for your friendship and help!

Thank my former officemates, Andreas and Tord, who helped me to get familiar with the local environment and made the working time not boring.

To my dearest parents, just saying 'thank you' cannot express my feeling. You are always behind me and unconditionally support me no matter where I am and what I am doing. I can complete my PhD study because of your love and understanding. Thank my sister, Min. Now we have more common topics to talk since we finally reached the same goal.

In the end, to my husband, Martin. When I face difficulties and feel frustrated, you always hold my hands and smile to me. You are my best friend and deeply understand my mind. This work records too many important memories in our lives.

Jing Xie
Gløshaugen, January 2011

Abbreviations

ACK	Acknowledgement
ATM	Asynchronous Transfer Mode
BEB	Binary Exponential Backoff
CCDF	Complementary Cumulative Distribution Function
CIF-Q	Channel-condition Independent Packet Fair Queueing
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CTS	Clear-to-Send
CW	Contention Window
DCF	Distributed Coordination Function
DiffServ	Differentiated Services
DIFS	Distributed Inter Frame Space
EBB	Exponentially Bounded Burstiness
FIFO	First-In-First-Out
FSMC	Finite-State Markov Channel
gSBB	generalized Stochastically Bounded Bursty
GCRA	Generic Cell Rate Algorithm
GPS	Generalized Processor Sharing
GR	Guaranteed Rate
GRC	Guaranteed Rate Clock
HOL	Head-of-Line
IntServ	Integrated Services
IWFQ	Idealized Wireless Fair Queueing
LR	Latency Rate
MAC	Medium Access Control
MGF	Moment Generating Function
PCF	Point Coordination Function
PDF	Probability Density Function

PDV	Packet Delay Variation
PER	Packet Error Rate
PMF	Probability Mass Function
QoS	Quality of Service
RTS	Request-to-send
SAC	Stochastic Arrival Curve
SBB	Stochastically Bounded Burstiness
SIFS	Short Inter-frame Space
StoNC	Stochastic Network Calculus
SSC	Stochastic Service Curve
SSQ	Single Server Queue
STA	Station
WFFQ	Wireless Fluid Fair Queueing
WFQ	Weighted Fair Queueing
WLAN	Wireless Local Area Network

Notations

$a(n)$	Arrival time of the $(n+1)$th packet
$\mathcal{A}$(t)	Cumulative amount of arrival traffic up to time t
$\mathcal{A}^*$(t)	Cumulative amount of departure traffic up to time t
$\mathcal{B}$(t)	Space-domain system backlog at time t
$B(t)$	Time-domain system backlog at time t
$\mathcal{D}$(t)	Space-domain system delay at time t
$d(n)$	Departure time of the $(n+1)$th packet.
$D(n)$	Time-domain system delay of packet $P(n)$
$E(t)$	Cumulative error by time t
$\hat{E}$(t)	Instant error at time t
$\mathcal{F}$	Set of non-negative wide-sense increasing functions
F_X	Cumulative distribution function of random variable X
$\bar{F}_X$	Complementary cumulative distribution function of X
$\bar{\mathcal{F}}$	Set of non-negative wide-sense decreasing functions
$\bar{\mathcal{G}}$	Subset of $\bar{\mathcal{F}}$
$I(t)$	Cumulative amount of impaired service up to time t
$\mathbb{I}(m,n)$	Cumulative impairment in the cumulative service time
L_n	Length of packet $P(n)$
M_X	Moment generating function of random variable X
$P(n)$	The $(n+1)$th packet
$\mathcal{S}$(t)	Cumulative amount of provided service up to time t
$W(n)$	Waiting delay of packet $P(n)$
δ_n	Service time of packet $P(n)$
τ_n	Inter-arrival time between packets $P(n-1)$ and $P(n)$
τ_n^*	Inter-departure time between packets $P(n-1)$ and $P(n)$
ε_n	Error term associated with serving packet $P(n)$
$\Delta(m,n)$	Cumulative service time between two packets
$\Gamma(m,n)$	Cumulative inter-arrival time between two packets

Table of Contents

Chapter 1

Introduction

Computer networks have dramatically evolved over the past several decades and significantly influenced the way of life, communication methods and working methods. Users can share resources of devices connected by communication channels. Compared to the user requirements, the network resources including bandwidth and buffers are scarce. Efficiently utilizing the resources while fulfilling the desired performance metrics at the same time is thus one of the key considerations for network planning and design and for developing new technologies. Performance analysis is required when constructing computer network systems that can fulfill the desired performance metrics. In addition, analyzing the performance of the existing network technologies and the current network systems is helpful for developing new technologies and improving the current systems.

1.1 Performance Analysis

To analyze the performance of computer networks is a challenging task. It requires an intimate knowledge of the network system which is analyzed, and a careful selection of the methodology and tools [59]. Among various performance analysis techniques, this thesis focuses on the analytical modeling technique.

Analytical modeling abstracts the features of a computer network system as a set of parameters or parameterized functions in order to make the modeling task tractable [69]. A computer network system is mathematically described so that certain information about the system behavior can be yielded [89], such as establishing some system equations. Network performance is then derived by solving these equations [97].

Modeling provides a framework for gathering, organizing, evaluating, and understanding information about a system [77]. In order to capture the essential characteristics of the analyzed systems but exclude extraneous information, proper assumptions and hypotheses are necessary for building models. Considering that computer networks behave non-deterministically, statistical models are needed to represent the random events happening in computer networks, such as the randomly generated network traffic and the time-varying service delivered to a traffic flow. Most assumptions and hypotheses therefore imply the underlying stochastic nature.

1.1. Performance Analysis

Statistical modeling tools such as variable distributions, queueing models and Markov models are commonly used for characterizing the behavior of computer networks. Stochastic models are first set up, and the performance metrics are then determined. The mostly concerned performance metrics include:

- **Throughput** refers to the average rate of successful data or message delivery over a communication link or system. It is usually measured in bits per second (bit/s or bps).
- **Latency** refers to the time delay experienced in a system. The definition may vary depending on the system. It is usually measured in millisecond (ms).
- **Delay** in a general sense refers to a lapse of time. It is usually measured in millisecond (ms).
- **Packet delay variation** (PDV) refers to the difference in end-to-end delay between selected packets in a flow with any lost packets being ignored [38].
- **Bandwidth metric** contains four sub-metrics as listed below:
  - Bandwidth capacity
  - Achievable bandwidth
  - Available bandwidth
  - Bandwidth utilization

Queueing phenomena are very common in computer network systems, where the various computers or devices can be modeled as individual queues. The whole system itself can be modeled as a queueing network providing the required service to the traffic that needs to be transmitted. In queueing systems, the shared resources are called servers and the customers arriving at a queue may be messages and/or packets.

In order to conduct the performance analysis for a network system, properly defined traffic models and service models are needed. The traffic model characterizes the traffic arrival process. The well-known arrival processes having been widely applied to the analysis of queueing systems include processes with exponential inter-arrival time distribution, Erlang-k inter-arrival time distribution or deterministic inter-arrival time distribution. The service model describes the

service process, particularly, the service time distribution. The above mentioned distributions are applicable for the service time distribution as well.

Having constructed the queueing model for a computer network system, we can analyze the system mathematically. For example, a number of performance metrics can be derived [3]:

- The distribution of the waiting time and the system time[1] of a customer. The system time is the waiting time plus the service time.
- The distribution of the number of customers in the system (including or excluding the one or those in service).
- The distribution of the amount of the work in the system. That is the sum of service times of the waiting customers and the residual service time of the customer in service.
- The distribution of the busy period of the server. This is a period of time during which the server is working continuously.

A queueing system is a stochastic system, yet its time-dependent or transient behavior is difficult to analyze. *Statistical equilibrium* is a significant role in the analysis of stochastic systems. This represents a state of stochastic processes, the behavior of which is independent of time and the initial state [12]. The *equilibrium or limiting behavior* appears to be much easier to analyze. Under certain conditions, the system of interest is assumed to enter a steady state or a state of equilibrium after enough time has elapsed. The concerned performance metrics have the limiting distributions as time goes to infinity. These distributions are independent of the initial condition of the system.

Having derived the limiting distribution of the concerned performance metrics, the expected values of these metrics are of interest as well:

- the mean waiting time;
- the mean system time;
- the mean number of customers in the queue;

[1]It is also called sojourn time in the queueing theory literature.

- the mean number of customers in the system;
- the probability that the system is in a particular state.

A common situation is that the limiting distribution of the number of customers in the system is firstly derived. To further derive the expected value of the other metrics, the Little's Theorem plays a pivotal role. Under the steady state assumption, Little's Theorem reveals the relationship between the number of customers in a system and the time that a customer spends in this system under a given arrival rate.

For several decades, extensive efforts have been devoted to analysis of queueing systems. Significant results in the classical queueing theory have been widely used to study queueing type problems in computer networks. For example, the Lindely equation is fundamental and yet elementary for computing the queue length at an arbitrary instant for a queueing system, in which the arrival process is characterized by a general inter-arrival time distribution and the service process is described by a general distribution. Moreover, deriving the moments of the distributions of interest is a common way to find and describe the behavior of queueing systems.

As diverse network-based applications and services emerge continuously, the existing queueing models are sometimes difficult to capture the unique customer and service characteristics and requirements in modern packet-switched computer networks. To analyze the complicated queueing systems, such as integrated services networks, needs to take into account the correlations among two or more stochastic processes. The study of such queueing systems has become more challenging because it needs both a good understanding of the analyzed network system and a deep knowledge of mathematics. The engineering specialists may lack strong mathematical skills and the mathematicians may not thoroughly understand the network systems. To shorten the gap between engineering and applied mathematics gives rise to the need of developing new analytical theories for performance analysis of computer networks.

Network Calculus is one of the new analytical theories. It was introduced in early 1990s to deal with the performance analysis issues in modern packet-switched computer networks. Alternate algebras such as the min-plus algebra and the max-plus algebra are used to transform complex non-linear network system into analytically tractable linear systems.

Central to network calculus is properly defining the traffic model and service model. The traffic model called *arrival curve* in network calculus describes the traffic characterization and was initially developed from the (σ, ρ)-traffic characterization [33] [34]. The service model called *service curve* in network calculus characterizes the service behavior of a network element and was originated from the service characterization of Generalized Processor Sharing (GPS) [86] [87]. The core concept of model definitions in network calculus is to find a bound on the cumulative traffic or service. One representative class of traffic models is *envelope processes* which are comprehensively reviewed in [84]. The state-of-the-art of the service models is nicely summarized in [47].

In order to facilitate performance analysis, network calculus has explored some properties, including the analysis of single-node service guarantees, extending the single-node analysis to a sequence of nodes, the aggregate flow analysis and the per-flow analysis.

Unlike the conventional queueing theory which focuses on the quantities in an equilibrium state, network calculus focuses on the analysis of either deterministic or stochastic bounds on performance metrics. Correspondingly, network calculus has been developed along two tracks - *deterministic* and *stochastic*.

Deterministic network calculus deals with deterministic queueing systems in computer networks and is based on worst-case scenario analysis. Although the deterministic service guarantee provides the highest *Quality of Service* (QoS) level that can be provided, a significant portion of network resources is unused on average. However, multimedia applications have gradually dominated data communications. These applications typically have diverse requirements on the service provided by computer networks and most of them can tolerate a certain amount of violation on the service requirements which are called *stochastic service guarantees*. A stochastic service guarantee allows the QoS objectives specified by a flow to be guaranteed with a probability less than 1 [44]. By allowing some packets to violate their required QoS measures, the stochastic service can exploit the statistical multiplexing gain on network links and hence improve the network utilization. To deal with the stochastic service guarantee issue, stochastic network calculus has attracted much research attention recently.

Excellent books [15] [21] not only summarize the significant progress of deterministic network calculus but also provide the details of the rel-

evant mathematical knowledge and application examples. In addition, the theory of effective bandwidth and the pioneer work on stochastic network calculus are covered [21]. A systematic review of stochastic network calculus is available [67].

Most available research effort on service characterization in both deterministic and stochastic network calculus is considered as generalized from the Latency Rate (LR) server which is a general model for analysis of scheduling algorithms [95]. Essentially, LR models the service process using the amount of service delivered by the server in a time period. The server behavior is characterized from the spatial perspective. Another general server model, called Guaranteed Rate (GR) server, has also been investigated for deterministic service guarantees [21]. The GR model captures the service characterization by comparing with a virtual time function in the time domain. The time-domain model has also been extended to analyze deterministic service guarantees for aggregate flows [31] [63]. However, to the best of our knowledge, it is unclear whether the virtual time function can be extended to a stochastic version and how to conduct the performance analysis. Similarly, while there is an extensive network calculus literature on performance analysis from the spatial perspective, the study from a temporal network calculus perspective is very limited.

An objective of this thesis is to develop a temporal network calculus which can be applied to model and analyze networks with stochastic service provision. The trend toward supporting multimedia services in wireless networks invokes research on analytically evaluating wireless network performance. *Another objective* of this thesis is to apply the proposed temporal network calculus to model and analyze the IEEE 802.11 network and the error-prone wireless channel.

1.2 Focus of This Thesis

We focus on analyzing the queueing behavior in computer networks from a temporal network calculus perspective. The temporal behavior of arrivals is described by the *cumulative* inter-arrival time. The service provided to arrivals is quantified using the *cumulative* service time correspondingly. The cumulative time interval can be represented by a stochastic process. The fundamental concept of network calculus is using a bound to characterize the distribution of a stochastic process.

The bound on the cumulative inter-arrival time is called *arrival curve*. The bound on the cumulative service time is called *service curve*.

The bound on a stochastic process is not unique. The tightness of bounds is a trade-off issue since we may face the situation that it is difficult to find a tighter bound. However, a loose bound without any insight on the analyzed system is not appreciated. Finding an optimal bound is thus a consideration of model definition as well.

In order to ease the derivation of performance bounds, we may enforce extra constraints on the arrival and service curves. This may result in the hardness of numerical computation. Thus, we need some transformations between models.

To examine the applicability of the defined time-domain models, we have to apply them to the analysis of real applications.

This thesis studies the above-mentioned problems and introduces some approaches to handle them.

1.3 Research Challenges and Contributions

Performance models capture the behavior characteristics of networks. The behavior of a computer network is often subject to many irregularities and stochastic fluctuations. The reason behind this phenomena is manifold. First of all, the diverse network applications incorporate some complicated dynamics which generate the varying traffic patterns, accordingly. Secondly, in many network systems, the service provided by the networks are non-deterministic. These networks are stochastic in nature. In addition, aggregate multiplexing has been employed extensively in order to improve resource utilization. From the perspective of individual flows, the service received dynamically changes over time because new flows join or existing flows leave.

With the aim to characterize the network behavior with the consideration of the above mentioned issues, we develop a temporal network calculus including generic model definition, property exploration, the concretization of generic models and application study.

1.3.1 Time-domain Modeling and Transformations

In order to capture the temporal behavior of the arrival process and the service process, we develop the time-domain models.

1. **Time-domain Stochastic Arrival Curve**

   The first model is based on a probabilistic extension of a deterministic lower-bound on the cumulative inter-arrival time. However, this basic and simple model without any additional constraint is difficult to be applied for deriving performance bounds. Another model is then introduced to enforce some constraint for characterizing the stochastic behavior of the cumulative inter-arrival time.

2. **Time-domain Stochastic Service Curve**

   The deterministic GR server model [52] is the root of modeling the service behavior in the time-domain. A guaranteed departure time is introduced to be the criteria for evaluating the service provided by a system. A stochastic service curve is defined to represent a bounded probability that the actual departure time is later than the guaranteed departure time. In order to explore the fundamental properties of the time-domain models, a stronger definition with some constraint is introduced.

3. **Transformations between Models**

   Having defined the above arrival curve and service curve models, new questions arise:

   I. What is the guidance of applying the appropriate model?

   II. If the available information abstracted from the system is not sufficient for constructing the appropriate model, can we first construct a model based on the available information and then transform this model to the appropriate one?

   To answer these two questions, we establish the relationships between the basic and the improved models.

Moreover, the available literature of stochastic network calculus mainly focuses on characterizing the spatial behavior of the arrival and the service. Particularly, many models of the arrival process which are the so-called *space-domain* arrival curves in this thesis have been extensively studied, including the $(\sigma(\theta), \rho(\theta))$ stochastic traffic model [21], the effective bandwidth model [42] [72], the exponentially bounded burstiness (EBB) model [102], the stochastically bounded burstiness (SBB) model [94], the generalized stochastically bounded bursty (gSBB) model and two generalized arrival curve models [61] [67]. It is worth investigating the underlying correlation between the time-domain arrival curve and the space-domain arrival curve.

1.3.2 Fundamental Properties

Model construction provides the fundamental elements to performance analysis. Based on the stochastic arrival curve and the stochastic service curve models, several questions of interest need to be answered.

Q1. How to acquire some insight about the behavior of the departure process?

Q2. How to obtain the stochastic performance bounds guaranteed by a system?

Q3. How to analyze the performance of a system consisting of multiple servers in series?

Q4. How to analyze the performance of the aggregate flow?

The above questions rely on four fundamental properties to solve.

P1. **Service Guarantees** provide the probabilistic bounds on backlog and system delay.

P2. **Output Characterization** property shows that the temporal behavior of the departure process can be described using the arrival curve and the service curve.

P3. **Concatenation** property can be used to represent a tandem system with multiple servers as a 'black box'. Then this system can be treated as a single server system when analyzing the system performance.

P4. **Superposition** property can be applied for treating multiple individual flows as a 'single' flow under the First-In-First-Out (FIFO) aggregate service discipline.

1.3.3 Concretization of Generic Models

The generic time-domain models have been defined. The further work is to concretize the generic models with linking some well-known stochastic processes to them and then conducting the performance analysis. In addition, we exemplify the temporal analysis approach by investigating the delay performance of a Gilbert-Elliott channel.

1. **Key technique: Moment Generating Function (MGF)**

   A key technique used in linking an arrival process or service process to the time-domain stochastic arrival curve characterization or stochastic service curve characterization is the MGF.

2. **Error-Prone Wireless Channel Analysis**

   The error-prone nature of wireless channels causes data transmission inherently stochastic and influences the link capacity over time. Thus, the service provided by wireless channels is non-deterministic.

   Gilbert-Elliott channel model [40] [50] is simple while still abstracts the essential properties of the real wireless channel. This channel model can be represented by a two-state homogeneous Markov chain, based on which, the time-domain stochastic service curve of the Gilbert-Elliott channel is obtained. Then given the arrival process characterization, the delay performance can be investigated by applying the service guarantee property (i.e., P1.). Moreover, the delay bound can be improved by taking into account the independence between the arrival process and the service process[2].

   We also compare the delay bounds obtained from the temporal analysis approach with those obtained from the spatial analysis approach. The numerical results show that these two approaches essentially yield very close results.

[2]This conclusion holds for this specific case while may not hold in some other cases.

1.3.4 Application Study

The IEEE 802.11 wireless network is studied to demonstrate how the temporal network calculus is applied to performance analysis.

IEEE 802.11 Medium Access Control (MAC) defines two access methods, the Distributed Coordination Function (DCF) and the Point Coordination Function (PCF). The former is the basic access method and investigated in this thesis. The DCF employs the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) mechanism which allows multiple terminals to share a wireless channel. To alleviate the collisions, the *Binary Exponential Backoff* (BEB) retransmission algorithm is implemented in coordination with the contention-based channel access mechanism.

The additional delay introduced by the collision and the following backoff can be characterized by a delay process. The temporal behavior of the shared wireless channel is thus modeled as a time-domain stochastic server. The crucial step of defining the stochastic service curve is to find an upper bound on the Complementary Cumulative Distribution Function (CCDF) of the cumulative service time. Then for an arrival process which has a stochastic arrival curve, the system delay bound can be readily obtained by applying the service guarantee property (i.e., P1.). Moreover, we also study the system delay bound under finite buffer size[3]. The numerical results using MATLAB are discussed to extensively examine the relevant parameters and provide insight into the obtained analytical bounds.

1.3.5 Service Model with Error Process

Many networks may only provide stochastic service due to some random impairment process. Wireless networks are the most known examples because the error-prone nature of wireless channels causes data transmission error. Such errors happen at the bit-level. In the available stochastic network calculus literature, transmission errors are considered implicitly. The amount of service consumed by transmission errors is simply treated as *impaired service* and deduced in performance analysis [61]. This simple way of treating errors is not sufficient to investigate networks where transmission errors influence the perfor-

[3]This condition is different from the important assumption in stochastic network calculus that assumes an infinite buffer size.

mance and some error handling schemes are adopted to adapt service provision based on the error information. In order to give some insight on how the bit-level error does influence the performance, we propose an error process to explicitly characterize the errors occurring at the bit-level.

1. **Error Characterization**

   We define two stochastic processes to capture the behavior of error occurrence. The snapshot observation of error occurrence records a temporal behavior and is represented by an instant error process. The cumulative number of errors is a spatial quantity and described by a cumulative error process.

2. **Concatenation Property**

   The concatenation property is investigated through analyzing three systems.

   - The study of a *single server system* shows that the instant error introduced by the system remains stochastically unchanged no matter how the error process and the ideal service process are ordered, so does the cumulative error.
   - Analyzing *a system consisting of two error processes in tandem* reveals that both the instant system error process and the cumulative system error process do not change regardless how the two error processes constituting the system are ordered.
   - *A system of multiple servers in tandem* can be viewed as multiple individual processes connected in series. Investigating this system illustrates that both the instant error and the cumulative error introduced by the system are stochastically equal regardless of the order of placing individual processes.

3. **Stochastic Error curve**

   The error process can be considered as a 'virtual error flow' competing the bandwidth with the arrival traffic. A stochastic error curve model is defined to describe the stochastic nature of the error process. The error curve model is proved to hold the concatenation property as well.

4. **Influence of Error Handling on Performance Bounds**

   A simple network is studied to demonstrate how to apply the introduced concepts and to show the influence of error handling on system performance. The approach of re-transmitting the unsuccessfully delivered units causes the transmitted units to delay longer. Another approach to handle the error is discarding the unsuccessfully delivered units and hence degrades the goodput.

1.4 Organization of thesis

This thesis is organized as follows. In Chapter 2 we review mathematical knowledge involved in this thesis and the relevant results of stochastic network calculus. In addition, notations that will be used in the sequel and the specification of system are introduced.

The time-domain model definitions and the model transformations are described in Chapter 3, the material in which has been partially published in Paper B and Paper C.

The four properties (P1. - P4.) are thoroughly investigated in Chapter 4. The relevant discussion reveals the reasons why we establish the transformations between models in Chapter 3. This chapter partially extends Paper B and Paper C.

In Chapter 5, we concretize the time-domain traffic and service models by linking some well-known stochastic processes to them. In addition, we exemplify the temporal analysis approach by investigating the delay performance of a Gilbert-Elliott channel. This chapter is based on the content of Paper E.

The detailed analysis of IEEE 802.11 DCF service time, the derivation of analytical bounds and the numerical evaluation are presented in the first part of Chapter 6. This chapter covers the content of Paper D.

In Appendix A, we define the service model with the error process and investigate the concatenation property of this model. The comparison of two error handling approaches are presented in detail as well. The main results have been published in Paper A.

1.5 Included Publications

This thesis is partially based on five peer-reviewed papers, all of which have been published. All papers are written under supervision and in cooperation with Professor Yuming Jiang.

[A]. *Jing Xie* and Yuming Jiang. "An Analysis on Error Servers for Stochastic Network Calculus." In *Proceedings of the 33rd IEEE Conference on Local Computer Networks* (LCN), Montreal, Canada, October 2008. (Regular paper)

[B]. *Jing Xie* and Yuming Jiang. "Stochastic Network Calculus Models under Max-Plus Algebra." In *Proceedings of IEEE Global Telecommunications Conference* (Globecom), Hawaii, USA, Nov. - Dec. 2009.

[C]. *Jing Xie* and Yuming Jiang. "Stochastic Service Guarantee Analysis Based on Time-Domain Models." In *Proceedings of the 17th Annual Meeting of the IEEE/ACM International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication Systems* (MASCOTS), London, UK, September 2009. (Extended paper)

[D]. *Jing Xie* and Yuming Jiang. "A Network Calculus Approach to Delay Evaluation of IEEE 802.11 DCF." In *Proceedings of the 35th IEEE Conference on Local Computer Networks* (LCN), Denver, USA, October 2010. (Regular paper)

[E]. *Jing Xie* and Yuming Jiang. "A Temporal Network Calculus Approach to Service Guarantee Analysis of Stochastic Networks." In *Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools* (ValueTools), Paris, France, May 2011. (Regular paper)

Chapter 2

Network Model and Background

This chapter describes the network model and gives some mathematical preliminaries that are needed for the analysis in the following chapters. A brief overview on stochastic network calculus of particular relevance to this thesis is presented as well. Section 2.1 defines the network model and introduces the notations used throughout this thesis. In Section 2.2, the fundamental operations of the min-plus algebra and the max-plus algebra and the relevant properties of these operations are reviewed. The knowledge of probability and stochastic process used throughout this thesis is given in Section 2.3. We end this chapter by summarizing the significant aspects of stochastic network calculus relevant to this thesis.

2.1 Notations and System Specification

In this thesis, we make the following assumptions unless stated otherwise.

- A packet is considered to be received by a network element when and only when its last bit has arrived to the network element.
- 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.
- Packets arriving to a network element are queued in the buffer and served in the FIFO order. All queues are assumed to be empty at time 0.
- We assume that systems are lossless and provide sufficient buffer space to store all incoming traffic.

In the following subsections, we define various processes to model a network from the spatial perspective and the temporal perspective, respectively.

2.1.1 Space-Domain Notations

A process of characterizing the network spatial behavior is defined to be a function of time t ($t \geq 0$). Particularly, several process are defined below.

- The cumulative amount of traffic arriving to a network element up to time t is represented as the *space-domain arrival process* denoted by $\mathcal{A}(t)$.
- The cumulative amount of traffic departing from the network element up to time t is represented as the *space-domain departure process* denoted by $\mathcal{A}^*(t)$.
- The cumulative amount of service provided by the network element up to time t is represented as the *space-domain service process* denoted by $\mathcal{S}(t)$.
- The cumulative amount of service consumed by some impairment up to time t is represented as the *space-domain impairment process* denoted by $I(t)$.

Assume that all processes are defined on $t \geq 0$ and by convention, have zero value at $t = 0$. All functions are assumed to be left-continuous[1].

For any $0 \leq s \leq t$, let

$$\begin{aligned}
&\mathcal{A}(s,t) \equiv \mathcal{A}(t) - \mathcal{A}(s),\\
&\mathcal{A}^*(s,t) \equiv \mathcal{A}^*(t) - \mathcal{A}^*(s),\\
&\mathcal{S}(s,t) \equiv \mathcal{S}(t) - \mathcal{S}(s),\\
&I(s,t) \equiv I(t) - I(s).
\end{aligned}$$

To differentiate the arrivals from different flows, we use $\mathcal{A}_i(t)$ and $\mathcal{A}_i^*(t)$ to denote the arrival and departure processes of the ith flow, where $i = 1, 2....$

When analyzing the performance of a network system in this thesis, we mainly focus on the **system backlog** and **system delay** which are defined as [15] [21] [35] [67]:

Definition 1. *Let $\mathcal{A}(t)$ and $\mathcal{A}^*(t)$ denote the space-domain arrival process and departure process of a lossless network system, respectively. The system backlog $B(t)$ at time $t \geq 0$ is defined as*

$$\mathcal{B}(t) = \mathcal{A}(t) - \mathcal{A}^*(t). \tag{2.1}$$

[1]Refer to Chapter 1.1 of [15] for the discussion about the left-continuous assumption.

Assume the arrival packets are served according to the FIFO discipline. The system delay of traffic arriving at time $t \geq 0$, $D(t)$, is defined as

$$\mathcal{D}(t) = \inf\{\tau : \mathcal{A}(t) \leq \mathcal{A}^*(t+\tau)\}. \tag{2.2}$$

2.1.2 Time-Domain Notations

A process of characterizing the network temporal behavior is defined to be a function of packet sequence number n ($n \geq 0$). In this thesis, we consider packets arriving to a network system according to some general inter-arrival distribution.

We use $P(n)$, $a(n)$, $d(n)$ and δ_n, to denote the $(n+1)th$ packet entering the system, its arrival time to the system, departure time from the system and the service time provided by the system, respectively, where $n = 0, 1, 2, ...$

The inter-arrival time and inter-departure time between packets $P(n)$ and $P(n+1)$ are denoted by τ_{n+1} and τ^*_{n+1}, respectively. Let $P(0)$ be the initial arrival, $\tau_0 = a(0)$ and $\tau^*_0 = d(0)$. Note that $a(n) = \sum_{k=0}^{n} \tau_k$ and $d(n) = \sum_{k=0}^{n} \tau^*_k$.

- From the temporal perspective, an arrival process counts the cumulative inter-arrival time between two arbitrary packets and is denoted by $\Gamma(m,n) = \sum_{k=m+1}^{n} \tau_k$. Note $\Gamma(n,n) = 0$.

- A service process describes the cumulative service time received between two arbitrary packets and is denoted by $\Delta(m,n) = \sum_{k=m}^{n} \delta_k$. Note that $\Delta(n,n) = \delta_n$.

- A departure process represents the cumulative inter-departure time between two arbitrary packets and is denoted by $\Gamma^*(m,n) = \sum_{k=m+1}^{n} \tau^*_k$. Note that $\Gamma^*(n,n) = 0$.

- The impairment process represents the cumulative impairment in the cumulative service time received between two arbitrary packets and is denoted by $\mathbb{I}(m,n) = \sum_{k=m}^{n} \varepsilon_k$.

All processes are defined on $0 \leq m \leq n$.

In this thesis, both $a(n)$ and $\Gamma(m,n)$ are used interchangeably to represent an arrival process.

To differentiate the packets of different flows, we use $P_i(n)$ to denote the $(n+1)$th packet of the ith flow, where $i = 1, 2, ...$ This subscript is also applicable for other notations, such as $a_i(n)$, $d_i(n)$, $\delta_{i,n}$ and $\tau_{i,n}$.

In the time-domain, the system backlog $B(t)$ and system delay $D(n)$ are defined as follows.

Definition 2. *Let $a(n)$ and $d(n)$ be the time of packet $P(n)$ arriving to a system and that of departing from the system, respectively. Let the departure time of packet $P(n)$ be $d(n) = t$ $(t \geq 0)$. Then the system backlog at time $t \geq 0$ is*

$$B(t) \leq \inf\big\{k \geq 0 : d(n) \leq a(n+k)\big\}. \tag{2.3}$$

The system delay of packet $P(n)$ experienced in the system is

$$D(n) = d(n) - a(n). \tag{2.4}$$

Moreover, the time that packet $P(n)$ has waited in queue is

$$W(n) = D(n) - \delta_n. \tag{2.5}$$

2.1.3 Other notations

In this thesis, the following function sets are often used. Particularly, the set of non-negative wide-sense increasing functions is denoted by $\mathcal{F}$, where for each function $f(\cdot)$, there holds

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

and for any function $f(\cdot) \in \mathcal{F}$, we set $f(x) = 0$ for all $x < 0$.

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

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

and for any function $f(\cdot) \in \mathcal{F}$, we set $f(x) = 1$ for all $x < 0$.

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

$$\bar{\mathcal{G}} = \big\{ f(\cdot) : \forall n \geq 0, (\int_x^\infty dy)^n f(y) \in \bar{\mathcal{G}} \big\}. \tag{2.6}$$

By definition, the processes defined in the space-domain, $\mathcal{A}(t)$, $\mathcal{A}^*(t)$, $S(t)$ and $I(t)$, belong to $\mathcal{F}$. Similarly, the processes defined in the time-domain, $a(n)$ and $d(n)$, belong to $\mathcal{F}$ as well. In addition, negative exponential functions belong to $\bar{\mathcal{F}}$.

For ease of exposition, we adopt

$$[x]^+ \equiv \max[0, x] \quad and \quad [x]_1 \equiv \min[1, x].$$

In addition, the ceiling and floor functions are used in this thesis as well.

- The ceiling function $\lceil x \rceil$ returns the smallest integer not less than x.
- The floor function $\lfloor x \rfloor$ returns the larget integer not greater than x.

2.2 Min-Plus Algebra and Max-Plus Algebra Basics

An essential idea of (stochastic) network calculus is to use alternate algebras particularly the min-plus and max-plus algebras [15] to transform complex non-linear network systems into analytically tractable linear systems [67]. To the best of our knowledge, the existing models and results of stochastic network calculus mainly focus on characterizing network behavior from the spatial perspective and are based on the min-plus algebra that has basic operations particularly suitable for characterizing the cumulative amount of arrival traffic and the cumulative amount of service. Interestingly, analytically modeling network behavior from the temporal perspective heavily relies on the max-plus algebra.

In the following, we review the basics of both min-plus algebra and max-plus algebra used in this thesis.

In the min-plus algebra, the 'addition' operation represents *infimum* or *minimum* when it exists, and the 'multiplication' operation is $+$.

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

- The min-plus convolution of functions $f, g \in \mathcal{F}$, denoted by $\otimes$, is defined as

$$(f \otimes g)(t) = \inf_{0 \le s \le t} \{f(s) + g(t-s)\}$$

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

- The min-plus deconvolution of functions $f, g \in \mathcal{F}$, denoted by $\oslash$, is defined as

$$(f \oslash g)(t) = \sup_{s \ge 0} \{f(s+t) - g(s)\}$$

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

It has been proved that the min-plus convolution operation is associative and commutative [9] [15] [22] [67].

- Associativity: for any $f_1, f_2, f_3 \in \mathcal{F}$, $(f_1 \otimes f_2) \otimes f_3 = f_1 \otimes (f_2 \otimes f_3)$.
- Commutativity: for any $f_1, f_2 \in \mathcal{F}$, $f_1 \otimes f_2 = f_2 \otimes f_1$.

In the max-plus algebra, the 'addition' operation represents *supremum* or *maximum* when it exists, and the 'multiplication' operation is $+$. For functions in the max-plus algebra, the following operations are often used.

- The max-plus convolution of functions $f, g \in \mathcal{F}$, denoted by $\bar{\otimes}$, is defined as

$$(f \bar{\otimes} g)(n) = \sup_{0 \le m \le n} \{f(m) + g(n-m)\}$$

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

- The max-plus deconvolution of functions $f, g \in \mathcal{F}$, denoted by $\bar{\oslash}$, is defined as

$$(f \oslash g)(n) = \inf_{m \geq 0} \{f(n+m) - g(m)\}$$

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

The max-plus convolution is associative and commutative [15].

- Associativity: for any $g_1, g_2, g_3 \in \mathcal{F}$, $(g_1 \bar{\otimes} g_2) \bar{\otimes} g_3 = g_1 \bar{\otimes} (g_2 \bar{\otimes} g_3)$.
- Commutativity: for any $g_1, g_2 \in \mathcal{F}$, $g_1 \bar{\otimes} g_2 = g_2 \bar{\otimes} g_1$.

2.3 Probability and Stochastic Process

2.3.1 Random Variables

For any random variable X,

- its cumulative distribution function (CDF) denoted by $F_X(x) \equiv P\{X \leq x\}$, belongs to $\mathcal{F}$;
- its complementary cumulative distribution function (CCDF) denoted by $\bar{F}_X(x) \equiv P\{X > x\}$, belongs to $\bar{\mathcal{F}}$.

$F_X(x)$ is monotone non-decreasing and right-continuous. In addition, we know

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to +\infty} F_X(x) = 1.$$

The Stieltjes convolution of two functions is often used in this thesis and thus the definition is given below. For two functions $f(x)$ and $g(x)$, their Stieltjes convolution is

$$(f * g)(x) = \int_{-\infty}^{+\infty} f(x-y) dg(y). \tag{2.7}$$

The Stieltjes convolution is commutative when $f(x)$ and $g(x)$ are CDFs.

After expanding the right-hand side of (2.7), we have

$$\begin{aligned}(f * g)(x) &= \int_{-\infty}^{+\infty} f(x-y)dg(y) \\ &= f(x-y)g(y)\Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} g(x-y)df(y),\end{aligned}$$

where the first term on the right-hand side must be zero in order to make the Stieltjes convolution is commutative.

For two independent random variables X and Y, it is well known that

$$F_{X+Y} = F_X * F_Y = \int_{-\infty}^{\infty} F_X(x-y)dF_Y(y)$$

and

$$\bar{F}_{X+Y} = 1 - F_X * F_Y.$$

If X and Y are non-negative, $F_X(x) = 0$ and $F_Y(x) = 0$ for $x < 0$. Suppose there exist $\bar{F}_X(x) \leq f(x)$ and $\bar{F}_Y(x) \leq g(x)$. The following lemma (Lemma 6.1 [67]) introduces a relation between $\bar{F}_{X+Y}$ and $f(x)$ and $g(x)$.

Lemma 1. *Consider non-negative random variables X and Y. Suppose they are independent and $\bar{F}_X(x) \leq f(x)$ and $\bar{F}_Y(x) \leq g(x)$, where $f, g \in \bar{\mathcal{F}}$. Then, for $x \geq 0$, there holds*

$$P\{X + Y > x\} \leq 1 - (\bar{f} * \bar{g})(x) \tag{2.8}$$

where $\bar{f} = 1 - [f(x)]_1$ and $\bar{g} = 1 - [g(x)]_1$.

Let $\mathbf{E}[X]$ denote the *expected value* of a random variable X. Then the *moment generating function* (MGF) of this random variable, denoted by $M_X(\theta)$, is defined as:

$$\begin{aligned}M_X(\theta) &\equiv \mathbf{E}[e^{\theta X}] \\ &= \begin{cases} \sum_x e^{\theta x}\ p_X(x), & \text{X is discrete} \\ \int_{-\infty}^{\infty} e^{\theta x}\ f_X(x)dx, & \text{X is continuous} \end{cases}\end{aligned}$$

where, θ is real variable, $p_X(x)$ and $f_X(x)$ represent the probability density function (PDF) of a discrete random variable X and probability mass function (PMF) of a continuous random variable X, respectively.

The well-known Chernoff bound gives an upper bound on the CCDF of a random variable X:

$$P\{X \geq x\} \leq e^{-\theta x}\mathbf{E}[e^{\theta X}] \tag{2.9}$$

for all $\theta \geq 0$.

In this thesis, we often concern about the sum of multiple of random variables $\{X_i\}$, namely,

$$Y = \sum_{i=1}^{N} X_i,$$

where, if $X_1, ..., X_N$ are independent, it is known that

$$M_Y(\theta) = M_{X_1}(\theta) \cdots M_{X_N}(\theta). \tag{2.10}$$

For $Y = \sum_{i=1}^{N} X_i$, if $X_1, ..., X_N$ are possibly dependent, the following lemma (Lemma 1.5 [67]) is important.

Lemma 2. *For the sum of multiple random variables $Y = \sum_{i=1}^{N} X_i$, no matter whether they are independent or not, there holds for the CCDF of Y,*

$$\bar{F}_Y(y) \leq \bar{F}_{X_1} \otimes \cdots \otimes \bar{F}_{X_N}(y). \tag{2.11}$$

2.3.2 Stochastic Processes

A *stochastic process* $\{X(t), t \in T\}$ is a collection of random variables defined for each t in the *index* set T. When T is countable set, the stochastic process is said to be a *discrete-time* process. If T is an interval of the real line, the stochastic process is said to be a *continuous-time* process.

The CDF of a stochastic process $X(t)$ is defined as for any (allowed) t:

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

The CCDF of the stochastic process $X(t)$ is defined as

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

The MGF of the stochastic process $X(t)$ is defined as

$$\begin{aligned} M_X(\theta(t),t) &\equiv \mathbf{E}[e^{\theta(t)X(t)}] \\ &= \begin{cases} \sum_x e^{\theta(t)x}\ p_X(x,t), & \text{X is discrete} \\ \\ \int_{-\infty}^{\infty} e^{\theta(t)x}\ f_X(x,t)dx, & \text{X is continuous} \end{cases} \end{aligned}$$

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

The *stationary* process is often considered in this thesis. For a stochastic process $\{X(t)\}$ with $F_X(x_{t_1+\tau},...,x_{t_n+\tau})$ representing the CDF of the joint distribution of $\{X(t)\}$ at times $t_1+\tau$,...,$t_n+\tau$, $\{X(t)\}$ is said to be stationary if for all n, τ and t_1,...τ_n, there holds

$$F_X(x_{t_1+\tau},...,x_{t_n+\tau}) = F_X(x_{t_1},...,x_{t_n}).$$

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

A *martingale* is a stochastic process, where, the conditional expected value of an observation at some time t, given all the observations up to some earlier time s, equals the observation at that time s.

Let U_n be a stochastic process. If U_n is a discrete-time process with finite mean, then it is a discrete-time martingale iff there holds for all $n = 1, 2, ...$

$$\mathbf{E}[U_{n+1}|U_1, U_2, ..., U_n] = U_n.$$

The stochastic process U_n is said to be a *supermartingale* iff for all $n = 1, 2, ...$

$$\mathbf{E}[U_{n+1}|U_1, U_2, ..., U_n] \leq U_n.$$

The stochastic process U_n is said to be a *submartingale* iff for all $n = 1, 2, ...$

$$\mathbf{E}[U_{n+1}|U_1, U_2, ..., U_n] \geq U_n.$$

A martingale is a supermartingale and a submartingale as well.

The following lemma (Theorem 3.2 in [39]) presents the Doob's submartingale inequalities which are useful when the supremum operation is involved.

Lemma 3. *If* $\{U_k, 1 \leq k \leq n\}$ *is a submartingale, then for any real number* x*, there holds:*

$$P\Big\{ \sup_{1\leq m\leq n} U_m \geq x \Big\} \leq \frac{\boldsymbol{E}[U_n^+]}{x}, \tag{2.12}$$

$$P\Big\{ \inf_{1\leq m\leq n} U_m \leq x \Big\} \geq \frac{\boldsymbol{E}[U_1] - \boldsymbol{E}[U_n^+]}{x}. \tag{2.13}$$

Lemma 4 presents an inequality of supermartingale and the corresponding proof [65].

Lemma 4. *If* $\{U_k, 1 \leq k \leq n\}$ *is a supermartingale and all* U_k*,* $k = 1, ..., n$*, are non-negative, then for any real number* $x > 0$*, there holds:*

$$P\Big\{ \sup_{1\leq m\leq n} U_m \geq x \Big\} \leq \frac{\boldsymbol{E}[U_1]}{x}. \tag{2.14}$$

Proof. Since $\{U_k, 1 \leq k \leq n\}$ is a supermartingale, it is trivially true that $\{-U_k, 1 \leq k \leq n\}$ is a submartingale. Then, from Eq.(2.13), we obtain

$$\begin{aligned} P\Big\{ \inf_{1\leq m\leq n} (-U_m) \leq -x \Big\} &\leq \frac{\mathbf{E}[-U_1] - \mathbf{E}[(-U_n)^+]}{-x} \\ &= \frac{\mathbf{E}[-U_1]}{-x} = \frac{\mathbf{E}[U_1]}{x} \end{aligned}$$

where we have applied the fact that $(-U_n)^+ = 0$ and hence $\mathbf{E}[(-U_n)^+] = 0$. Then

$$\Big\{ \sup_{1\leq m\leq n} U_m \geq x \Big\} = \Big\{ \inf_{1\leq m\leq n} (-U_m) \leq -x \Big\}$$

from which, the proof is completed. □

Remark. Note that while Lemma 3 holds under more general conditions, Lemma 4 requires that the supermartingale is comprised of non-negative random variables.

2.3.3 Stochastic Ordering

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

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

we say that X is stochastically smaller than Y [90], written as $X \leq_{st} Y$. The same notation applies when X and Y are random vectors.

Similarly, we say stochastic process $X(t)$ is stochastically smaller than $Y(t)$, written as $X(t) \leq_{st} Y(t)$, if for any t and all x, there holds

$$P\{X(t) > x\} \leq P\{Y(t) > x\}.$$

2.4 State of The Art in Stochastic Network Calculus

This section briefly reviews the important background on stochastic network calculus of particular relevance to this thesis. More specifically, the available literature on stochastic network calculus mainly focuses on modeling network behavior and analyzing network performance from the spatial perspective [17] [30] [45] [47] [61] [67] [68] [78] [80] [84]. We call the corresponding models and results *space-domain* models and results in this thesis.

2.4.1 Space-domain Traffic Models

In order to characterize the arrival process of a flow from the spatial perspective, let us consider the amount of traffic generated by this flow in a time interval $(s, t]$, denoted by $\mathcal{A}(s, t)$. In general, the amount of traffic generated by the flow should be limited so that a certain level of QoS for this flow can be guaranteed. A "famous" traffic model of (deterministic) network calculus is Cruz's (σ, ρ)-traffic characterization defined as below [33]:

$$\mathcal{A}(s, t) \leq \rho \cdot (t - s) + \sigma, \tag{2.15}$$

where σ is the burstiness allowed and ρ is an upper bound on the long term average rate of the traffic flow. The right-hand side of

Inequality (2.15) is a simple linear function while an upper bound on the cumulative amount of traffic $\mathcal{A}(s,t)$ could be any non-decreasing, non-negative function of time. Thus, let $\alpha(t)$ denote a deterministic bound on the cumulative amount of the generated traffic as shown in Figure 2.1, where always holds for all $0 \leq s \leq t$,

$$\mathcal{A}(s,t) \leq \alpha(t-s), \qquad \text{(Traffic amount property)}.$$

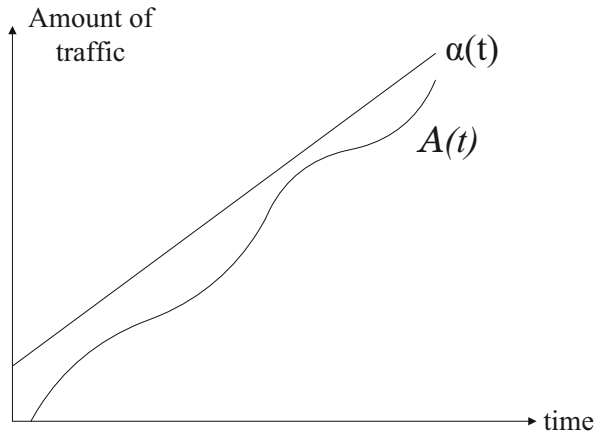

Figure 2.1: The cumulative amount of traffic $A(t)$ is bounded above by $\alpha(t)$

However, the stochastic nature of most network traffic may cause this deterministic relation hardly holding. Alternatively, we may use the probability distribution of $\mathcal{A}(s,t)$ to express the arrival process. The Stochastically bounded burstiness (SBB) traffic model is the probabilistic version of the (σ,ρ)-traffic characterization. The SBB model is defined as follows [94]:

$$P\{\mathcal{A}(s,t) \geq \rho \cdot (t-s) + \sigma\} \leq f(\sigma) \tag{2.16}$$

where ρ is the upper rate and $f(\sigma)$ is the bounding function which is non-increasing and non-negative. Similarly, we can define a traffic model in terms of a general function $\alpha(t)$.

The following traffic model is defined based on the *traffic amount property* and called the *t.a.c* stochastic arrival curve [67].

Definition 3. ***(t.a.c Stochastic Arrival Curve).***

A flow is said to have a traffic-amount-centric (t.a.c) stochastic arrival curve $\alpha(t) \in \mathcal{F}$ with bounding function $f(x) \in \bar{\mathcal{F}}$, if for all $0 \leq s \leq t$ and all $x \geq 0$, there holds

$$P\big\{\mathcal{A}(s,t) - \alpha(t-s) > x\big\} \leq f(x). \tag{2.17}$$

Remark. The left-hand side of Inequality (2.17) represents the violation probability that the actual amount of generated traffic $\mathcal{A}(s,t)$ exceeds the upper bound $\alpha(t-s)$. The right-hand side of Inequality (2.17) gives an upper-bound on the violation probability. The stochastic arrival curve $\alpha(t)$ is an upper bound on $\mathcal{A}(t)$ and not unique. Thus, finding a tighter bound is of concern.

While promising and intuitively simple, Definition 3 has limited use for deriving further results such as delay bound or backlog bound. Thus, another traffic model with more restriction called the *v.b.c* stochastic arrival curve is introduced to facilitate the derivation of performance bounds [67] .

Definition 4. ***(v.b.c Stochastic Arrival Curve).***

A flow is said to have a virtual-backlog-centric (v.b.c) stochastic arrival curve $\alpha(t) \in \mathcal{F}$ with bounding function $f(x) \in \bar{\mathcal{F}}$, if for all $0 \leq s \leq t$ and all $x \geq 0$, there holds

$$P\Big\{\sup_{0\leq s\leq t}\big[\mathcal{A}(s,t) - \alpha(t-s)\big] > x\Big\} \leq f(x). \tag{2.18}$$

Definition 4 solves the difficulty of Definition 3. However, Inequality (2.18) represents a property that may be hard to calculate [100]. A compromise way is to establish a general relation between the *t.a.c* stochastic arrival curve (SAC) and the *v.b.c* SAC. Then we can flexibly use any of them according to the need. The following theorem provides the relation between the *t.a.c* SAC and the *v.b.c* SAC [67].

Theorem 1. *(1) If a flow has a v.b.c SAC $\alpha(t) \in \mathcal{F}$ with bounding function $f(x) \in \bar{\mathcal{F}}$, then the flow has a t.a.c SAC $\alpha(t) \in \mathcal{F}$ with the same bounding function $f(x) \in \bar{\mathcal{F}}$.*

(2) Conversely, if a flow has a t.a.c SAC $\alpha(t) \in \mathcal{F}$ with bounding function $f(x) \in \bar{\mathcal{F}}$, then the flow has a v.b.c SAC $\alpha_\epsilon(t) \in \mathcal{F}$ with bounding function $f^\epsilon(x) \in \bar{\mathcal{G}}$, where

$$\begin{aligned} \alpha_\epsilon(t) &= \alpha(t) + \epsilon \cdot t \\ f^\epsilon(x) &= \left[f(x) + \frac{1}{\epsilon} \int_x^\infty f(y)dy \right]_1 \end{aligned}$$

for any $\epsilon > 0$.

Remark. The bounding function $f^\epsilon(x)$ belongs to $\bar{\mathcal{G}}$ which is a subset of $\bar{\mathcal{F}}$.

2.4.2 Space-domain Service Models

Many network systems may only provide stochastic service, such as wireless networks and multi-access networks. In order to ensure a certain level QoS for an admitted flow, a network system typically guarantees a minimum amount of service to this flow such as the GPS service discipline. In the context of stochastic network calculus, the service curve model is defined based on a stochastic lower bound on the cumulative amount of service provided by the system. In this thesis, the word 'server' is often used interchangeably with 'network system'.

The GPS service discipline [86] provides the basic concept of defining the service curve model. Consider a network node employing the GPS discipline. Let r denote the rate allocated to the arrival process $\mathcal{A}$. Then the departure process $\mathcal{A}^*$ can be expressed by

$$\mathcal{A}^*(t) - \mathcal{A}^*(t_0) \geq r(t - t_0) \tag{2.19}$$

where t_0 is the beginning of the last busy period for the arrival process up to time t. Recall that $\mathcal{A}$ is left-continuous. At time t_0, the backlog is 0, i.e., $\mathcal{A}(t_0) = \mathcal{A}^*(t_0)$. Combining this with Eq.(2.19), we have

$$\begin{aligned} & \mathcal{A}^*(t) - \mathcal{A}(t_0) \geq r(t - t_0) \\ \Rightarrow \quad & \mathcal{A}^*(t) \geq \inf_{0 \leq s \leq t} [\mathcal{A}(s) + r(t - s)] \end{aligned}$$

which can be written as the min-plus convolution:

$$\mathcal{A}^*(t) \geq \mathcal{A} \otimes r(t).$$

The service curve, like the arrival curve, could be any non-decreasing and non-negative function of time. Let $\beta(t)$ denote a deterministic lower bound on the cumulative amount of service up to time t. Then there always holds for all $t \geq 0$,

$$\mathcal{A}^*(t) \geq \mathcal{A} \otimes \beta(t). \tag{2.20}$$

Based on Inequality (2.20), a probabilistic version of the deterministic lower bound is defined below [67].

Definition 5. ***(Weak Stochastic Service Curve).***

A network system is said to provide a weak stochastic service curve $\beta(t) \in \mathcal{F}$ with bounding function $g(x) \in \bar{\mathcal{F}}$, if for all $t \geq 0$ and all $x \geq 0$, there holds

$$P\big\{\mathcal{A} \otimes \beta(t) - \mathcal{A}^*(t) > x\big\} \leq g(x). \tag{2.21}$$

Remark. Similar to the stochastic arrival curve, the stochastic service curve of a network system is not unique.

Definition 5 does not explicitly define the stochastic service curve $\beta(t)$ because Inequality (2.21) couples the arrival process, the service curve and the departure process. Thus, Definition 6 defines a stochastic strict service curve [61] to explicitly describe the relation between the service process and the stochastic service curve.

Definition 6. ***(Stochastic Strict Service Curve).***

A network system is said to provide stochastic strict service curve $\beta(t) \in \mathcal{F}$ with bounding function $g(x) \in \bar{\mathcal{F}}$, if during any period $(s,t]$, the amount of service $\mathcal{S}(s,t)$ provided by this system satisfies, for any $x \geq 0$,

$$P\big\{\mathcal{S}(s,t) < \beta(t-s) - x\big\} \leq g(x). \tag{2.22}$$

The following model [61] defines an important type of stochastic strict server. In such a stochastic server, the stochastic nature of service is due to some impairment process.

Definition 7. ***(Stochastic Strict Service Curve with Impairment).***

Consider a network system providing strict service curve $\hat{\beta}(t) \in \mathcal{F}$ with impairment process $I(t)$. If the impairment process has a stochastic arrival curve $\alpha_I(t) \in \mathcal{F}$ with bounding function $f_I(x) \in \bar{\mathcal{F}}$, which can be either the t.a.c SAC or the v.b.c SAC, then the system provides stochastic strict service curve $\beta(t)$ with bounding function $f_I(x)$, where $\beta(t) = \hat{\beta}(t) - \alpha_I(t)$.

Remark. Here $\hat{\beta}(t)$ characterizes the ideal service process without the impairment process. It is a deterministic lower bound on the cumulative amount of service that the system would have provided if there had been no service impairment.

2.4.3 Five Basic Properties

This section reviews the five basic properties of stochastic network calculus which have been throughly investigated in [67]. These properties can ease tractable network analysis and are explored under the various traffic models and service models reviewed in Section 2.4.1 and Section 2.4.2.

(P1.) Service Guarantees

The service guarantee property means that the performance bounds such as delay bound and backlog bound can be derived under the given traffic model and service model. In order to facilitate the derivation of performance bounds, we introduce two important concepts which are often used in the rest of the thesis.

Definition 8. *Consider two functions $\alpha(t), \beta(t) \in \mathcal{F}$. The maximum horizontal distance between them, denoted by $h(\alpha, \beta)$, is defined as (e.g., [15] [36] [67])*

$$h(\alpha, \beta) = \sup_{t \geq 0} \big\{ \inf\{\tau \geq 0 : \alpha(t) \leq \beta(t + \tau)\} \big\}, \tag{2.23}$$

and the maximum vertical distance between them, denoted by $v(\alpha, \beta)$*, is defined as (e.g., [15] [36] [67])*

$$v(\alpha, \beta) = \sup_{t \geq 0} \big\{ \alpha(t) - \beta(t) \big\} \equiv \alpha \oslash \beta(0). \tag{2.24}$$

Figure 2.2 gives an intuitive explanation of these two concepts using functions $\alpha(t)$ and $\beta(t)$.

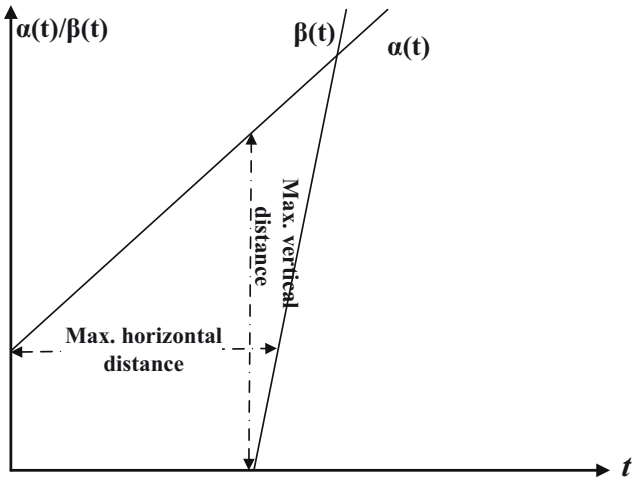

Figure 2.2: Maximum horizontal and vertical distances between two functions

The following theorem [67] gives the backlog bound under the condition that the arrival process has a *v.b.c* SAC and the network system provides a weak stochastic service curve (SSC).

Theorem 2. *Consider a network system providing a weak SSC* $\beta(t) \in \mathcal{F}$ *with bounding function* $g(x) \in \bar{\mathcal{F}}$*. The arrival process has a v.b.c SAC* $\alpha(t) \in \mathcal{F}$ *with bounding function* $f(x) \in \bar{\mathcal{F}}$*. Then for all* $t \geq 0$ *and* $x \geq 0$*, the system backlog* $\mathcal{B}(t)$ *at time* t *is bounded by*

$$P\big\{ \mathcal{B}(t) > x \big\} \leq f \otimes g(x - \alpha \oslash \beta(0)). \tag{2.25}$$

Note that $\alpha \oslash \beta(0)$ denotes the min-plus deconvolution and represents the maximal vertical distance between functions $\alpha(t)$ and $\beta(t)$.

Assume the network system provides a strict service curve[2] $\hat{\beta}(t)$ with impairment process $I(t)$ which has a *v.b.c* SAC $\alpha_I(t)$ with bounding function $f_I(x)$. If the arrival process is independent of the impairment process, Theorem 3 [67] gives the backlog bound by taking into account the independence.

Theorem 3. *Consider a network system providing a strict service curve $\hat{\beta}(t) \in \mathcal{F}$ with impairment process $I(t)$ which has a v.b.c SAC $\alpha_I(t) \in \mathcal{F}$ with bounding function $f_I(x) \in \bar{\mathcal{F}}$. The arrival process $\mathcal{A}(t)$ has a v.b.c SAC $\alpha(t) \in \mathcal{F}$ with bounding function $f(x) \in \bar{\mathcal{F}}$. If $\mathcal{A}(t)$ and $I(t)$ are independent of each other, then for all $t \geq 0$ and $x \geq 0$, the system backlog $\mathcal{B}(t)$ at time t is bounded by*

$$P\big\{\mathcal{B}(t) > x\big\} \leq 1 - \bar{f} * \bar{f}_I\big(x - \sup_{s \geq 0}[\alpha(s) - \beta(s)]\big). \tag{2.26}$$

where $\beta(t) = \hat{\beta}(t) - \alpha_I(t)$, $\bar{f} = 1 - [f(x)]_1$ and $\bar{f}_I = 1 - [f_I(x)]_1$.

Under the same condition as for Theorem 2, we have the following result [67] for the system delay bound.

Theorem 4. *Consider a network system providing a weak SSC $\beta(t) \in \mathcal{F}$ with bounding function $g(x) \in \bar{\mathcal{F}}$. The arrival process has a v.b.c SAC $\alpha(t) \in \mathcal{F}$ with bounding function $f(x) \in \bar{\mathcal{F}}$. Then for all $t \geq 0$ and $x \geq 0$, the system delay $\mathcal{D}(t)$ is bounded by*

$$P\big\{\mathcal{D}(t) > h(\alpha + x, \beta)\big\} \leq f \otimes g(x). \tag{2.27}$$

Note that $h(\alpha + x, \beta)$ represents the maximal horizontal distance between functions $\alpha(t) + x$ and $\beta(t)$.

Under the same assumption as for Theorem 3, the system delay bound is given below [67].

[2]This is a special case of stochastic strict service curve with bounding function $\hat{g}(x) = 0$.

Theorem 5. *Consider a network system providing a strict service curve $\hat{\beta}(t) \in \mathcal{F}$ with impairment process $I(t)$ which has a v.b.c SAC $\alpha_I(t) \in \mathcal{F}$ with bounding function $f_I(x) \in \bar{\mathcal{F}}$. The arrival process $\mathcal{A}(t)$ has a v.b.c SAC $\alpha(t) \in \mathcal{F}$ with bounding function $f(x) \in \bar{\mathcal{F}}$. If $\mathcal{A}(t)$ and $I(t)$ are independent of each other, then for all $t \geq 0$ and $x \geq 0$, the system delay $\mathcal{D}(t)$ is bounded by*

$$P\big\{\mathcal{D}(t) > h(\alpha + x, \beta)\big\} \leq 1 - \bar{f} * \bar{g}(x), \tag{2.28}$$

where $\beta(t) = \hat{\beta}(t) - \alpha_I(t)$, $\bar{f} = 1 - [f(x)]_1$ and $\bar{f}_I = 1 - [f_I(x)]_1$.

(P2.) Output Characterization

In order to analyze the end-to-end performance, we should be able to characterize the traffic behavior after the traffic departs from the previous node. Particularly, the focus is on using the same arrival traffic model to represent the departure traffic. The relevant result is presented below [67].

Theorem 6. *Consider a network system providing a weak SSC $\beta(t) \in \mathcal{F}$ with bounding function $g(x) \in \bar{\mathcal{F}}$. The arrival process has a v.b.c SAC $\alpha(t) \in \mathcal{F}$ with bounding function $f(x) \in \bar{\mathcal{F}}$. Then the departure flow has a t.a.c SAC $\alpha \oslash \beta(t)$ with bounding function $f \otimes g$.*

Note that the arrival flow has a *v.b.c* SAC while the departure flow has a *t.a.c* SAC. However, the *t.a.c* SAC can be transformed into the *v.b.c* SAC according to Theorem 1. Thus, the output characterization can be applied iteratively.

(P3.) Concatenation Property

As shown in the upper sub-figure of Figure 2.3, if a flow traverses a path consisting of n nodes, the end-to-end performance can be obtained by the node-by-node analysis, i.e., analyzing the performance at each node on the path. However, such an approach generally yields looser performance bounds [15] [67]. The concatenation property is thus introduced to represent a series of nodes in tandem as a 'black

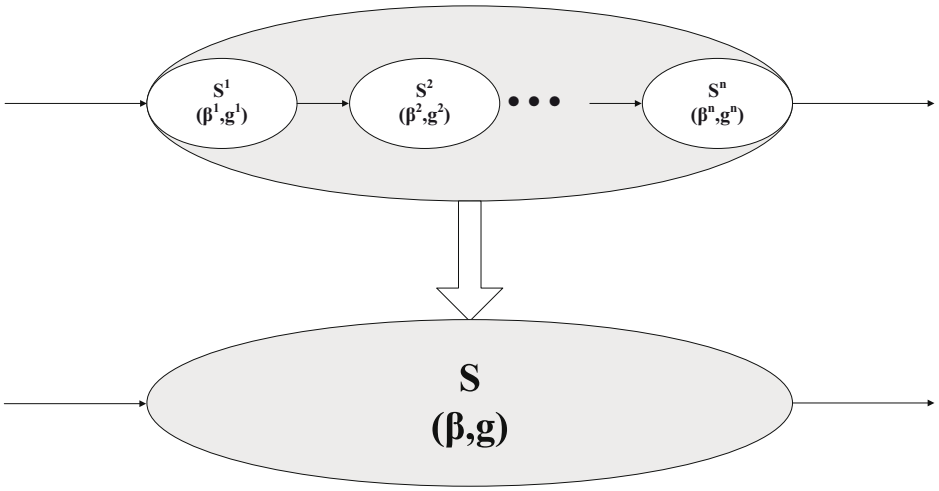

Figure 2.3: Concatenation property of stochastic service curve

box' which can be treated as a single node. This property can facilitate the end-to-end QoS performance analysis and improve results obtained from the node-by-node analysis. Theorem 7 [67] presents one relevant result of the concatenation property.

Theorem 7. *Consider a flow passing through a network of N nodes in tandem. If each node $n(=1,2,...,N)$ provides weak SSC $\beta^n(t) \in \mathcal{F}$ with bounding function $g^n(x) \in \bar{\mathcal{G}}$, then the network guarantees to the arrival process a weak SSC $\beta(t) \in \mathcal{F}$ with bounding function $g(x) \in \bar{\mathcal{G}}$, where*

$$\begin{aligned} \beta(t) &= \beta^1 \otimes \beta^2_{-\epsilon} \otimes \cdots \otimes \beta^N_{-(N-1)\epsilon}(t), \\ g(x) &= g^{1,\epsilon_1} \otimes g^{2,\epsilon_2} \otimes \cdots \otimes g^{N,\epsilon_N}(x) \end{aligned}$$

with

$$\beta^n_{-(n-1)\epsilon}(t) = \beta^n(t) - (n-1)\epsilon \cdot t$$

for $n = 1,...,N$ and $\epsilon > 0$,

$$g^{n,\epsilon_n}(x) = g^n(x) + \frac{1}{\epsilon_n} \int_x^\infty g^n(y)dy$$

for $n = 1, ..., N-1$, *and* $g^{N,\epsilon_N}(x) = g^N(x)$, *for any* $\epsilon_1, ..., \epsilon_{N-1} > 0$.

(P4.) Leftover Service

The leftover service property characterizes the service available to a flow at a server with competing flows. A simple example shown in Figure 2.4, where flows F_1 and F_2 compete for the resource at server S^1 under aggregate scheduling, then flow F_1 passes through server S^2 while flow F_2 leaves. In order to analyze the end-to-end performance of flow F_1, it needs to characterize the service received by flow F_1 at server S^1. From the following theorem [67], we can compute the leftover service provided to the constituent flow.

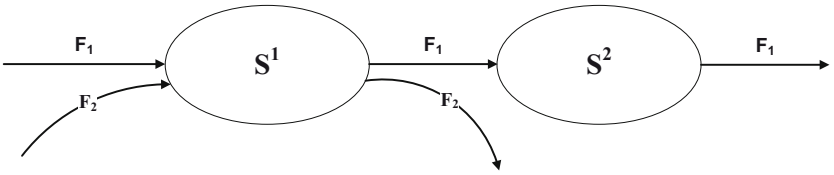

Figure 2.4: Leftover service of flow F_1

Theorem 8. *Consider a system with arrival process* $\mathcal{A}$ *which is the aggregation of two constituent arrival processes* $\mathcal{A}_1$ *and* $\mathcal{A}_2$. *Suppose* $\mathcal{A}_2$ *has a v.b.c SAC* $\alpha_2(t) \in \mathcal{F}$ *with bounding function* $f_2(x) \in \bar{\mathcal{F}}$ *and the system provides to the aggregation arrival process* $\mathcal{A}$ *a weak SSC* $\beta(t) \in \mathcal{F}$ *with bounding function* $g(x) \in \bar{\mathcal{F}}$. *Then if* $\beta(t) - \alpha_2(t) \in \mathcal{F}$, $\mathcal{A}_1$ *receives a weak SSC* $\beta(t) - \alpha_2(t) \in \mathcal{F}$ *with bounding function* $f_2 \otimes g(x) \in \bar{\mathcal{F}}$.

This property is very useful for deriving per-flow performance bounds under aggregate scheduling. When focusing on a specific flow such as F_1 in Figure 2.4, all the other flows can be considered together as an aggregate flow F_2.

(P5.) Superposition Property

The superposition property means that the superposition of multiple flows under the FIFO scheduling can be treated together as a

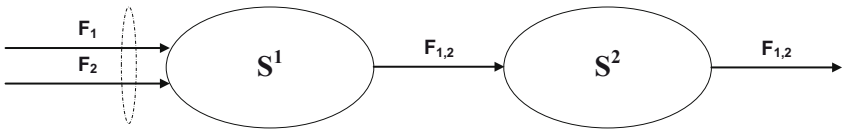

Figure 2.5: Superposition of two constituent flows

single aggregate flow. A simple example shown in Figure 2.5, where two individual flows F_1 and F_2 are aggregated into flow $F_{1,2}$ at server 1. Since two flows are treated equally at both server S^1 and server S^2, from analyzing the end-to-end performance of the aggregate flow $F_{1,2}$, the performance of each individual flow is readily obtained. Theorem 9 [67] provides the result of deriving the arrival process for the aggregate flow.

Theorem 9. *Consider N flows with arrival processes $\mathcal{A}_i$, $i = 1, ..., N$, respectively. Let $\mathcal{A}$ denote the aggregate arrival process. If for all i, the arrival process has a t.a.c (or v.b.c) SAC $\alpha_i(t) \in \mathcal{F}$ with bounding function $f_i(x) \in \bar{\mathcal{F}}$, then the aggregate arrival process has a t.a.c (or v.b.c) SAC $\alpha(t) \in \mathcal{F}$ with bounding function $f(x) \in \bar{\mathcal{F}}$, where*

$$\begin{aligned} \alpha(t) &= \sum_{i=1}^{N} \alpha_i(t), \\ f(x) &= f_1 \otimes \cdots \otimes f_N(x). \end{aligned}$$

Chapter 3

Time-domain Modeling and Transformations

The material in this chapter has been partially published as follows:

- *Jing Xie* and Yuming Jiang. "Stochastic Service Guarantee Analysis Based on Time-domain Models." In *Proceedings of* 17^{th} *Annual Meeting of the IEEE/ACM International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication System (MASCOTS)*, London, UK, September 2009. (Extended paper)
- *Jing Xie* and Yuming Jiang. "Stochastic Network Calculus Models under Max-plus Algebra." In *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM))*, Honolulu, US, December 2009.

3.1 Introduction

Modeling the behavior of queues existing in computer networks is a fundamental issue of network performance analysis. One of related research topics is QoS [49]. Stochastic Network calculus focuses on QoS guarantee analysis [20] [45] [61] [67]. It particularly studies networks where service guarantees are provided stochastically. Such networks include wireless networks, multi-access networks, and multimedia networks where applications can tolerate some certain violation of the desired performance [44].

One open research challenge related to stochastic network calculus is *time-domain modeling and analysis* [67]. Time-domain modeling for service guarantee analysis has its root from the deterministic Guaranteed Rate (GR) server model [52], where the service guarantee is captured by comparing with a (deterministic) virtual time function in the time-domain. This time-domain model has been extended to design aggregate-scheduling networks to support per-flow (deterministic) service guarantees [31] [63], while few such results are available from space-domain models. Other network scenarios where time-domain modeling may be preferable include wireless networks and multi-access networks.

In wireless networks, the varying link condition may cause failed transmission when the link is in 'bad' condition. The sender may hold until the link condition becomes 'good' or re-transmits. For such cases, it is difficult to directly find the stochastic service curve in the space-domain because we need to characterize the stochastic nature of the impaired service caused by the 'bad' link condition. A possible way is that we use an impairment process [61] to characterize the impaired service. However, how to define and find the impairment process arises another difficulty. Even though we can define an impairment process, we need to first convert the impairment process into some existing stochastic network calculus models, and then further analyze the performance bounds. The obtained performance bounds may become loose because of such conversion. If we characterize the service process in the time-domain, we can use random variables to represent the time intervals when the link is in 'bad' condition. Analyzing the stochastic nature of such random variables would be easier. In addition, this way can avoid the difference introduced by the intermediate conversion.

In contention-based multi-access networks, backoff schemes are of-

ten employed to reduce collision occurrence. Because the backoff process is characterized by backoff windows which may vary with the different backoff stages, it is quite cumbersome for a space-domain server model to characterize the service process with the consideration of the backoff process. This also prompts the possibility of characterzing the service process in the time-domain. Having said this, however, how to define a stochastic version of the virtual time function and then perform the corresponding analysis is yet open [67].

In this chapter, we define traffic and service models in the *time-domain*. Particularly, traffic models are defined based on probabilistic lower bounds on the *cumulative packet inter-arrival time*. Service models are defined in terms of the virtual time function and probabilistic upper bounds on the *cumulative packet service time*. Moreover, we establish the transformation between two traffic models or two service models. In order to bridge the gap between the newly defined time-domain models and the existing space-domain models, the transformations between them are established as well.

The chapter is organized as follows. Section 3.2 introduces the preliminary results to be used throughout this chapter. In Section 3.3 and Section 3.4, based on the introduction of the time-domain (deterministic) traffic and service models, we extend them to stochastic versions. In addition, the relationships among them as well as with some existing space-domain models are established.

3.2 Preliminary Results

3.2.1 Distance between Two Functions

The maximum horizontal and vertical distance between two functions have been defined in Definition 8. In the time-domain, we use $\lambda(n)$ instead of $\beta(t)$ and $\gamma(n)$ instead of $\alpha(t)$. Then the maximum horizontal distance between functions $\lambda(n)$ and $\gamma(n)$, denoted by $H(\gamma, \lambda)$, is defined as

$$H(\gamma, \lambda) = \sup_{n \geq 0} \big\{ inf[k \geq 0 : \gamma(n-k) \leq \lambda(n)] \big\}, \tag{3.1}$$

and the maximal vertical distance between them is defined as

$$V(\gamma, \lambda) = \sup_{n \geq 0} \{\gamma(n) - \lambda(n)\} \equiv \gamma \oslash \lambda(0). \tag{3.2}$$

The intuitive illustration of the above two distances is shown in Figure 3.1.

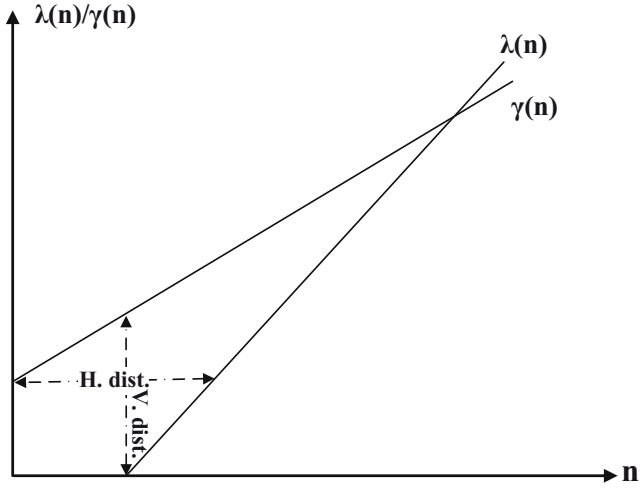

Figure 3.1: Maximum horizontal/vertical distance between two functions

It is worth noticing that in Figure 3.1, the x-axis represents the number of cumulative arrival or departure packets and the y-axis represents the cumulative inter-arrival time or service time when $\lambda(n)$ and $\gamma(n)$ are interpreted as the arrival curve and service curve defined in the time-domain, respectively. While in Figure 2.2, the x-axis represents the discrete time and the y-axis represents the cumulative amount of arrival or service when $\alpha(t)$ and $\beta(t)$ are interpreted as the arrival curve and service curve defined in the space-domain, respectively.

3.2.2 A Fundamental Transformation

Observing the arrivals from the temporal perspective captures the characteristics of the cumulative inter-arrival time. A natural question invoked here is to find the relationship between the cumulative inter-arrival time and the number of cumulative arrivals.

Recall that $\mathcal{A}(t)$ denotes the cumulative amount of arrival traffic up to time t. For ease of exposition, $\mathcal{A}(t)$ denotes **the number of cumulative arrival packets** throughout this chapter unless stated

otherwise. If $\mathcal{A}(t)$ has a deterministic upper-bound $\alpha(t) \in \mathcal{F}$, then the packet arrival time $a(n)$ can be determined according to $\alpha(t)$.

Lemma 5. *For function $\alpha(t) \in \mathcal{F}$, there holds the following relationships:*

1. *the following statements are equivalent:*

   a) for all $0 \leq s \leq t$, $\mathcal{A}(s,t) \leq \alpha(t-s) + x$ for any $x \geq 0$;

   b) for all $t \geq 0$, $\mathcal{A}(t) \leq \mathcal{A} \otimes \alpha(t) + x$ for any $x \geq 0$;

2. *if $\mathcal{A}(t) \leq \mathcal{A} \otimes \alpha(t) + x$ holds for any $t, x \geq 0$, we have $a(n) \geq a\bar{\otimes}\lambda(n) - y$, where $\lambda(n) \in \mathcal{F}$ and y are defined as follows*

$$\lambda(n) = \inf\{\tau : \alpha(\tau) \geq n\}, \tag{3.3}$$

$$y = \sup_{k \geq 0}[\lambda(k) - \lambda(k-x)]. \tag{3.4}$$

Proof. **(1).** For $(a) \rightarrow (b)$, from the condition, we obtain:

$$\mathcal{A}(s,t) - \alpha(t-s) - x \leq 0$$

for any $0 \leq s \leq t$.

Thus there holds

$$\sup_{0 \leq s \leq t}[\mathcal{A}(s,t) - \alpha(t-s) - x] \leq 0$$

which implies

$$\mathcal{A}(t) - \inf_{0 \leq s \leq t}[\mathcal{A}(s) + \alpha(t-s)] - x \leq 0.$$

Then we can conclude for any $t, x \geq 0$,

$$\mathcal{A}(t) \leq \mathcal{A} \otimes \alpha(t) + x.$$

For $(b) \rightarrow (a)$, from the condition, we have

$$\mathcal{A}(t) - \inf_{0 \le s \le t}[\mathcal{A}(s) + \alpha(t-s)] - x \le 0$$

which implies

$$\sup_{0 \le s \le t}[\mathcal{A}(s,t) - \alpha(t-s) - x] \le 0.$$

Then there must hold, for any $0 \le s \le t$,

$$\mathcal{A}(s,t) - \alpha(t-s) - x \le 0.$$

Thus, for any $0 \le s \le t$ and $x \ge 0$, the following holds

$$\mathcal{A}(s,t) \le \alpha(t-s) + x.$$

(2). From (1), we know that

$$\mathcal{A}(t) \le \mathcal{A} \otimes \alpha(t) + x \Leftrightarrow \mathcal{A}(s,t) \le \alpha(t-s) + x$$

for any $0 \le s \le t$ and $x \ge 0$.

Then for any $0 \le m \le n$, we have

$$\mathcal{A}\big(a(m), a_+(n)\big) \le \alpha\big(a_+(n) - a(m)\big) + x$$

where $a_+(n) = a(n) + \epsilon$ with $\epsilon \rightarrow 0$. We also know

$$n - m \le \mathcal{A}\big(a(m), a_+(n)\big) \le \alpha\big(a_+(n) - a(m)\big) + x.$$

Taking the inverse function of $\alpha\big(a_+(n) - a(m)\big)$ yields

$$\begin{aligned}
a_+(n) - a(m) &\ge \lambda(n-m-x) \\
&= \lambda(n-m) - [\lambda(n-m) - \lambda(n-m-x)] \\
&\ge \lambda(n-m) - \sup_{n-m \ge 0}[\lambda(n-m) - \lambda(n-m-x)] \\
&= \lambda(n-m) - \sup_{k \ge 0}[\lambda(k) - \lambda(k-x)] \\
\Longrightarrow a(n) &\ge a(m) + \lambda(n-m) - y,
\end{aligned}$$

which holds because $\epsilon \to 0$ and $y = \sup_{k\geq 0}[\lambda(k) - \lambda(k-x)]$.

Since the above inequality holds for any $0 \leq m \leq n$, we conclude

$$a(n) \geq \sup_{0\leq m\leq n} [a(m) + \lambda(n-m) - y] = a\bar{\otimes}\lambda(n) - y.$$

□

Example 1.

Suppose the number of cumulative arrival packets of a flow, $\mathcal{A}(t)$, is upper-bounded by $\alpha(t) + x$ for $t, x \geq 0$, where $\alpha(t) = \rho \cdot t$. Let $\alpha(t) \equiv n$. We get the inverse function of $\alpha(t)$, $\lambda(n) = \frac{n}{\rho}$. Inserting λ into Eq.(3.4) results in

$$y = \sup_{k\geq 0}\left\{\frac{k}{\rho} - \frac{(k-x)^+}{\rho}\right\} \quad = \begin{cases} \frac{x}{\rho} & \mathrm{k} \geq x, \\ < \frac{x}{\rho} & 0 \leq \mathrm{k} < x, \end{cases}$$

from which we get $y = \frac{x}{\rho}$. Then, for any packet, its arrival time satisfies

$$a(n) \geq \sup_{0\leq m\leq n}\left[a(m) + \frac{n-m}{\rho}\right] - \frac{x}{\rho}.$$

From the view of the packet arrival time, if $a(n)$ has a deterministic lower-bound $\lambda(n) \in \mathcal{F}$, the following lemma can be used to compute the cumulative number of arrival packets according to λ.

Lemma 6. *For function $\lambda(n) \in \mathcal{F}$, there holds:*

1. *the following statements are equivalent:*

   a) *for any $0 \leq m \leq n$, $a(n) - a(m) \geq [\lambda(n-m) - y]^+$ for any $y \geq 0$.*

   b) *for any $n \geq 0$, $a(n) \geq [a\bar{\otimes}\lambda(n) - y]^+$ for $y \geq 0$;*

2. *if $a(n) \geq [a\bar{\otimes}\lambda(n) - y]^+$ holds for any $n, y \geq 0$, we have $\mathcal{A}(t) \leq \mathcal{A} \otimes \alpha(t) + x$, where $\alpha(t) \in \mathcal{F}$ and x are defined as follows*

$$\begin{aligned} \alpha(t) &= sup\{k : \lambda(k) \leq t\} \\ x &= \sup_{u\geq 0}[\alpha(u+y) - \alpha(u) + 1]. \end{aligned} \tag{3.5}$$

Proof. **(1).** The $(a) \to (b)$ part has been proved in Lemma 5(2). We only prove the $(b) \to (a)$ part. From the condition, we know

$$a(n) - \sup_{0 \le m \le n} \{a(m) + \lambda(n-m)\} + y \ge 0$$

$$\Longrightarrow \qquad \inf_{0 \le m \le n} \{a(n) - a(m) - \lambda(n-m)\} + y \ge 0.$$

Thus there holds

$$a(n) - a(m) \ge [\lambda(n-m) - y]^+$$

for any $0 \le m \le n$ and $y \ge 0$.

(2) For any $0 \le s \le t$, let $\mathcal{A}(s) = m$ and $\mathcal{A}(t) = n$, where m and n can be obtained by

$$m = \sup\{k : a(k) \le s\},$$

$$n = \sup\{k : a(k) \le t\}.$$

Then we know $\mathcal{A}(s,t) = n - m$ and $a(n) - a(m+1) \le t - s$. Part (1) shows that $a(n) \ge a \bar{\otimes} \lambda(n) - y$ is equivalent to $a(n) - a(m) \ge \lambda(n-m) - y$. Then we have

$$t - s \ge a(n) - a(m+1) \ge \lambda(n-m-1) - y.$$

Taking the inverse function of $\lambda(n-m-1)$ yields $n-m-1 \le \alpha(t-s+y)$. Because $\mathcal{A}(s,t) = n - m$, we have

$$\begin{aligned}
\mathcal{A}(s,t) &\le \alpha(t-s+y) + 1 \\
&= \alpha(t-s) + [\alpha(t-s+y) - \alpha(t-s) + 1] \\
&\le \alpha(t-s) + \sup_{t-s \ge 0} [\alpha(t-s+y) - \alpha(t-s) + 1] \\
&= \alpha(t-s) + \sup_{u \ge 0} [\alpha(u+y) - \alpha(u) + 1] = \alpha(t-s) + x.
\end{aligned}$$

Since $\mathcal{A}(s,t) - \alpha(t-s) - x \leq 0$ holds for any $0 \leq s \leq t$, we have

$$\begin{aligned} & \sup_{0\leq s\leq t} [\mathcal{A}(s,t) - \alpha(t-s) - x] \leq 0 \\ \Longrightarrow \quad & \mathcal{A}(t) - \inf_{0\leq s\leq t} [\mathcal{A}(s) + \alpha(t-s)] - x \leq 0, \end{aligned}$$

from which, we conclude $\mathcal{A}(t) \leq \mathcal{A} \otimes \alpha(t) + x$. □

Remark. Lemma 6(1) reveals that if (a) holds, so does (b) and vice versus. We hence call Lemma 6(1) the *duality principle* of the time-domain arrival process with respect to the lower bound $\lambda(n)$.

3.3 Time-domain Traffic Models

This section first reviews the (deterministic) arrival curve model defined in the time-domain [21]. Then we generalize the deterministic model and define *time-domain* stochastic arrival curve models.

3.3.1 Deterministic Arrival Curve

Consider an arrival process that specifies packets arriving to a system at time $a(n)$. In order to deterministically guarantee a certain level of QoS to this flow, the traffic sent by this flow must be constrained. The (deterministic) network calculus traffic model in the time-domain characterizes packet inter-arrival times using a lower-bound function [23], called *time-domain (deterministic) arrival curve* in this thesis and defined as follows.

Definition 9. ***(Arrival Curve).***

A flow is said to have a (deterministic) arrival curve $\lambda(n) \in \mathcal{F}$, if its arrival process satisfies, for any $m, n \geq 0$ and $\tau \geq 0$,

$$a(m+n) - a(m) \geq \lambda(n) - \tau, \tag{3.6}$$

where $a(m+n) - a(m) = \Gamma(m, m+n)$ represents the inter-arrival time between packets $P(m)$ and $P(m+n)$.

The following example explains the concept of Definition 3.3.1.

Example 2.

The Generic Cell Rate Algorithm (GCRA) [58] with parameter (T, τ) is a parallel algorithm to the Leaky Bucket algorithm and has been used in fixed-length packet networks such as Asynchronous Transfer Mode (ATM) networks.

The GCRA measures cell rate at a specified time scale and assumes that cells will have a minimum interval between them. Here, T denotes the assumed minimum interval between cells and τ denotes the maximum acceptable excursion that quantifies how early cells may arrive with respect to T. It can be verified that if a flow is GCRA(T, τ)-constrained, it has an arrival curve

$$\lambda(n) = \big(T \cdot n - \tau\big)^+.$$

Thus, for any packet that conforms to the traffic contract, its arrival time satisfies $a(n) \geq \lambda(n+1)$[1]. In addition, for any two conformed packets, $P(m)$ and $P(m+n)$, their inter-arrival time satisfies

$$\begin{aligned} & a(m+n) \geq a(m) + n \cdot T - \tau \\ \Rightarrow \quad & a(m+n) - a(m) \geq n \cdot T - \tau. \end{aligned}$$

3.3.2 Inter-arrival-time Stochastic Arrival Curve

Definition 9 defines a (deterministic) arrival curve $\lambda(n)$ which is the lower-bound of the inter-arrival time between two arbitrary packets. However, the real network traffic characterization is complicated and Inequality (3.6) may not hold in general. Considering this, we extend the deterministic bound into a probabilistic bound.

Definition 10. ***(i.a.t Stochastic Arrival Curve).***

A flow is said to have an inter-arrival-time (i.a.t) stochastic arrival curve $\lambda(n) \in \mathcal{F}$ with bounding function $h(x) \in \bar{\mathcal{F}}$, if for any $m, n \geq 0$ and $x \geq 0$, there holds

$$P\Big\{a(m+n) - a(m) < \big[\lambda(n) - x\big]^+\Big\} \leq h(x). \tag{3.7}$$

[1]Recall that $a(n)$ represents the arrival time of the $n+1$th packet.

It is readily proved that Definition 9 is a special case of Definition 10 with $h(x) = 0$ for all $x \geq 0$.

Queueing theory typically characterizes the arrival process in terms of the probability distribution of the inter-arrival time between two consecutive customers:

$$P\{a(n) - a(n-1) = \tau_n \leq x\} = F(x).$$

Comparing $F(x)$ with Inequality (3.7), we notice that Inequality (3.7) gives a more general probability expression of the (cumulative) inter-arrival time. Thus, $F(x)$ is a special case of Inequality (3.7).

Example 3.

Consider a flow with fixed unit packet size. Suppose its packet inter-arrival times follow an exponential distribution with mean $1/\mu$. Then, the packet arrival time has an Erlang distribution with parameter (n, μ) [2]. For any two packets $P(m)$ and $P(m+n)$, their inter-arrival time $\Gamma(m, m+n)$ satisfies, for $x \geq 0$,

$$\begin{aligned} & P\Big\{a(m+n) - a(m) < \frac{n}{\mu} - x\Big\} \\ \leq \; & P\Big\{a(m+n) - a(m) \leq \big[\frac{n}{\mu} - x\big]^+\Big\} \\ = \; & 1 - \sum_{k=0}^{n-1} \frac{e^{-\mu y}(\mu y)^k}{k!} \end{aligned}$$

where $y = \frac{n}{\mu} - x$.

The *i.a.t* SAC is intuitively simple, but it has limited use if no additional constraint is enforced. We study a simple example to understand this problem. Consider that a single node provides the constant service time T to its input which has an *i.a.t* SAC $\lambda(n)$ with bounding function $h(x)$, where $\lambda(n) \geq T \cdot n$ for any $n > 0$. We are interested in the system delay $D(n)$, where, by definition, $D(n) = d(n) - a(n)$. Because the node provides the constant service time T, the service process has a (deterministic) service curve $\gamma(n) = T \cdot n$ which implies the departure time (see Eq.(3.13))

$$d(n) = \sup_{0 \leq m \leq n} [a(m) + T \cdot (n - m + 1)],$$

where $a(m)$ is the beginning of the backlogged period within which packet $P(n)$ is transmitted. Then the waiting delay is

$$\begin{aligned} W(n) &= D(n) - T \\ &= \sup_{0\le m\le n} \big\{a(m) + T\cdot(n-m+1)\big\} - a(n) - T \\ &= \sup_{0\le m\le n} \big\{a(m) + T\cdot(n-m) - a(n)\big\} \\ &\le \sup_{0\le m\le n} \big\{\lambda(n-m) - [a(n) - a(m)]\big\}. \end{aligned} \tag{3.8}$$

From Inequality (3.8), it is difficult to derive more results if no additional constraint is added because we only know

$$P\{\lambda(n-m) - [a(n) - a(m)] > x\} \le h(x)$$

according to Inequality (3.7). When investigating the performance metrics such as delay bound and backlog bound in Section 4.1, we face the similar difficulty.

3.3.3 Virtual-waiting-delay Stochastic Arrival Curve

The previous subsection has stated the difficulty of applying the *i.a.t* SAC to service guarantee analysis. To avoid such difficulty, we introduce another stochastic arrival curve model which is called *virtual-waiting-delay* (*v.w.d*) stochastic arrival curve. Before introducing the definition of this new arrival curve model, we need to know what is the *virtual-waiting-delay* property.

Consider a flow F of which packets arrive to a system at time $a(n)$. Suppose the arrival process of this flow has a (deterministic) arrival curve $\lambda(n)$. We are interested in the following function:

$$\sup_{0\le m\le n} \Big\{\lambda(n-m) - [a(n) - a(m)]\Big\} \tag{3.9}$$

which can be interpreted as follows. Let us consider a virtual single server queue (SSQ) system fed with the same flow F. The SSQ system has infinite buffer space and is initially empty. Suppose the virtual SSQ provides a constant service time $\lambda(1)$ for each packet. The time that packet $P(n)$ departs from the virtual SSQ is (see Definition 12):

$$d(n) = \sup_{0\le m\le n} [a(m) + \lambda(n-m+1)].$$

The waiting delay of $P(n)$ experienced in the virtual SSQ system is

$$\begin{aligned} W(n) &= D(n) - \lambda(1) \\ &= d(n) - a(n) - \lambda(1) \\ &= \sup_{0\le m\le n} [\lambda(n-m) + a(m)] - a(n) \\ &= \sup_{0\le m\le n} \big\{\lambda(n-m) - \Gamma(m,n)]\big\} \end{aligned}$$

where the last step is Eq.(3.9). We thus call Eq.(3.9) the *virtual waiting delay* property of the arrival process $\Gamma(m,n)$.

Recall Figure 3.1, if we replace $\gamma(n)$ by $\lambda(n-m)$ and $\lambda(n)$ by $\Gamma(m,n)$, Eq.(3.2) becomes Eq.(3.9). Thus, Eq.(3.2) can be used to express the system delay with $\Gamma(m,n)$ as the arrival process and $\lambda(n)$ as the service process.

By extending Eq.(3.9) into a probabilistic version, we define the $v.w.d$ stochastic arrival curve in the following.

Definition 11. ***(v.w.d Stochastic Arrival Curve).***

A flow is said to have a virtual-waiting-delay (v.w.d) stochastic arrival curve $\lambda(n) \in \mathcal{F}$ with bounding function $h(x) \in \bar{\mathcal{F}}$, if for any $0 \le m \le n$ and $x \ge 0$, there holds

$$P\Big\{ \sup_{0\le m\le n} \big\{\lambda(n-m) - \big[a(n) - a(m)\big]\big\} > x\Big\} \le h(x). \tag{3.10}$$

Through some manipulation, Inequality (3.10) can be expressed as the max-plus convolution:

$$P\big\{a(n) < a\bar{\otimes}\lambda(n) - x\big\} \le h(x). \tag{3.11}$$

Here, $a\bar{\otimes}\lambda(n)$ can be considered as the expected time that the packet would arrive to the head-of-line (HOL) if the flow has passed through a virtual SSQ with the (deterministic) service curve $\lambda(n)$. The packet is expected to arrive not earlier than the expected HOL time and x is introduced to denote the difference between the expected HOL time and the actual arrival time. The vioaltion probability is bounded by the function $h(x)$ which should decrease as x increases.

Example 4.

Consider a flow with the same fixed packet size. Suppose all packet inter-arrival times are exponentially distributed with mean $\frac{1}{\mu}$. Based on the steady-state PMF of the queue-waiting time for an M/D/1 queue [93], we say that the flow has a *v.w.d* SAC $\lambda(n) = \hbar \cdot n$ with bounding function h^{exp} for $0 < \hbar < \frac{1}{\mu}$. Let $\rho = \mu \cdot \hbar$. We can obtain the bounding function of the probability that the waiting delay $W(n)$ exceeds $x(\geq 0)$

$$h^{exp}(x) = 1 - (1-\rho) \sum_{i=0}^{\lfloor \frac{x}{\hbar} \rfloor} e^{-\mu(i\hbar - x)} \frac{[\mu(i\hbar - x)]^i}{i!}$$

where, $\lfloor y \rfloor$ denotes the floor function.

Lemma 6(1) demonstrates a duality principle for the (deterministic) arrival curves. It is interesting to study whether there exists some similar relationship between the *i.a.t* and the *v.w.d* models.

Theorem 10. *1. If a flow has a v.w.d SAC $\lambda(n) \in \mathcal{F}$ with bounding function $h(x) \in \bar{\mathcal{F}}$, then the flow has an i.a.t SAC $\lambda(n) \in \mathcal{F}$ with the same bounding function $h(x) \in \bar{\mathcal{F}}$.*

2. Conversely, if a flow has an i.a.t SAC $\lambda(n) \in \mathcal{F}$ with bounding function $h(x) \in \bar{\mathcal{G}}$, it also has a v.w.d SAC $\lambda_{-\eta}(n) \in \mathcal{F}$ with bounding function $h_\eta(x) \in \bar{\mathcal{G}}$, where for $\eta > 0$[2]

$$\begin{aligned} \lambda_{-\eta}(n) &= [\lambda(n) - \eta \cdot n]^+, \\ h_\eta(x) &= \Big[h(x) + \frac{1}{\eta} \int_x^\infty h(y)dy\Big]_1. \end{aligned}$$

Proof. The first part follows from that for any $0 \leq m \leq n$, there trivially holds

$$\lambda(n-m) - [a(n) - a(m)] \leq \sup_{0 \leq m \leq n} \big\{\lambda(n-m) - [a(n) - a(m)]\big\}.$$

[2]Note that η should not be greater than $\lim_{n\to\infty} \frac{\lambda(n)}{n}$.

For the second part, there holds

$$\begin{aligned} & P\Big\{ \sup_{0\le m\le n} \big\{\lambda_{-\eta}(n-m) - [a(n)-a(m)]\big\} > x\Big\} \\ \le\ & P\Big\{ \sup_{0\le m\le n} \big\{\lambda_{-\eta}(n-m) - [a(n)-a(m)]\big\}^+ > x\Big\}. \end{aligned}$$

For any $x \ge 0$,

$$\begin{aligned} & P\Big\{\{\lambda(n-m) - \eta\cdot(n-m) - [a(n)-a(m)]\}^+ > x\Big\} \\ =\ & P\Big\{\lambda(n-m) - \eta\cdot(n-m) - [a(n)-a(m)] > x\Big\} \\ =\ & P\Big\{\lambda(n-m) - [a(n)-a(m)] > x + \eta\cdot(n-m)\Big\} \\ \le\ & h\big(x+\eta\cdot(n-m)\big). \end{aligned}$$

Based on the above steps, we have

$$\begin{aligned} & P\Big\{ \sup_{0\le m\le n} \{\lambda_{-\eta}(n-m) - [a(n)-a(m)]\} > x\Big\} \\ \le\ & \sum_{m=0}^{n} P\Big\{\{\lambda_{-\eta}(n-m) - [a(n)-a(m)]\}^+ > x\Big\} \\ \le\ & \sum_{m=0}^{n} h(x+\eta\cdot(n-m)) \\ =\ & \sum_{k=0}^{n} h(x+\eta\cdot k) \\ \le\ & \sum_{k=0}^{\infty} h(x+\eta\cdot k) \\ =\ & h(x) + \sum_{k=1}^{\infty} h(x+\eta\cdot k) \\ \le\ & h(x) + \frac{1}{\eta}\int_x^{\infty} h(y)dy. \end{aligned}$$

The right-hand side of the last inequality still belongs to $\bar{\mathcal{G}}$. The second part follows from the above inequality and the fact that the probability is always not greater than one. □

Remark. In the second part of Theorem 10, $h(x) \in \bar{\mathcal{G}}$ while not $\in \bar{\mathcal{F}}$. If the requirement on the bounding function is relaxed to $h(x) \in \bar{\mathcal{F}}$, the above relationship may not hold in general.

Theorem 10 implies that under the condition that the bounding function is in $\bar{\mathcal{G}}$, the *v.w.d* SAC is as general as the *i.a.t* SAC since if a traffic source can be modeled by an *i.a.t* SAC, it can also be modeled by a *v.w.d* SAC but may be with a more loose bounding function.

It is worth highlighting that the *v.w.d* SAC looks similar as the *v.b.c* SAC (see Definition 4) defined in the space-domain. Since these two models play an important role for performance analysis in their respective domains, we also establish their relationship in the following theorem.

Theorem 11. *1. If a flow has a space-domain v.b.c SAC $\alpha(t) \in \mathcal{F}$ with bounding function $f(x) \in \bar{\mathcal{F}}$, the flow has a time-domain v.w.d SAC $\lambda(n) \in \mathcal{F}$ with bounding function $h(y) \in \bar{\mathcal{F}}$, where*

$$\lambda(n) = \inf\{\tau : \alpha(\tau) \geq n\}, \quad \textit{and} \quad h(y) = f\big(z^{-1}(y)\big)$$

with $z^{-1}(y)$ denoting the inverse function of y, where

$$y = z(x) \equiv \sup_{k \geq 0}\{\lambda(k) - \lambda(k - x)\}.$$

Specifically, if $\lambda(\cdot)$ is sub-additive, $z(x) = \lambda(x)$.

2. Conversely, if a flow has a time-domain v.w.d SAC $\lambda(n) \in \mathcal{F}$ with bounding function $h(y) \in \bar{\mathcal{F}}$, the flow has a space-domain v.b.c SAC $\alpha(t) \in \mathcal{F}$ with bounding function $f(x) \in \bar{\mathcal{F}}$, where

$$\alpha(t) = \sup\{k : \lambda(k) \leq t\}, \quad \textit{and} \quad f(x) = h\big(z^{-1}(x)\big)$$

with $z^{-1}(x)$ denoting the inverse function of x, where

$$x = z(y) \equiv \sup_{\tau \geq 0}\{\alpha(\tau + y) - \alpha(\tau) + 1\}.$$

Specifically, if $\alpha(\cdot)$ is sub-additive[3]*, $z(y) = \alpha(y) + 1$.*

[3] [14] clarifies that $\alpha(t)$ defines a meaningful constraint only if it is subadditive. If $\alpha(t)$ is not subadditive, it can be replaced by its subadditive closure.

Proof. (1) From Lemma 5, we know that for any $t, x \geq 0$, event

$$\{\mathcal{A}(t) \leq \mathcal{A} \otimes \alpha(t) + x\}$$

implies event

$$\{a(n) \geq a\bar{\otimes}\lambda(n) - y\}$$

where y is obtained from Eq.(3.4):

$$y = \sup_{k \geq 0} \big\{\lambda(k) - \lambda(k-x)\big\} \equiv z(x).$$

Thus, there holds

$$\begin{array}{rrcl} & P\big\{\mathcal{A}(t) \leq \mathcal{A} \otimes \alpha(t) + x\big\} & \leq & P\big\{a(n) \geq a\bar{\otimes}\lambda(n) - y\big\} \\ \Longrightarrow & P\big\{a(n) < a\bar{\otimes}\lambda(n) - y\big\} & \leq & P\big\{\mathcal{A}(t) > \mathcal{A} \otimes \alpha(t) + x\big\} \\ & & \leq & f(x), \end{array}$$

where the relationship between y and x is decided by function Eq.(3.4) as shown above. Particularly, if λ is sub-additive, i.e. $\lambda(a+b) \leq \lambda(a) + \lambda(b)$ for any a and b, we then have:

$$\begin{array}{rl} & P\big\{a(n) < a\bar{\otimes}\lambda(n) - \lambda(x)\big\} \\ \leq & P\big\{a(n) < a\bar{\otimes}\lambda(n) - \sup_{k \geq 0}[\lambda(k) - \lambda(k-x)]\big\} \\ \leq & f(x). \end{array}$$

Hence, the first part follows.

(2) From Lemma 6, we know that for any $n, y \geq 0$, event

$$\{a(n) \geq a\bar{\otimes}\lambda(n) - y\}$$

implies event

$$\{\mathcal{A}(t) \leq \mathcal{A} \otimes \alpha(t) + x\}$$

where x is obtained from Eq.(3.5) as:

$$x = \sup_{u \geq 0}\{\alpha(u+y) - \alpha(u) + 1\} \equiv z(y).$$

Thus, there holds

$$\begin{aligned} & P\big\{a(n) \geq a\bar{\otimes}\lambda(n) - y\big\} \;\leq\; P\big\{\mathcal{A}(t) \leq \mathcal{A} \otimes \alpha(t) + x\big\} \\ \Longrightarrow \quad & P\big\{\mathcal{A}(t) > \mathcal{A} \otimes \alpha(t) + x\big\} \;\leq\; P\big\{a(n) < a\bar{\otimes}\lambda(n) - y\big\} \\ & \qquad\qquad\qquad\qquad\quad \leq\; h(y), \end{aligned}$$

where the relationship between x and y is decided by function Eq.(3.5) as shown above. Particularly, if α is sub-additive, we have

$$\begin{aligned} & P\big\{\mathcal{A}(t) > \mathcal{A} \otimes \alpha(t) + \alpha(y) + 1\big\} \\ \leq\; & P\big\{\mathcal{A}(t) > \mathcal{A} \otimes \alpha(t) + \sup_{u \geq 0}[\alpha(u+y) - \alpha(u) + 1]\big\} \\ \leq\; & h(y), \end{aligned}$$

which ends the proof. □

The generalized stochastically bounded burstiness (gSBB) [103] is a special case of the space-domain *v.b.c* SAC. A summarization of some well-known traffic belonging to gSBB is given [67], including both Gaussian self-similar processes [4] [27] [73] [83], such as fractional Brownian motion, and non-Gaussian self-similar processes, such as $\alpha-$stable self-similar process [5] [70], and the $(\sigma(\theta), \rho(\theta))$ stochastic traffic model [19] [21]. With Theorem 11, the following example shows that gSBB can be readily represented using the time-domain *v.w.d* stochastic arrival curve.

Example 5.

If the arrival process of a flow $\mathcal{A}(t)$ can be described by gSBB with upper rate ρ and bounding function $f(x) \in \bar{\mathcal{F}}$, i.e., for any $t, x \geq 0$, there holds

$$P\Big\{\sup_{0 \leq s \leq t}\big\{\mathcal{A}(s,t) - \rho \cdot (t-s)\big\} > x\Big\} \leq f(x),$$

then the process $\mathcal{A}(t)$ has a $v.b.c$ SAC $\alpha(t) = \rho \cdot t$ with the bounding function $f(x)$. With Theorem 11 (1), the arrival process has a $v.w.d$ SAC $\lambda(n) = \frac{n}{\rho}$ which is sub-additive and the bounding function $h(y) = f(\rho \cdot y)$, i.e.,

$$P\Big\{ \sup_{0 \le m \le n} \big\{ \frac{1}{\rho} \cdot (n-m) - [a(n) - a(m)] \big\} > y \Big\} \le f(\rho \cdot y).$$

Remark. Theorem 11 provides a bridge through which we can readily utilize the available results of gSBB traffic. On the other hand, for some traffic types, they may be more suitable for being characterized using the time-domain traffic models rather than using the space-domain traffic models. We thus believe that the transformation between the two domains can facilitate the analysis.

3.4 Time-domain Service Models

Queueing theory characterizes the service process of a system based on the per customer service time. The time-domain service models borrow the similar concept from queueing theory to describe the cumulative service times.

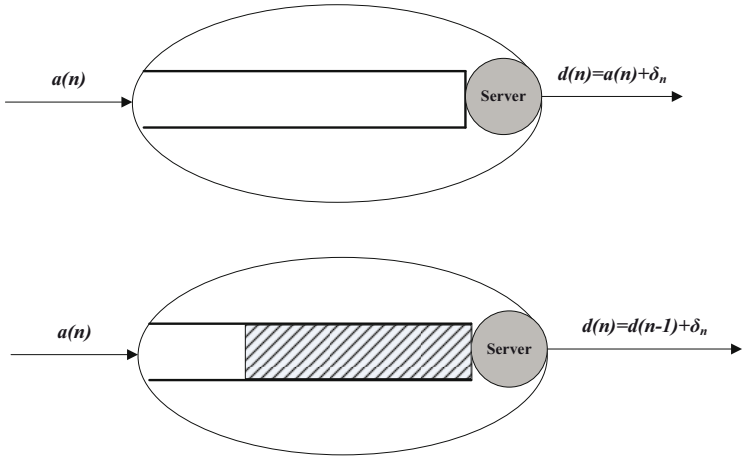

Figure 3.2: Packet departure time

As the upper part of Figure 3.2 shows, if packet $P(n)$ arrives to a system after packet $P(n-1)$ has departed from the system, the

departure time of $P(n)$ is $a(n) + \delta_n$. However, if $P(n)$ arrives to the system and 'see' $P(n-1)$ still in the system, then its departure time will be $d(n-1)+\delta_n$ as shown in the lower part of Figure 3.2. Combining both cases, we have

$$d(n) = \max[a(n), d(n-1)] + \delta_n \tag{3.12}$$

with $d(0) = 0$.

Applying Eq.(3.12) iteratively to its right-hand side results in

$$d(n) = \sup_{0 \le m \le n} \big[a(m) + \sum_{k=m}^{n} \delta_k\big]. \tag{3.13}$$

3.4.1 Deterministic Service Curve

In order to provide deterministic service guarantees to an arrival flow, the system usually allocates a minimum service rate to the flow. A guaranteed minimum service rate is equivalent to a guaranteed maximum service time for each packet of the flow, and accordingly the packet's departure time from the system is bounded. The *Guaranteed Rate Clock* (GRC) is defined based on the guaranteed maximum service time $\hat{\delta}_n$ [52] [53]:

$$GRC(n) = \max[a(n), GRC(n-1)] + \hat{\delta}_n. \tag{3.14}$$

with $GRC(0) = 0$.

The above equation looks very similar to Eq.(3.12). The only difference between Eq.(3.12) and Eq.(3.14) is that $GRC(n)$ represents the guaranteed departure time[4] of packet $P(n)$ while $d(n)$ is the actual departure time of packet $P(n)$.

Plugging the guaranteed maximum service time $\hat{\delta}_n$ into Eq.(3.13) results in the following expression for $GRC(n)$:

$$GRC(n) = \sup_{0 \le m \le n} \big[a(m) + \sum_{k=m}^{n} \hat{\delta}_k\big]. \tag{3.15}$$

[4]The guaranteed departure time is actually $GRC(n)$+error term [52]. However, the underlying service discipline considered throughout this thesis is FIFO, under which, the error term is *zero*.

Suppose we can use a function $\gamma(n-m+1)$ to denote $\sum_{i=m}^{n}\hat{\delta}_i$. Then, Eq.(3.15) becomes

$$\begin{aligned} GRC(n) &= \sup_{0\le m\le n}\left[a(m)+\gamma(n-m+1)\right] \\ &= a\bar{\otimes}\gamma(n) \end{aligned} \tag{3.16}$$

which is the basis for defining the (deterministic) service model [23] as follows. The right-hand side of Eq.(3.16) is the max-plus convolution of the arrival time $a(n)$ and function $\gamma(n)$.

Definition 12. ***(Service Curve).***

Consider a system $\mathcal{S}$ with the arrival process $a(n)$ and the departure process $d(n)$. The system is said to provide to the arrival a (deterministic) service curve $\gamma(n)\in\mathcal{F}$, if for any $n\ge 0$,

$$d(n)\le a\bar{\otimes}\gamma(n). \tag{3.17}$$

Inequality (3.17) illustrates that the actual packet departure time will not be later than the guaranteed departure time. Definition 12 implies that $\gamma(n)$ is an upper bound on the cumulative service time.

The (deterministic) service curve has the following duality principle:

Lemma 7. *For any $n, x\ge 0$, there holds*

$$d(n)-a\bar{\otimes}\gamma(n)\le x,$$

if and only if

$$\sup_{0\le m\le n}[d(n)-a\bar{\otimes}\gamma(n)]\le x$$

for any $n\ge 0$, where $\gamma(n)\in\mathcal{F}$.

Proof. For the "if" part, it trivially holds because

$$d(n)-a\bar{\otimes}\gamma(n)\le\sup_{0\le m\le n}[d(n)-a\bar{\otimes}\gamma(n)].$$

For the "only if" part, from $d(n) - a\bar{\otimes}\gamma(n) \leq x$ for any $n \geq 0$, we have

$$\sup_{0\leq m\leq n}[d(n) - a\bar{\otimes}\gamma(n)] \leq \sup_{0\leq m\leq n}[x] = x.$$

□

The first part of Lemma 7 defines a (deterministic) service curve $\gamma(n) + x$. Lemma 7 states that if a server provides a service curve $\gamma(n) + x$, then $\sup_{0\leq m\leq n}[d(m) - a\bar{\otimes}\gamma(m)] \leq x$ holds, and vice versa. In this sense, we call Lemma 7 the *duality principle* of service curve.

3.4.2 Stochastic Service Curve

For systems that only provide service guarantees stochastically or applications that require only stochastic QoS guarantees, the service time may not be deterministically guaranteed. Accordingly, Inequality (3.17) does not hold in general. Then we extend the (deterministic) service curve into a probabilistic version.

Definition 13. ***(i.d Stochastic Service Curve).***

A system is said to provide an inter-departure time (i.d) stochastic service curve $\gamma(n) \in \mathcal{F}$ *with bounding function* $j(x) \in \bar{\mathcal{F}}$*, if for any* $n, x \geq 0$*, there holds*

$$P\Big\{d(n) - a\bar{\otimes}\gamma(n) > x\Big\} \leq j(x). \tag{3.18}$$

Remark. The stochastic service curve (SSC) of a service process is not unique. It implies that some optimization techniques may be needed when we try to find the SSC of a specific system.

Example 6.

Consider two nodes, the sender and the receiver, communicate through an error-prone wireless link which is modeled as a slotted system. The wireless link can be considered as a stochastic server. Packets have fixed-length and are served in FIFO manner at the sender.

To simplify the analysis, we assume the length of time slot equals one packet transmission time[5].

The sender can send the head-of-queue packet only at the beginning of a time slot. Due to the error-prone characteristics of the wireless link, the probability that a packet can be sent correctly is determined by packet error rate (PER). Here, we assume that packet errors happen independently in every transmission with the fixed PER denoted by P_e. The successful transmission probability of one packet is hence $1 - P_e$. If error happens, the unsuccessfully transmitted packet can be retransmitted in the next time slot immediately. One packet can be retransmitted unlimited times until it is successfully received by the receiver.

The per-packet service time, δ_n, is a geometric random variable with parameter $1 - P_e$. The cumulative service time of sending packets $P(m)$ to $P(n)$ is $\sum_{k=m}^{n} \delta_k$ which follows the negative binomial distribution with parameter $1 - P_e$. The mean per-packet service time is $\bar{\delta} = \frac{1}{1-P_e}$.

According to the CCDF of the negative binomial distribution, the cumulative service time for two arbitrary packets $P(m)$ and $P(m+n)$ is given by

$$\begin{aligned} & P\Big\{ \sum_{k=m}^{m+n} \delta_k > \bar{\delta} \cdot (n+1) + x \Big\} \\ \leq & \sum_{i=\lceil \gamma(n+1)+x \rceil}^{\infty} \binom{i-1}{n} (1-P_e)^{n+1} P_e^{i-(n+1)}, \end{aligned} \tag{3.19}$$

for any $x \geq 0$, where $\lceil \cdot \rceil$ is the ceiling function.

The right-hand side of Inequality (3.19) represents the bound on the probability that the cumulatively actual service time exceeds the cumulative mean service time. From Inequality (3.19), we can obtain the *i.d* SSC provides to the arrival packets. Let $\gamma_\eta(n) = \bar{\delta} \cdot n + \eta \cdot n$ for $\eta > 0$. The right-hand side of Inequality (3.19) is denoted by $j(x)$.

[5]It means we only compute the number of time slots in this example.

According to Definition 13, we know

$$\begin{aligned} & d(n) - a\bar{\otimes}\gamma_\eta(n) \\ = & \sup_{0\le m\le n}\big[a(m) + \sum_{k=m}^{n}\delta_k\big] - \sup_{0\le m\le n}\big[a(m) + (\bar{\delta}+\eta)\cdot(n-m+1)\big] \\ \le & \sup_{0\le m\le n}\big[\sum_{k=m}^{n}\delta_k - \bar{\delta}\cdot(n-m+1) - \eta\cdot(n-m+1)\big], \end{aligned}$$

from which, we have

$$\begin{aligned} & P\Big\{\sup_{0\le m\le n}\big[\sum_{k=m}^{n}\delta_k - \bar{\delta}\cdot(n-m+1) - \eta\cdot(n-m+1)\big] > x\Big\} \\ \le & \sum_{m=0}^{n} P\Big\{\sum_{k=m}^{n}\delta_k - \bar{\delta}\cdot(n-m+1) > x + \eta\cdot(n-m+1)\Big\} \\ \le & \sum_{m=0}^{n} j(x+\eta\cdot(n-m+1)) \\ = & \sum_{k=1}^{n+1} j(x+\eta\cdot k) \\ \le & \Big[\frac{1}{\eta}\int_x^\infty j(y)dy\Big]_1. \end{aligned}$$

Thus, we conclude that this error-prone wireless link provides an $i.d$ SSC $\gamma_\eta(n)$ with the bounding function $j_\eta(x)$ for $\eta > 0$, where

$$\begin{aligned} \gamma_\eta(n) &= \bar{\delta}\cdot n + \eta\cdot n, \\ j_\eta(x) &= \Big[\frac{1}{\eta}\int_x^\infty j(y)dy\Big]_1. \end{aligned}$$

Inequality (3.19) is only relevant to the cumulative service time and does not invovle the arrival process. Thus it provides a way to find the $i.d$ SSC.

Remark. Example 6 illustrates that we can obtain the SSC from analyzing per packet service time. However, if applying the space-domain results for such case, we need an impairment process [67] to characterize the cumulative amount of service consumed by unsuccessful transmissions. In other words, we still need to compute the

cumulative slots due to failed transmission and then convert it into the amount of service. Such conversion may introduce error or result in looser bounds.

In Chapter 4, we show that many results can be derived from the *i.d* SSC. However, without additional constraints, we have difficulty in proving the concatenation property for the *i.d* SSC. To address this difficulty, we introduce another service curve model in the following subsection.

3.4.3 η-Stochastic Service Curve

Lemma 7 reveals the duality principle of the (deterministic) service curve γ, which is provided by a system with input $a(n)$ and output $d(n)$ if and only if for all $n \geq 0$, there holds

$$\sup_{0 \leq m \leq n} \left\{ d(m) - a \bar{\otimes} \gamma(m) \right\} \leq x,$$

which can be generalized to the η-stochastic service curve as defined below.

Definition 14. ***(η-Stochastic Service Curve).***

A system is said to provide an η-stochastic service curve $\gamma(n) \in \mathcal{F}$, *with respect to* η, *with bounding function* $j_\eta(x) \in \bar{\mathcal{F}}$, *if for any* $n, x \geq 0$, *there holds*

$$P\Big\{ \sup_{0 \leq m \leq n} \left[d(m) - a \bar{\otimes} \gamma(m) - \eta \cdot (n-m) \right] > x \Big\} \leq j^\eta(x), \qquad (3.20)$$

for any small $\eta > 0$.

Note that the left-hand side of Inequality (3.20) represents a property that is typically hard to calculate. It means that Definition 14 is more strict than Definition 13. Thus it is important to find the relationship between the *i.d* SSC and the η-stochastic service curves.

Theorem 12. *1. If a system provides to its arrival process an* η*-stochastic service curve* $\gamma(n)$ *with bounding function* $j_\eta(x) \in \bar{\mathcal{F}}$,

it provides to the arrival process an i.d SSC $\gamma(n)$ with the same bounding function $j_\eta(x) \in \bar{\mathcal{F}}$;

2. *If a system provides to its arrival process an i.d SSC $\gamma(n)$ with bounding function $j(x) \in \bar{\mathcal{G}}$, it provides to the arrival process an η-stochastic service curve $\gamma(n)$ with bounding function $j_\eta(x) \in \bar{\mathcal{G}}$ for $\eta > 0$, where*

$$j_\eta(x) = \Big[j(x) + \frac{1}{\eta}\int_x^\infty j(y)dy\Big]_1.$$

Proof. The first part follows since there always holds

$$d(n) - a\bar{\otimes}\gamma(n) \leq \sup_{0\leq m\leq n}\big[d(m) - a\bar{\otimes}\gamma(m) - \eta\cdot(n-m)\big]$$

by letting $m = n$ on the right hand side.

For the second part, there holds

$$\begin{aligned} & \sup_{0\leq m\leq n}\big[d(m) - a\bar{\otimes}\gamma(m) - \eta\cdot(n-m)\big] \\ \leq & \sup_{0\leq m\leq n}\{d(m) - a\bar{\otimes}\gamma(m) - \eta\cdot(n-m)\}^+. \end{aligned}$$

Hence for any $x \geq 0$, there exists

$$\begin{aligned} & P\big\{\sup_{0\leq m\leq n}\{d(m) - a\bar{\otimes}\gamma(m) - \eta\cdot(n-m)\} > x\big\} \\ \leq & \sum_{m=0}^{n} P\big\{d(m) - a\bar{\otimes}\gamma(m) - \eta\cdot(n-m) > x\big\} \\ \leq & \sum_{u=0}^{n} j(x+\eta\cdot u) \\ \leq & \Big[j(x) + \frac{1}{\eta}\int_x^\infty j(y)dy\Big]_1. \end{aligned}$$

The right-hand side of the above inequality still belongs to $\bar{\mathcal{G}}$ and is always not greater than 1. The proof of the second part is completed.

□

In the second part of the above theorem, $j(x) \in \bar{\mathcal{G}}$ while not $\in \bar{\mathcal{F}}$. If the requirement on the bounding function is relaxed to $j(x) \in \bar{\mathcal{F}}$, the above relationship may not hold in general.

3.4.4 Stochastic Strict Service Curve

Definition 13 explores the relationship between the arrival process and the departure process, but it does not explicitly characterize the service process. From Inequality (3.18), it is difficult to directly find the stochastic service curve $\gamma(n)$ for a specific system. Example 6 illustrates this difficulty since we have to add some increment η to the stochastic service curve $\gamma(n)$. To solve this problem, we expand Eq.(3.16) as follows:

$$d(n) - a\bar{\otimes}\gamma(n) = \sup_{0\le m\le n}\big[a(m) + \Delta(m,n)\big] - a\bar{\otimes}\gamma(n). \tag{3.21}$$

Without loss of generality, assume $a(m_0)$ $(0 \le m_0 \le n)$ is the beginning of the backlogged period in which packet $P(n)$ is served. Then,

$$\sup_{0\le m\le n}\big[a(m) + \Delta(m,n)\big] = a(m_0) + \Delta(m_0,n)$$

and $a\bar{\otimes}\gamma(n) \ge a(m_0) + \gamma(n - m_0 + 1)$.

We rewrite the right-hand side of Eq.(3.21) as

$$\begin{aligned} & a(m_0) + \Delta(m_0,n) - a\bar{\otimes}\gamma(n) \\ \le\ & a(m_0) + \Delta(m_0,n) - a(m_0) - \gamma(n - m_0 + 1) \\ =\ & \Delta(m_0,n) - \gamma(n - m_0 + 1). \end{aligned} \tag{3.22}$$

Note that Inequality (3.22) holds for arbitrary $m_0 \le n$. Inspired by this, we define a new service curve model as below.

Definition 15. ***(Stochastic Strict Service Curve).***

A system is said to provide stochastic strict service curve $\gamma(n) \in \mathcal{F}$ with bounding function $j(x) \in \bar{\mathcal{F}}$, if the cumulative service time between two arbitrary packets $P(m)$ and $P(n)$[6] satisfies

$$P\Big\{\Delta(m,n) - \gamma(n - m + 1) > x\Big\} \le j(x) \tag{3.23}$$

[6]If $P(m)$ and $P(n)$ are in the same backlogged period, $\Delta(m,n) = d(n) - d(m-1) = \Gamma^*(m-1,n)$.

for any $x \geq 0$.

Moreover, Inequality (3.22) reveals a relationship between the $i.d$ SSC and the stochastic strict service curve. And from Theorem 12(2), the relationship between the stochastic strict service and the $\eta-$stochastic service curve is indirectly obtained.

Theorem 13. *Consider a system providing stochastic strict service curve* $\gamma(n) \in \mathcal{F}$ *with bounding function* $j(x) \in \bar{\mathcal{F}}$.

1. *It provides an i.d SSC* $\gamma(n)$ *with the same bounding function* $j(x)$.

2. *If* $j(x) \in \bar{\mathcal{G}}$, *it provides an* $\eta-$*stochastic service curve* $\gamma(n)$ *with bounding function* $j_\eta(x) \in \bar{\mathcal{G}}$, *where*

$$j_\eta(x) = \left[j(x) + \frac{1}{\eta} \int_x^\infty j(y) dy \right]_1.$$

3.5 Conclusion

This chapter first defines two traffic models, the $i.a.t$ SAC and the $v.w.d$ SAC. The former is a straightforward extension of the deterministic arrival curve which characterizes the cumulative inter-arrival time. The $i.a.t$ arrival curve model is simple while not applicable for exploring the basic properties, such as service guarantees. The $v.w.d$ arrival curve model is thus introduced to facilitate the property exploration.

The $v.w.d$ arrival curve is defined based on the virtual-waiting-delay property, which is generally hard to compute. In order to flexibly apply these two models, the transformation between them is established. Moreover, the transformation between the time-domain $v.w.d$ arrival curve model and the space-domain $v.b.c$ arrival curve model is established. With this transformation, it is possible to map the available examples of space-domain traffic models into the time-domain traffic models.

The second part of this chapter focuses on defining three service models. The $i.d$ SAC model is the probabilistic extension of the guaranteed rate clock function. This service curve model is simple but has

the same problem as the $i.a.t$ arrival curve model, i.e., inapplicable for exploring the basic properties, such as the concatenation property. The $\eta-$stochastic service curve model is hence defined to deal with this problem. The $\eta-$stochastic service curve is based on the supremum of a given set, which is a more strict condition compared to the condition of finding the $i.d.$ service curve. The $i.d$ stochastic service curve can be transformed into the $\eta-$stochastic service curve with a more lose bounding function.

In addition, the $i.d$ service model couples the service process of a system with both the arrival process and the departure process. Such connection makes it difficult to explicitly characterize the service process. Therefore, the stochastic strict service curve is introduced to help find the $i.d$ stochastic service curve of a system. The corresponding relationships between these three service models are investigated as well.

Chapter 4

Fundamental Properties

The material in this chapter has been partially published as follows:

- *Jing Xie* and Yuming Jiang. "Stochastic Service Guarantee Analysis Based on Time-domain Models." In *Proceedings of* 17^{th} *Annual Meeting of the IEEE/ACM International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication System (MASCOTS)*, London, UK, September 2009. (Extended paper)
- *Jing Xie* and Yuming Jiang. "Stochastic Network Calculus Models under Max-plus Algebra." In *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM))*, Honolulu, US, December 2009.

This chapter presents further results derived from the time-domain traffic and service models defined in Chapter 3. Particularly, we investigate the four basic properties reviewed in Section 2.4.3, including service guarantees, output characterization, concatenation property and superposition property. Some properties can directly be proved only for the combination of a specific traffic model and a specific service model. This clarifies why we established the various transformations between models in Section 3.3 and Section 3.4. With these transformation, we can flexibly apply the corresponding models to characterizing specific systems.

4.1 Service Guarantees

This section investigates the delay bound and backlog bound under the scenario that the arrival process has a *v.w.d* stochastic arrival curve and the service process has an *i.d* stochastic service curve.

4.1.1 Delay Bound

The system delay significantly impacts QoS and is an important performance metric. The following theorem shows that given the stochastic arrival curve and stochastic service curve, we can obtain a bound on system delay readily.

Theorem 14. ***(System Delay Bound).***

Consider a system $\mathcal{S}$ providing an i.d SSC $\gamma(n) \in \mathcal{F}$ with bounding function $j(x) \in \bar{\mathcal{F}}$ to the input which has a v.w.d SAC $\lambda(n) \in \mathcal{F}$ with bounding function $h(x) \in \bar{\mathcal{F}}$. Let $D(n) = d(n) - a(n)$ be the delay in the system of packet $P(n)$. For $x \geq 0$, $D(n)$ is bounded by

$$P\{D(n) > x\} \leq j \otimes h([x - \gamma \oslash \lambda(1)]^+). \tag{4.1}$$

Proof. For any $n \geq 0$, according to the definition of $D(n)$, there holds

$$\begin{aligned} D(n) &= d(n) - a(n) \\ &= \big[d(n) - a\bar{\otimes}\gamma(n)\big] + \big[a\bar{\otimes}\gamma(n) - a(n)\big] \end{aligned}$$

$$
\begin{aligned}
&= \big[d(n) - a\bar{\otimes}\gamma(n)\big] + \sup_{0\le m\le n}\big\{a(m) + \gamma(n-m+1) - a(n)\big\} \\
&= \big[d(n) - a\bar{\otimes}\gamma(n)\big] + \sup_{0\le m\le n}\big\{\lambda(n-m) - [a(n) - a(m)] + \\
&\qquad\qquad \gamma(n-m+1) - \lambda(n-m)\big\} \\
&\le d(n) - a\bar{\otimes}\gamma(n) + \sup_{0\le m\le n}\big\{\lambda(n-m) - [a(n) - a(m)]\big\} \\
&\quad + \sup_{0\le m\le n}\{\gamma(n-m+1) - \lambda(n-m)\} \\
&\le d(n) - a\bar{\otimes}\gamma(n) + \sup_{0\le m\le n}\big\{\lambda(n-m) - [a(n) - a(m)]\big\} \\
&\quad + \sup_{k\ge 0}\{\gamma(k+1) - \lambda(k)\}.
\end{aligned}
$$

To ensure system stability, we require

$$\lim_{k\to\infty}\frac{1}{k}[\gamma(k) - \lambda(k)] \le 0. \tag{4.2}$$

In the rest of the thesis, without explicitly stating, we shall assume inequality Eq.(4.2) holds. In addition, the following results are given

$$P\big\{d(n) - a\bar{\otimes}\gamma(n) > x\big\}$$

and

$$P\Big\{\sup_{0\le m\le n}\big\{\lambda(n-m) - \big[a(n) - a(m)\big]\big\} > x\Big\}.$$

From Lemma 2 and

$$\sup_{k\ge 0}\big\{\gamma(k+1) - \lambda(k)\big\} = \gamma \oslash \lambda(1),$$

we can conclude

$$P\{D(n) > x\} \le j \otimes h(x - \gamma \oslash \lambda(1)).$$

□

If the arrival process and the service process are independent, according to Lemma 1, we can obtain the corresponding bound on the system delay as follows.

Lemma 8. ***(System delay bound: independent condition)***

Consider a system $\mathcal{S}$ providing an i.d SSC $\gamma(n) \in \mathcal{F}$ with bounding function $j(x) \in \bar{\mathcal{F}}$ to the arrival process which has a v.w.d SAC $\lambda(n) \in \mathcal{F}$ with bounding function $h(x) \in \bar{\mathcal{F}}$. Suppose the arrival process and the service process are independent. Then for $x \geq 0$, the system delay $D(n)$ is bounded by

$$P\{D(n) > x\} \leq 1 - \bar{j} * \bar{h}([x - \gamma \oslash \lambda(1)]^+), \tag{4.3}$$

where $\bar{j}(x) = 1 - [j(x)]_1$ and $\bar{h}(x) = 1 - [h(x)]_1$.

Recall Equation (3.13) which expresses the packet departure time $d(n)$ in terms of the beginning of the latest backlogged period $a(m)$ $(m \leq n)$ and the cumulative service time $\sum_{k=m}^{n} \delta_k$. Replacing the packet departure time $d(n)$ with Equation (3.13) results in another expression for the system delay $D(n)$:

$$\begin{aligned} D(n) &= \sup_{0 \leq m \leq n} \Big[a(m) + \sum_{k=m}^{n} \delta_k \Big] - a(n) \\ &= \sup_{0 \leq m \leq n} \big[\sum_{k=m}^{n} \delta_k - \sum_{k=m+1}^{n} \tau_k \big]. \end{aligned} \tag{4.4}$$

Assume $\gamma(n) \leq \lambda(n)$ for any $n > 0$ and $\gamma(n) = \lambda(n) = 0$ for all $n \leq 0$. In addition, $\lambda(n)$ and $\gamma(n)$ are subadditive. Adding $\lambda(n-m+1) - \gamma(n-m+1)$ to the right-hand side of Eq.(4.4) leads to

$$\begin{aligned} D(n) &\leq \sup_{0 \leq m \leq n} \big[\Delta(m,n) - \gamma(n-m+1) + \lambda(n-m+1) - \sum_{k=m+1}^{n} \tau_k \big] \\ &\leq \sup_{0 \leq m \leq n} \big[\Delta(m,n) - \gamma(n-m+1) \big] + \\ &\quad \sup_{0 \leq m \leq n} \big[\lambda(n-m+1) - \Gamma(m,n) \big] \\ &\leq \sup_{0 \leq m \leq n} \big[\Delta(m,n) - \gamma(n-m+1) \big] + \\ &\quad \sup_{0 \leq m \leq n} \big[\lambda(n-m) + \lambda(1) - \Gamma(m,n) \big], \end{aligned}$$

where, the last step is obtained from the condition that $\lambda(n)$ is subadditive. Suppose there exists the following inequalities:

$$P\Big\{ \sup_{0\le m\le n} \big[\Delta(m,n) - \gamma(n-m+1)\big] > x \Big\} \le j(x), \quad (4.5)$$

$$P\Big\{ \sup_{0\le m\le n} \big[\lambda(n-m) - \Gamma(m,n)\big] > x \Big\} \le h(x), \quad (4.6)$$

with which, we obtain another delay bound:

$$P\{D(n) > x\} \le j \otimes h\big(x - \lambda(1)\big). \quad (4.7)$$

Remark. Inequality (4.6) represents a *v.w.d* SAC $\lambda(n)$ with bounding function $h(x)$. From Inequality (4.5), we can prove that $\gamma(n)$ is an *i.d* SSC with the same bounding function $j(x)$.

Proof. Let $d(n)$ be instead of Eq.(3.13). According to Definition 13, we can write

$$\begin{aligned} & d(n) - a\bar{\otimes}\gamma(n) \\ = & \sup_{0\le m\le n} \big[a(m) + \Delta(m,n)\big] - \sup_{0\le m\le n} \big[a(m) + \gamma(n-m+1)\big] \\ \le & \sup_{0\le m\le n} \big[\Delta(m,n) - \gamma(n-m+1)\big]. \end{aligned}$$

From the above inequality and the condition of Inequality (4.5), we know

$$\begin{aligned} & P\big\{d(n) - a\bar{\otimes}\gamma(n) > x\big\} \\ \le & P\Big\{ \sup_{0\le m\le n} \big[\Delta(m,n) - \gamma(n-m+1)\big] > x \Big\} \\ \le & j(x). \end{aligned}$$

Thus, we conclude $\gamma(n)$ is an *i.d* SSC with bounding function $j(x)$. □

The left-hand side of Inequality (4.5) requires finding the supremum of a set, i.e., $P\{\sup_n f_n > x\}$. The right-hand side of Inequality (4.5) is an upper bound on the probability computed by the left-hand side. In order to find this upper bound, we consider a virtual SSQ

with δ_n as packet service times and $\gamma(1)$ as packet inter-arrival times. Then the waiting time of packet $P(n+1)$ in such virtual SSQ is

$$\begin{aligned} W(n+1) &= d(n+1) - a(n+1) - \delta_n \\ &= \sup_{0\le m\le n+1}\Big[a(m) + \sum_{k=m}^{n+1}\delta_k\Big] - a(n+1) - \delta_{n+1} \\ &= \sup_{0\le m\le n+1}\Big[\sum_{k=m}^{n}\delta_k - [a(n+1) - a(m)]\Big] \\ &= \sup_{0\le m\le n}\Big[\sum_{k=m}^{n}\delta_k - \gamma(n-m+1)\Big], \end{aligned}$$

from which, Inequality (4.5) can be interpret as the bound on the CCDF of waiting time distribution in $D/G/1$ queue. If the packet service times are independent, a bound based on martingale can be derived for $GI/GI/1$ queue (e.g. [64]). Fixing n, we define a stochastic process

$$V(l) = e^{\theta\sum_{k=n-l}^{n}[\delta_k - \gamma(1)]}$$

with $\theta > 0$ and $0 \le l < n$. Note that $e^{\theta[\delta_{n-1-(l+1)} - \gamma(1)]}$ is independent of all $e^{\theta[\delta_{n-1-v} - \gamma(1)]}$ for all $v = 0, 1, ..., l$. It can be proved that $\{V(l)\}$, $l = 0, 1, ..., n$, is a supermartingale[1]. In this way, with Doob's inequality for martingale, we have for $\theta > 0$ and $\mathbf{E}[e^{\theta(\delta_0 - \gamma(1))}] \le 1$:

$$\begin{aligned} & P\Big\{\sup_{0\le m\le n}\big[\sum_{k=m}^{n}\delta_k - \gamma(n-m+1)\big] > x\Big\} \\ = \;& P\Big\{e^{\theta\sup_{0\le m\le n}\sum_{k=m}^{n}[\delta_k - \gamma(1)]} > e^{\theta x}\Big\} \\ = \;& P\Big\{\sup_{0\le l<n} V(l) > e^{\theta x}\Big\} \\ \le \;& \mathbf{E}[e^{\theta(\delta_0 - \gamma(1))}]e^{-\theta x}, \end{aligned} \tag{4.8}$$

where $\mathbf{E}[e^{\theta(\delta_0 - \gamma(1))}]$ is the moment generating function of $\delta_0 - \gamma(1)$.

Inequality (4.8) shows that θ is an optimization parameter. The following way is used to get an optimized bound for the waiting time in queue.

[1]Readers refer to the literature, e.g. [64] for the complete proof. Note that we can find similar result of this martingale proof back to Kingman in 1970s.

Lemma 9. *Consider a $D/GI/1$ queue. Assume that $M_{\delta_0-\gamma(1)}$ exists for small $\theta > 0$. Let $\theta^* = \sup\{\theta : M_{\delta_0-\gamma(0)} \leq 1\}$. Then, we have the following bound for waiting time in queue:*

$$P\{W > x\} \leq \inf_{0<\theta\leq\theta^*} M_{\delta_0-\gamma(1))} e^{-\theta x}. \tag{4.9}$$

Following the same approach, Inequality (4.6) also represents a virtual SSQ with $\lambda(1)$ as packet service times and τ_n as packet inter-arrival times. Then Inequality (4.6) can be interpreted as the waiting time in $GI/D/1$ queue. For such virtual system, the bounding function $h(x)$ is

$$h(x) = \inf_{0<\theta\leq\theta^*} \mathbf{E}[e^{\theta(\lambda(1)-\tau_0)}] e^{-\theta x}, \tag{4.10}$$

where $\theta^* = \sup\{\theta : M_{\lambda(1)-\tau_0)} \leq 1\}$.

The following example shows how to obtain a system delay bound based on Inequalities (4.5) and (4.6).

Example 7.

Consider an $E_2/M/1$ system. The packet inter-arrival times follow a 2-stage hypo-exponential distribution with two different arrival rates μ and 2μ. The packet service times are distributed exponentially with mean $\frac{1}{\mu}$.

Then, taking $\gamma(n) = \frac{n}{r_s}$ where $r_s < \mu$, the right-hand side of Inequality (4.8) becomes

$$j(x) = \inf_{0<\theta\leq\theta^*} \frac{\mu}{\mu-\theta} e^{-\frac{\theta}{r_s}} e^{-\theta x},$$

where

$$\theta^* = \sup\{\theta : \frac{\mu}{\mu-\theta} e^{-\frac{\theta}{r_s}} \leq 1\}.$$

Taking $\lambda(n) = \frac{n}{r_a}$ where $r_a > \frac{2\mu}{3}$, the right-hand side of Inequality (4.6) becomes

$$h(x) = \inf_{0<\theta\leq\theta^*} \frac{\mu}{\mu+\theta} \frac{2\mu}{2\mu+\theta} e^{\frac{\theta}{r_a}} e^{-\theta x},$$

where

$$\theta^* = \sup\{\theta : \frac{\mu}{\mu+\theta} \frac{2\mu}{2\mu+\theta} e^{\frac{\theta}{r_a}} \leq 1\}.$$

Considering the arrival process and the service process are independent, applying Lemma 1 may yield a tighter system delay bound compared with the bound given in Inequality (4.7). Note that we use

$$j(x) = \frac{\mu}{\mu - \theta} e^{-\frac{\theta}{r_s}} e^{-\theta x}$$

and

$$h(x) = \frac{\mu}{\mu + \theta} \frac{2\mu}{2\mu + \theta} e^{\frac{\theta}{r_a}} e^{-\theta x}$$

to implement Stieltjes convolution when applying Inequality (2.7) here. Then, we have the following bound on the system delay

$$\begin{aligned} & P\{D > x\} \\ \leq\ & 1 - \big((1-j) * (1-h)\big)(x) \\ =\ & \inf_{0<\theta\leq\theta^*} 1 - \frac{\mu}{\mu-\theta} e^{-\frac{\theta}{r_s}} \big(1 - e^{-\theta x^*} - \frac{2\theta\mu^2 x^*}{(2\mu+\theta)(\mu+\theta)} e^{\theta(\frac{1}{r_a} - x^*)}\big) \end{aligned} \quad (4.11)$$

where $x^* = x - \frac{1}{r_a}$.

When computing the bound based on Inequality (4.11), r_s and r_a are confined by $\frac{2\mu}{3} < r_a \leq r_s < \mu$. For the allowed combinations of r_s and r_a, their corresponding θ^* are listed in Table 4.1.

r_a	0.7μ	0.7μ	0.7μ	0.8μ	0.8μ	0.9μ
r_s	0.7μ	0.8μ	0.9μ	0.8μ	0.9μ	0.9μ
θ^*	0.1225μ	0.1225μ	0.1225μ	0.3713μ	0.1931μ	0.1931μ

Table 4.1: Combination of r_a and r_s vs. θ_0

The right-hand side of Inequality (4.11) is influenced by r_a, r_s and θ^*. To study the influence of these three parameters, we compare the bound with the exact CCDF of system delay. First, we know the CDF of the waiting time in the G/M/1 queue [74]:

$$\begin{aligned} F_w(x) &= P\{W \leq x\} \\ &= 1 - (2 - \sqrt{2}) e^{-\mu(\sqrt{2}-1)x}. \end{aligned}$$

The system delay equals the waiting time in queue plus the service time, then the CCDF of system delay is expressed as[2]

$$\begin{aligned} P\{D > x\} &= P\{W + \delta > x\} \\ &= 1 - F_w * F_{exp}(x) \end{aligned} \quad (4.12)$$

[2]The waiting time in queue and the service time are independent.

where $F_{exp}(x) = 1 - e^{-\mu x}$ is the CDF of the exponential distribution. Thus we get the CCDF of the system delay in the $G/M/1$ system:

$$P\{D > x\} = e^{-\mu(\sqrt{2}-1)x}. \tag{4.13}$$

Let $\mu = 1$. Figure 4.1 shows that the curve obtained with $r_a = 0.8\mu, r_s = 0.8\mu$ is the most tight bound. Revisit Table 4.1, the corresponding θ^* is 0.3713μ. The two curves obtained with $r_a = 0.8\mu, r_s = 0.9\mu$ and $r_a = 0.9\mu, r_s = 0.9\mu$ are very close. The corresponding θ^* for these two combinations are the same, 0.1931μ. Similarly, if $r_a = 0.7\mu$, the three combinations have the same $\theta^* = 0.1225\mu$. The corresponding curves are also very close. It reveals that the curve becomes tighter as θ^* increases.

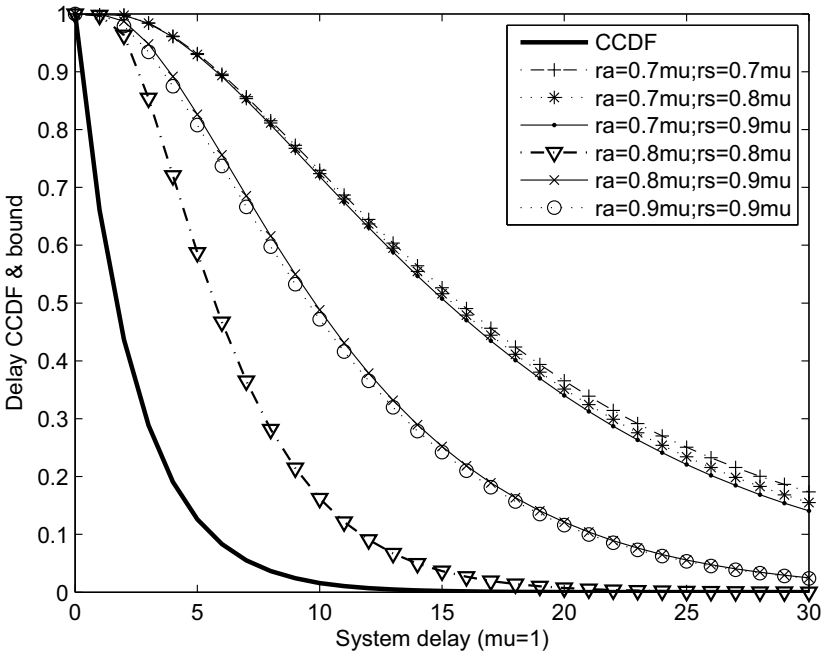

Figure 4.1: CCDF and bound of E_2/M/1 system delay

4.1.2 Backlog Bound

When investigating the backlog in the system at time t, we first recall the definition of the system backlog:

$$\mathcal{B}(t) = \mathcal{A}(t) - \mathcal{A}^*(t),$$

where $\mathcal{A}(t)$ and $\mathcal{A}^*(t)$ represent the cumulative number of arrival packets and the cumulative number of departure packets up to time t, respectively.

Let the departure time of packet $P(m)$ be $d(m) = t$. Then $\mathcal{B}(t)$ can be determined by:

$$\mathcal{B}(t) \leq \inf\big\{k \geq 0 : d(m) \leq a(m+k)\big\}.$$

The following theorem provides a probabilistic bound on the system backlog in terms of the arrival process and the service process.

Theorem 15. ***(Backlog Bound)***

Consider a system providing an i.d SSC $\gamma(n) \in \mathcal{F}$ with bounding function $j(x) \in \bar{\mathcal{F}}$ to the input which has a v.w.d SAC $\lambda(n) \in \mathcal{F}$ with bounding function $h(x) \in \bar{\mathcal{F}}$. The backlog at time t (≥ 0) is bounded as below

$$P\{\mathcal{B}(t) > x\} \leq j \otimes h\big(\gamma \bar{\oslash} \lambda([x-1]^+)\big) \tag{4.14}$$

for $x \geq 1$.

Let

$$H(\lambda, \gamma + x) = \sup_{m \geq 0} \big\{ \inf[k \geq 0 : \gamma(m) + x \leq \lambda(m+k)] \big\}$$

represent the maximum horizontal distance between functions $\lambda(n)$ and $\gamma(n) + x$. The probability that $\mathcal{B}(t)$ exceeds $H(\lambda, \gamma + x)$ is bounded by

$$P\{\mathcal{B}(t) > H(\lambda, \gamma + x) + 1\} \leq j \otimes h(x). \tag{4.15}$$

Proof. From the condition, we have

$$\begin{aligned}
& d(m) - a(m+x) \\
= \;& \big[d(m) - a\bar{\otimes}\gamma(m)\big] + \big[a\bar{\otimes}\gamma(m) - a(m+x)\big] \\
= \;& \big[d(m) - a\bar{\otimes}\gamma(m)\big] + \sup_{0\le k\le m}\big\{a(k)+\gamma(m-k+1)\big\} - a(m+x) \\
= \;& \big[d(m) - a\bar{\otimes}\gamma(m)\big] + \sup_{0\le k\le m}\big\{\lambda(m+x-k) - [a(m+x)-a(k)] \\
& \qquad\qquad + \gamma(m-k+1) - \lambda(m+x-k)\big\} \\
\le \;& \big[d(m) - a\bar{\otimes}\gamma(m)\big] + \sup_{0\le k\le m+x}\big\{\lambda(m+x-k) - [a(m+x)-a(k)]\big\} \\
& - \inf_{0\le k\le m}\big\{\lambda(m-k+x) - \gamma(m-k+1)\big\}
\end{aligned}$$

Let $v = m-k+1$. The above inequality is written as

$$\begin{aligned}
& d(m) - a(m+x) \\
\le \;& \big[d(m) - a\bar{\otimes}\gamma(m)\big] + \sup_{0\le k\le m+x}\big\{\lambda(m+x-k) - [a(m+x)-a(k)]\big\} \\
& - \inf_{1\le v\le m+1}\big\{\lambda(v+x-1) - \gamma(v)\big\}.
\end{aligned}$$

Because there holds

$$\begin{aligned}
\inf_{1\le v\le m+1}\big\{\lambda(v+x-1)-\gamma(v)\big\} &\ge \inf_{v\ge 1}\big\{\lambda(v+x-1)-\gamma(v)\big\} \\
&= \lambda\bar{\oslash}\gamma([x-1]^+),
\end{aligned}$$

with the same conditions as analyzing the delay, we obtain

$$P\{\mathcal{B}(t) > x\} \le j \otimes h\big(\lambda\bar{\oslash}\gamma([x-1]^+)\big).$$

To prove Inequality (4.15), we replace $x = H(\lambda, \gamma+y)+1$ in event $\{\mathcal{B}(t) > x\}$ and have

$$\begin{aligned}
& d(m) - a(m + H(\lambda,\gamma+y)+1) \\
\le \;& \big[d(m) - a\bar{\otimes}\gamma(m)\big] + a\bar{\otimes}\lambda\big(m+H(\lambda,\gamma+y)+1\big) \\
& - a\big(m+H(\lambda,\gamma+y)+1\big) + \sup_{v\ge 0}\big\{\gamma(v) - \lambda(v+H(\lambda,\gamma+y))\big\}.
\end{aligned}$$

The definition of $H(\lambda, \gamma + y)$ implies

$$\gamma(v) + y \leq \lambda(v + H(\lambda, \gamma + y))$$

for any $v \geq 0$, i.e.,

$$\sup_{v \geq 0} \{\gamma(v) - \lambda(v + H(\lambda, \gamma + y))\} \leq -y.$$

Then we conclude

$$P\{\mathcal{B}(t) > H(\lambda, \gamma + x) + 1\} \leq j \otimes h(x).$$

□

Remark. $H(\lambda, \gamma + x)$ can be considered as the maximum system backlog in a (deterministic) virtual system, where the arrival process is $\lambda(n)$ and the service process is $\gamma(n) + x$. Inequality (4.15) is thus a bound on such maximum system backlog.

If the arrival process and the service process are independent of each other, from Lemma 1, another bound on the system backlog is provided as follows.

Lemma 10. ***(Backlog Bound: independent condition)***

Consider a system providing an i.d SSC $\gamma(n) \in \mathcal{F}$ with bounding function $j(x) \in \bar{\mathcal{F}}$ to the input which has a v.w.d SAC $\lambda(n) \in \mathcal{F}$ with bounding function $h(x) \in \bar{\mathcal{F}}$. Suppose the input process and the service process are independent. Then the backlog at time t (≥ 0) is bounded by:

$$P\{\mathcal{B}(t) > x\} \leq 1 - \bar{j} * \bar{h}\big(\gamma \bar{\oslash} \lambda([x-1]^+)\big) \tag{4.16}$$

for $x \geq 1$.

Let $H(\lambda, \gamma + x)$ represent the maximum horizontal distance between functions $\lambda(n)$ and $\gamma(n) + x$. The probability that $\mathcal{B}(t)$ exceeds $H(\lambda, \gamma + x)$ is bounded by

$$P\{\mathcal{B}(t) > H(\lambda, \gamma + x) + 1\} \leq 1 - \bar{j} * \bar{h}(x). \tag{4.17}$$

4.2 Output Characterization

The previous section introduces the service guarantees in a single node. Another common scenario which performance analysis deals with is the end-to-end service guarantees. An intuitive and simple approach is called *node-by-node* analysis [60]. The idea of node-by-node analysis is: it first analyzes service guarantees provided to a flow at each node along its path and then integrates obtained analytical results for all nodes of its path to get the end-to-end service guarantees provided by this path. In order to conduct the node-by-node analysis, it is necessary to know how to characterize the departure process from a single node.

Let us consider a simple network shown in Figure 4.2. After a flow passes through Server 1, its departure process is the arrival process for Server 2.

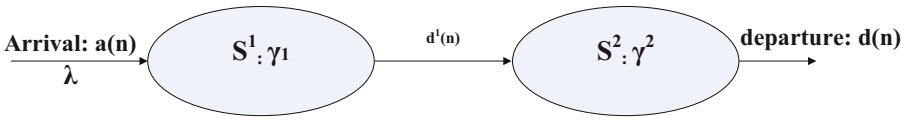

Figure 4.2: Output characterization

Assume the initial arrival $a(n)$ has a stochastic arrival curve $\lambda(n)$ and each server provides service with a stochastic service curve γ^k, $k = 1, 2$. In order to analyze the end-to-end performance for the initial arrival flow, such as delay bound, the intuitive method is to derive the delay bound in Server 1 and Server 2, respectively. Adding these two delay bounds together gives the end-to-end delay bound.

Deriving the delay bound in Server 1 directly follows the result of Section 4.1.1. When deriving the delay bound in Server 2, we need to characterize the arrival process of Server 2. Clearly, the arrival process of Server 2 is the departure process of Server 1. Now the question is how to characterize the departure process of Server 1 based on the given conditions?

Theorem 16. ***(Output Characterization)***

Consider a system provides an i.d SSC $\gamma(n) \in \mathcal{F}$ with bounding function $j(x) \in \bar{\mathcal{F}}$ to its arrival which has a v.w.d SAC $\lambda(n) \in \mathcal{F}$ with

bounding function $h(x) \in \bar{\mathcal{F}}$. *The output has an i.a.t SAC* $\lambda \bar{\oslash} \gamma(n-m-1)$ *with bounding function* $j \otimes h(x) \in \bar{\mathcal{F}}$, *i.e., for any* $0 \leq m < n-1$, *there holds*

$$P\Big\{\lambda \bar{\oslash} \gamma(n-m-1) - [d(n) - d(m)] > x\Big\} \leq j \otimes h(x). \qquad (4.18)$$

Proof. For any two departure packets $m < n$, there holds

$$\begin{aligned} d(m) - d(n) &\leq d(m) - a(n) \\ &= d(m) - a(n) + a \bar{\otimes} \gamma(m) - a \bar{\otimes} \gamma(m) \\ &= \big[d(m) - a \bar{\otimes} \gamma(m)\big] + \sup_{0 \leq k \leq m} \big\{a(k) + \gamma(m-k+1)\big\} - a(n) \\ &= \big[d(m) - a \bar{\otimes} \gamma(m)\big] + \sup_{0 \leq k \leq m} \big\{\gamma(m-k+1) - [a(n) - a(k)]\big\} \end{aligned}$$

$$\begin{aligned} &= \big[d(m) - a \bar{\otimes} \gamma(m)\big] + \sup_{0 \leq k \leq m} \Big\{\gamma(m-k+1) - \lambda(n-k) \\ &\qquad\qquad + \lambda(n-k) - [a(n) - a(k)]\Big\} \\ &\leq \big[d(m) - a \bar{\otimes} \gamma(m)\big] + \sup_{0 \leq k \leq m} \Big\{\lambda(n-k) - [a(n) - a(k)]\Big\} \\ &\quad + \sup_{0 \leq k \leq m} \big\{\gamma(m-k+1) - \lambda(n-k)\big\}. \end{aligned}$$

Let $v = m - k + 1$. Then the above inequality is written as

$$\begin{aligned} d(m) - d(n) &\leq \big[d(m) - a \bar{\otimes} \gamma(m)\big] + \sup_{0 \leq k \leq n} \Big\{\lambda(n-k) - [a(n) - a(k)]\Big\} \\ &\quad - \inf_{1 \leq v \leq m+1} \big\{\lambda(n-m-1+v) - \gamma(v)\big\} \\ &\leq \big[d(m) - a \bar{\otimes} \gamma(m)\big] + \sup_{0 \leq k \leq n} \Big\{\lambda(n-k) - [a(n) - a(k)]\Big\} \\ &\quad - \inf_{0 \leq v \leq m+1} \big\{\lambda(n-m-1+v) - \gamma(v)\big\} \end{aligned}$$

where the last step is because

$$\inf_{0 \leq k \leq m+1} [f_k] \leq \inf_{1 \leq k \leq m+1} [f_k].$$

Adding $\inf_{0\le v\le m+1}\{\lambda(n-m-1+v)-\gamma(v)\}$ to both sides of the above inequality results in

$$\begin{aligned}&\inf_{0\le v\le m+1}\{\lambda(n-m-1+v)-\gamma(v)\}-[d(n)-d(m)]\\ \le\ &[d(m)-a\bar{\otimes}\gamma(m)]+\sup_{0\le k\le n}\{\lambda(n-k)-[a(n)-a(k)]\}.\end{aligned}$$

In addition, there holds

$$\begin{aligned}\lambda\bar{\oslash}\gamma(n-m-1) &= \inf_{v\ge 0}\{\lambda(n-m-1+v)-\gamma(v)\}\\ &\le \inf_{0\le v\le m+1}\{\lambda(n-m-1+v)-\gamma(v)\}.\end{aligned}$$

To ensure that the right-hand side of the above inequality is meaningful, it requires $n-m-1>0$. With the same conditions as analyzing delay, we conclude

$$\begin{aligned}&P\Big\{\lambda\bar{\oslash}\gamma(n-m-1)-[d(n)-d(m)]>x\Big\}\\ \le\ &P\Big\{[d(m)-a\bar{\otimes}\gamma(m)]+\sup_{0\le k\le n}\{\lambda(n-k)-[a(n)-a(k)]\}>x\Big\}\\ \le\ &j\otimes h(x).\end{aligned}$$

□

Remark. In the above theorem, the initial arrival process has a *v.w.d* SAC while the departure process has an *i.a.t* SAC. In order to derive the service guarantees in Server 2, we need Theorem 10 (2) to transform the *i.a.t* SAC into a *v.w.d* SAC. Such transformation introduces a more loose bounding function. The node-by-node analysis thus generates a more loose end-to-end delay bound. Alternatively, network calculus possesses an attractive property - the concatenation property which is discussed in the following section. The comparison between the node-by-node analysis and the concatenation analysis has proven that the latter can yield a tighter bound on the end-to-end delay bound [62].

The output characterization property however is very useful when analyzing complicated network scenarios, such as Figure 4.3, where

flows join or leave dynamically. In order to analyze the per-flow service guarantees, the departure process from each single node should be characterized using the same traffic model as the arrival process and the service model of this node.

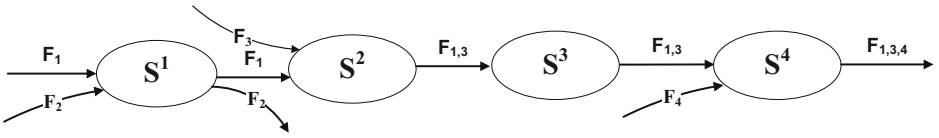

Figure 4.3: Complicated network scenario

Moreover, if the arrival process and the service process are independent of each other, the following lemma characterizes the departure process.

Lemma 11. ***(Output Characterization: independent condition.)***

Consider a system provides an i.d SSC $\gamma(n) \in \mathcal{F}$ *with bounding function* $j(x) \in \bar{\mathcal{F}}$ *to its arrival which has a v.w.d SAC* $\lambda(n) \in \mathcal{F}$ *with bounding function* $h(x) \in \bar{\mathcal{F}}$*. The output has an i.a.t SAC* $\lambda^* \in \mathcal{F}$ *with bounding function* $h^*(x) \in \bar{\mathcal{F}}$*, where*

$$\begin{aligned} \lambda^*(n) &= \lambda \bar{\oslash} \gamma(n-1), \\ h^*(x) &= 1 - \bar{j} * \bar{h}(x). \end{aligned} \tag{4.19}$$

4.3 Concatenation Property

The concatenation property uses an equivalent system to represent a system of multiple servers connected in tandem, each of which provides a stochastic service curve to the input. Then such equivalent system provides the initial input a stochastic service curve, which is derived from the stochastic service curve provided by all involved individual servers.

In the following discussion, γ^k and j^k denote the stochastic service curve and bounding function of the kth server. For packet $P(n)$, the time arriving to the kth server is $a^k(n)$ and the time departing from the kth server is $d^k(n)$. For a network of N tandem servers, the initial arrival is $a(n) = a^1(n)$ and the final departure is $d(n) = d^N(n)$.

Theorem 17. ***(Concatenation Property)***

Consider a flow passing through a network of N nodes connected in tandem. If each node $k(= 1,2,...,N)$ provides an i.d SSC $\gamma^k(n) \in \mathcal{F}$ with bounding function $j^k(x) \in \bar{\mathcal{G}}$ to its input, the network provides to the initial input an i.d SSC $\gamma(n)$ with bounding function $j(x)$, where

$$\begin{aligned} \gamma(n) &= \gamma^1 \bar{\otimes} \gamma_\eta^2 \bar{\otimes} \cdots \bar{\otimes} \gamma_{(N-1)\eta}^N(n) \\ j(x) &= j^{1,\eta_1} \otimes j^{2,\eta_2} \otimes \cdots \otimes j^N(x), \end{aligned}$$

with

$$\gamma_{(k-1)\eta}^k(n) = \gamma^k(n) + (k-1) \cdot \eta \cdot n$$

for $k = 2, ..., N$ and $\eta > 0$, and

$$j^{k,\eta_k}(x) = \left[j^k(x) + \frac{1}{\eta_k} \int_x^\infty j^k(y) dy \right]_1$$

for $k = 1, ..., N-1$ and $\eta_k > 0$.

Proof. We shall only prove the three-node case, from which, the proof can be easily extended to the N-node case.

The departure of the first node is the arrival to the second node,

so $d^1(n) = a^2(n)$ and $d^2(n) = a^3(n)$. We then have,

$$
\begin{aligned}
& d(n) - a\bar{\otimes}\gamma^1\bar{\otimes}\gamma_\eta^2\bar{\otimes}\gamma_{2\eta}^3(n) \\
= \; & d(n) - \sup_{0\le m\le n}\Big\{a\bar{\otimes}\gamma^1(m) + \gamma_\eta^2\bar{\otimes}\gamma_{2\eta}^3(n-m+1)\Big\} + d^1(m) - d^1(m) \\
\le \; & d(n) - \sup_{0\le m\le n}\Big\{\gamma_\eta^2\bar{\otimes}\gamma_{2\eta}^3(n-m+1) + d^1(m) - \eta\cdot(n-m+1) \\
& \qquad\qquad - [d^1(m) - a\bar{\otimes}\gamma^1(m) - \eta\cdot(n-m)]\Big\} \\
\le \; & d(n) - \sup_{0\le m\le n}\Big\{\gamma_\eta^2\bar{\otimes}\gamma_{2\eta}^3(n-m+1) + d^1(m) - \eta\cdot(n-m+1)\Big\} \\
& + \sup_{0\le m\le n}\Big\{d^1(m) - a\bar{\otimes}\gamma^1(m) - \eta\cdot(n-m)]\Big\} \\
= \; & d(n) - \sup_{0\le m\le n}\{a^2(m) + \sup_{0\le k\le n-m+1}[\gamma^2(k) + \eta\cdot k + \gamma^3(n-m+1-k) \\
& +2\eta\cdot(n-m+1-k)] - \eta\cdot(n-m+1)\big\} + \\
& + \sup_{0\le m\le n}\big\{d^1(m) - a\bar{\otimes}\gamma^1(m) - \eta\cdot(n-m)\big\} \\
= \; & d(n) - \sup_{0\le m\le n}\big\{a^2(m) + \sup_{0\le k\le n-m+1}[\gamma^2(k) + \gamma^3(n-m+1-k) \\
& +\eta\cdot(n-m+1-k)]\big\} + \sup_{0\le m\le n}\big\{d^1(m) - a\bar{\otimes}\gamma^1(m) - \eta\cdot(n-m)\big\} \\
= \; & d(n) - a^2\bar{\otimes}\gamma^2\bar{\otimes}\gamma_\eta^3(n) + \sup_{0\le m\le n}\big\{d^1(m) - a\bar{\otimes}\gamma^1(m) - \eta\cdot(n-m)\big\} \\
\le \; & d(n) - \sup_{0\le m\le n}\Big\{a^2\bar{\otimes}\gamma^2(m) + \gamma_\eta^3(n-m+1)\Big\} - a^3(m) + \eta\cdot(n-m+1) \\
& +d^2(m) - \eta\cdot(n-m) + \sup_{0\le m\le n}\big\{d^1(m) - a\bar{\otimes}\gamma^1(m) - \eta\cdot(n-m)\big\} \\
\le \; & d(n) - \sup_{0\le m\le n}\big\{a^3(m) + \gamma_\eta^3(n-m+1) - \eta\cdot(n-m+1)\big\} \\
& + \sup_{0\le m\le n}\big\{d^2(m) - a^2\bar{\otimes}\gamma^2(m) - \eta\cdot(n-m)\big\} \\
& + \sup_{0\le m\le n}\big\{d^1(m) - a\bar{\otimes}\gamma^1(m) - \eta\cdot(n-m)\big\} \\
= \; & d(n) - a^3\bar{\otimes}\gamma^3(n) + \sup_{0\le m\le n}\big\{d^2(m) - a^2\bar{\otimes}\gamma^2(m) - \eta\cdot(n-m)\big\} \\
& + \sup_{0\le m\le n}\big\{d^1(m) - a\bar{\otimes}\gamma^1(m) - \eta\cdot(n-m)\big\}.
\end{aligned}
$$

Based on the relationship between the $i.d$ SSC and the η-stochastic service curve presented in Theorem 12(2), the following inequality

holds

$$P\{d(n) - a\bar{\otimes}\gamma^1\bar{\otimes}\gamma^2_\eta\bar{\otimes}\gamma^3_{2\eta}(n) > x\} \leq j^3 \otimes j^{2,\eta_2} \otimes j^{1,\eta_1},$$

which completes the proof.

Remark. Note that both the max-plus convolution and the min-plus convolution are associative and commutative. □

The proof of Theorem 17 utilizes the relationship between the *i.d* SSC and the $\eta-$stochastic service curve. The following lemma directly describes the service characterization of a network of nodes connected in tandem, where each single node provides an $\eta-$stochastic service curve to its input.

Lemma 12. *Consider a flow passing through a network of N nodes connected in tandem. If each node $k(= 1,2,...,N)$ provides an η-stochastic service curve $\gamma^k(n) \in \mathcal{F}$ with bounding function $j^k(x) \in \bar{\mathcal{F}}$ to its input, i.e.,*

$$P\Big\{\sup_{0\leq m\leq n}\big\{d^k(m) - a^k\bar{\otimes}\gamma^k(m) - \eta\cdot(n-m)\big\} > x\Big\} \leq j^k(x),$$

then the network guarantees to the initial arrival process an i.d SSC $\gamma(n)$ with bounding function $j(x)$ with

$$\gamma(n) = \gamma^1\bar{\otimes}\gamma^2_\eta\bar{\otimes}\cdots\bar{\otimes}\gamma^N_{(N-1)\eta}(n)$$

$$j(x) = j^1 \otimes j^2 \otimes \cdots \otimes j^N(x),$$

where $\gamma^k_{(k-1)\eta}(n) = \gamma^k(n) + (k-1)\cdot\eta\cdot n$, $k = 2,...,N$, for any small $\eta > 0$.

Proof. We shall only prove two-node case, from which, the proof can be extended to the N-node case. Keep in mind that $a^2(n) = d^1(n)$.

For the two-node case, we have

$$\begin{aligned}
& d(n) - a\bar{\otimes}\gamma^1\bar{\otimes}\gamma_\eta^2(n) \\
= \; & d(n) - \sup_{0\le m\le n}\big\{a\bar{\otimes}\gamma^1(m) + \gamma^2(n-m+1) + \eta\cdot(n-m+1)\big\} \\
\le \; & d(n) - \sup_{0\le m\le n}\big\{a\bar{\otimes}\gamma^1(m) + \gamma^2(n-m+1) + \eta\cdot(n-m)\big\} \\
& + d^1(m) - a^2(m) \\
\le \; & d(n) - \sup_{0\le m\le n}\big\{a^2(m) + \gamma^2(n-m+1)\big\} \\
& + \sup_{0\le m\le n}\big\{d^1(m) - a\bar{\otimes}\gamma^1(m) - \eta\cdot(n-m)\big\} \\
= \; & d(n) - a^2\bar{\otimes}\gamma^2(n) + \sup_{0\le m\le n}\big\{d^1(m) - a\bar{\otimes}\gamma^1(m) - \eta\cdot(n-m)\big\} \\
\le \; & \sup_{0\le m\le n}\big\{d(m) - a^2\bar{\otimes}\gamma^2(m) - \eta\cdot(n-m)\big\} \\
& + \sup_{0\le m\le n}\big\{d^1(m) - a\bar{\otimes}\gamma^1(m) - \eta\cdot(n-m)\big\}.
\end{aligned}$$

The last step holds because of Theorem 12(1). From the condition, we conclude

$$P\big\{d(n) - a\bar{\otimes}\gamma^1\bar{\otimes}\gamma_\eta^2(n) > x\big\} \le j^1 \otimes j^2(x).$$

□

If the arrival process and all individual service processes are independent of each other, the following lemma summarizes the service characterization of such a network.

Lemma 13. ***(Concatenation property: independent condition).***

Consider a flow passing through a network of N nodes connected in tandem. Suppose each node k(= 1,2,...,N) provides an η-stochastic service curve $\gamma^k(n) \in \mathcal{F}$ *with bounding function* $j^k(x) \in \bar{\mathcal{F}}$ *to its input. The initial arrival process and all individual service processes are independent of each other. Then the network provides to the initial*

input an i.d SSC $\gamma(n)$ *with bounding function* $j(x)$ *with*

$$\begin{aligned} \gamma(n) &= \gamma^1 \bar{\otimes} \gamma_\eta^2 \bar{\otimes} \cdots \bar{\otimes} \gamma_{(N-1)\eta}^N(n) \\ \\ j(x) &= 1 - \bar{j^1} * \bar{j^2} * \cdots * \bar{j^N}(x), \end{aligned} \tag{4.20}$$

where $\bar{j^k} = 1 - [j^k]_1$ *for* $k = 1, ..., N$.

Remark. The proof of the concatenation property reveals another reason of defining the $\eta-$stochastic service curve model.

4.4 Superposition Property

The superposition property can be applied for treating multiple individual flows as an aggregate flow under the FIFO aggregate scheduling. Particularly, if the arrival process of each individual flow can be stochastically characterized by a stochastic arrival curve, we also can find a stochastic arrival curve to describe the arrival process of the aggregate flow. Then we only need to analyze the service guarantees for the aggregate flow since all constituent flows are served equally.

4.4.1 Superposition of Renewal Processes

The superposition of multiple flows essentially falls into the research issue - superposition of renewal processes. In queueing networks, an individual server may receive inputs from different sources. Therefore, it may be reasonable to postulate that the arrival process to a server is a superposition of statistically independent constituent processes [76]. The individual constituent processes are typically considered as renewal processes. A *renewal process* is a counting process for which the times between successive events are independent and identically distributed possibly with an arbitrary distribution [91].

The arrival process of each constituent flow F_i is a renewal process denoted by $\{\mathcal{A}_i(t), t \geq 0\}$ which is a non-negative integer-valued stochastic process. This renewal process has inter-arrival times $\tau_{i,1}$,$\tau_{i,2}$,... and the $(n+1) - th$ renewal $a_i(n) = \sum_{k=0}^{n} \tau_{i,k}$. The superposition of

renewal processes has been widely studied since the original investigation by Cox and Smith [32]. However, the renewal property is not preserved under superposition except for Poisson sources. More precisely, the inter-arrival times in the superposition process are statistically dependent, a property that cannot be captured by a renewal model [96].

One important work has pointed out that if the number of constituent processes tends to infinity, the superposition converges to a Poisson process [6] [7]. Although the approximate renewal model [6] shows very low errors in the analysis of a representative set of $\sum_i GI_i/G/1$ queueing systems, it requires that the number of constituent processes should be large enough. There is a lot of research work on the superposition of renewal processes [41] [71].

The available literature on this research issue mainly focuses on finding the exact stationary distribution of the interval between two consecutive events for the superposition process. Moreover, much mathematical knowledge is typically needed to conduct the analysis.

In the following, we introduce an approach to characterize the superposition processes of multiple flows from a network calculus viewpoint.

4.4.2 Arrival Time Determination

First, we only consider the superposition of two flows, F_1 and F_2. Let $a_1(n)$, $a_2(n)$ and $a(n)$ be the arrival process of F_1, F_2 and the aggregate flow, respectively. As shown in Figure 4.4, F_1 and F_2 are aggregated in the FIFO manner at the server.

Figure 4.4: Aggregation of two flows

Figure 4.5 illustrates that the arrival process of the aggregate flow is dependent on the arrival process of two individual flows. Packet

$P(n)$[3] of the aggregate flow is either the mth packet of flow F_1 or the $(n+1-m)$th packet of flow F_2, where $0 \leq m \leq n+1$.

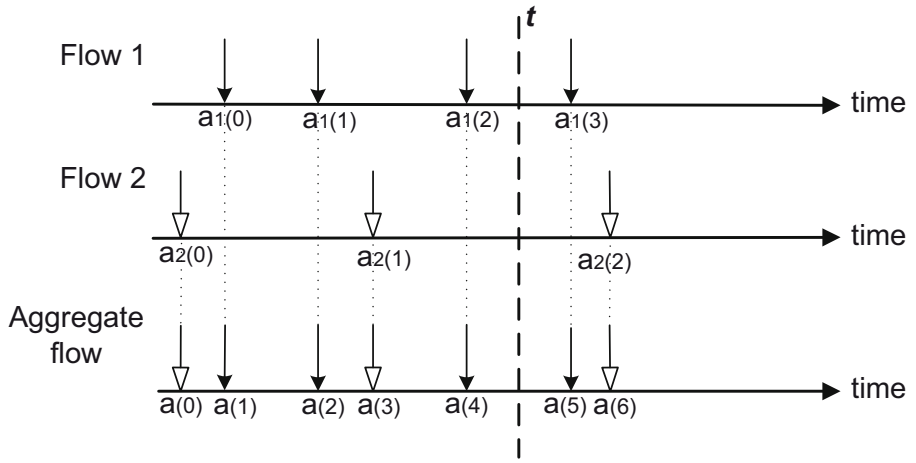

Figure 4.5: Packet arrival time

For instance, when observing the arrival process of the aggregate flow at time t, we find that packet $P(4)$ (arrival time: $a(4) < t$) has arrived to the server. Packet $P(4)$ is the $3rd$ packet of F_1 and arrives to the server later than the $2nd$ packet of F_2, i.e., $a_1(2) = a(4) < a_2(1)$. Thus, we have the generalized expression for the arrival time of packet $P(n)$:

$$a(n) = \max\big[a_1(m-1), a_2(n-m)\big]. \tag{4.21}$$

However, from Eq.(4.21), we can have $n+1$ combinations for $0 \leq m \leq n+1$. By convention, we adopt $a_i(m) = 0$ if $m < 0$. The minimum among all combinations represents the time when the first packet of these combinations arrives to the server. Then this packet is inserted into the FIFO queue as packet $P(n)$ and its arrival time is

$$a(n) = \min_{0 \leq m \leq n+1} \Big\{ \max\big[a_1(m-1), a_2(n-m)\big] \Big\} \tag{4.22}$$

with

$$a(0) = \min\Big\{ \max\big[0, a_2(0)\big], \max\big[a_1(0), 0\big] \Big\} = \min[a_1(0), a_2(0)].$$

[3]Recall that $P(n)$ denotes the $(n+1)$th packet of the aggregate flow. The same notation is also used for single flows F_1 and F_2.

We again examine the arrival time of packet $P(4)$ according to Eq.(4.22). The arrival time $a(4)$ is the minimum of the following set:

$$\Big\{ \max\big[a_1(4), 0\big] = a_1(4), \max\big[0, a_2(4)\big] = a_2(4),$$
$$\max[a_1(0), a_2(3)] = a_2(3), \max[a_1(1), a_2(2)] = a_2(2),$$
$$\max[a_1(2), a_2(1)] = a_1(2), \max[a_1(3), a_2(0)] = a_1(3)\Big\},$$

among which, the combination $\max[a_1(2), a_2(1)] = a_1(2)$ is the minimum and thus $a(4) = a_1(2)$.

Eq.(4.22) can be generalized to the aggregation of $N(\geq 2)$ flows:

$$a(n) = \min_{\sum m_i = n+1} \Big\{ \max[a_1(m_1 - 1), a_2(m_2 - 1), ..., a_N(n - \sum_{i=1}^{N-1} m_i)] \Big\} \tag{4.23}$$

for $0 \leq m_i \leq n + 1$.

4.4.3 Superposition Process Characterization

Eq.(4.22) can compute the packet arrival time of the aggregate flow, whereas we still have the difficulty in characterizing the packet inter-arrival time of the aggregate flow if the packet inter-arrival times of two constituent flows follow the general distribution. For this reason, it is difficult to directly characterize the arrival process of the aggregate flow from the temporal perspective. Alternatively, we rely on the available results of the superposition property explored in the space-domain (see Theorem 9).

In Figure 4.5, the cumulative arrival packets of the aggregate flow by time t, $\mathcal{A}(t)$, equals $\mathcal{A}_1(t) + \mathcal{A}_2(t)$, from which we can indirectly find the stochastic arrival curve for the aggregate flow.

As shown in Figure 4.6, the condition is that the time-domain stochastic arrival curve of all constituent flows are known, and the target is to verify that the aggregate flow also has a time-domain stochastic arrival curve.

Since it is difficult to directly reach the target, we try to find a bypass. If a flow has a time-domain $v.w.d$ SAC, with Theorem 11(2), this flow has a space-domain $v.b.c$ SAC, for which the superposition property holds, i.e., Theorem 9. Applying Theorem 11(1) can get the $v.w.d$ SAC for the aggregate flow.

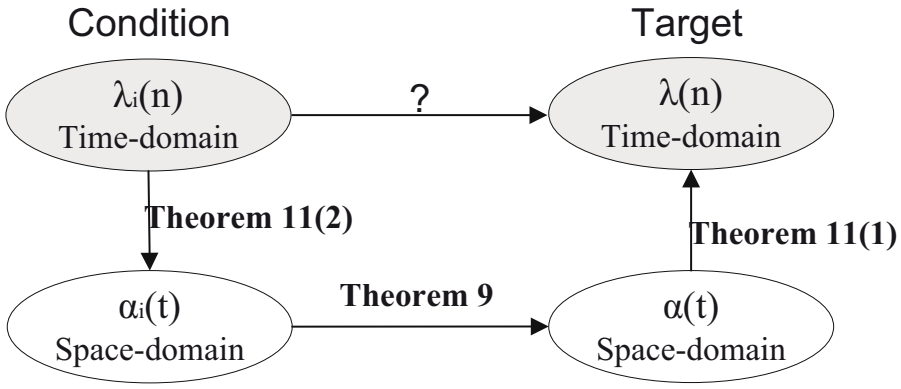

Figure 4.6: Transformation in Theorem 18

If flow F_i has a *v.w.d* SAC $\lambda_i(n)$ with bounding function $h_i(x)$, $i = 1, 2, ..., N$, from Theorem 11(2), we can verify that flow F_i has a *v.b.c* SAC $\alpha_i(t)$ with bounding function $f_i(x) = h_i\big(z_i^{-1}(x)\big)$, where $\alpha_i(t)$ and $z_i^{-1}(x)$ are given in Theorem 11(2). Furthermore, according to Theorem 9, the aggregate flow has a *v.b.c* SAC $\alpha(t) = \sum_{i=1}^{N} \alpha_i(t)$ with bounding function $f(x) = f_1 \otimes \cdots \otimes f_N(x)$. Finally, we apply Theorem 11(1) and can verify that the aggregate flow also has a *v.w.d* SAC.

Theorem 18. ***(Superposition property)***

Consider the aggregate of N flows. If the arrival process of each flow has a v.w.d SAC $\lambda_i(n) \in \mathcal{F}$ *for* $i = 1, 2, ..., N$*, i.e.,*

$$P\{a_i(n) < a_i \bar{\otimes} \lambda_i(n) - y\} \leq h_i(y),$$

which implies that every flow also has a v.b.c SAC

$$\alpha_i(t) = \sup\{k : \lambda_i(k) \leq t\}$$

with bounding function

$$f_i(x) = h_i\big(z_i^{-1}(x)\big)$$

where $z_i^{-1}(x)$ *denote the inverse function of* x*:*

$$x = z_i(y) \equiv \sup_{\tau \geq 0}\{\alpha_i(\tau + y) - \alpha_i(\tau) + 1\}.$$

Then the aggregate arrival process $a(n)$ *has a v.w.d SAC* $\lambda(n)$ *with bounding function* $h(y)$*, where*

$$\lambda(n) = \inf\{\tau : \sum_{i=1}^{N} \alpha_i(\tau) \geq n\}, \qquad h(y) = f\big(z^{-1}(y)\big),$$

with $f(x) = f_1 \otimes \cdots \otimes f_N(x)$ *and* $z^{-1}(y)$ *denoting the inverse function of* y*:*

$$y = z(x) \equiv \sup_{k \geq 0}\{\lambda(k) - \lambda(k - x)\}.$$

4.4.4 Special Case: Superposition of Poisson Processes

As we have mentioned in Section 4.4.1, the Poisson process is a special case of renewal processes because its renewal property is still preserved under superposition. On the hand, the superposition of multiple Poisson processes is still a Poisson process. If describing a Poisson process from the temporal perspective, the inter-arrival times between two arbitrary events follow the Gamma distribution. We can readily examine the superposition property of Poisson processes from the temporal perspective.

Example 8. *Consider the superposition process of two independent Poisson arrival processes. Suppose all packets of both arrival processes have the same size. The packet inter-arrival times follow exponential distributions with mean* $\frac{1}{\mu_1}$ *and* $\frac{1}{\mu_2}$*, respectively. Find the time-domain v.w.d stochastic arrival curve for the superposition process.*

It is well-known that the superposition of two Poisson processes is still a Poisson process. According to the condition, the superposition

process is a Poisson process with mean arrival rate $\mu_1 + \mu_2$. If characterizing the superposition process in terms of the inter-arrival time between two arbitrary packet, it is a Gamma process. Recall Example 4 which has given the $v.w.d$ stochastic arrival curve for a Gamma process. We thus know that the superposition process has a $v.w.d$ stochastic arrival curve $\lambda_s(n) = T_s \cdot n$ $(0 < T_s < \frac{1}{\mu_1+\mu_2})$ with bounding function $h_s(x)$:

$$h_s(x) = 1 - (1 - \rho_s) \sum_{i=0}^{\lfloor \frac{x}{T_s} \rfloor} e^{-(\mu_1+\mu_2)(iT_s - x)} \frac{[(\mu_1 + \mu_2)(iT_s - x)]^i}{i!},$$

where $\rho_s = (\mu_1 + \mu_2) \cdot T_s$.

Remark. It is readily to generalize the above example into the superposition of multiple independent Poisson processes.

4.5 Conclusion

This section investigated four basic properties of stochastic network calculus under the time-domain models defined in Chapter 3.

Deriving the delay bound is intuitively straightforward because we directly applied the definition of the $v.w.d$ SAC and the $i.d$ SSC. Whereas both the $v.w.d$ SAC and the $i.d$ SSC are difficult to find. This difficulty can be solved for some special cases by constructing two virtual SSQs, such as $GI/GI/1$ queue. It requires that the packet inter-arrival times are independent and identically distributed. Since the per packet service times are typically independent, this condition is often satisfied.

The system backlog is essentially a space-domain property. By fixing the packet departure time, the number of backlogged packets can be determined. This property also holds under the combination of the $v.w.d$ SAC and the $i.d$ SSC. Moreover, the bound on the maximum system backlog in a (deterministic) virtual system is given. This bound may be applied as a threshold to measure whether the system is stable.

The output characterization is very useful for analyzing the end-to-end performance. Particularly, the node-by-node analysis requires that the arrival process to each node along a path is known. The arrival process to the current node is the departure process from the previous node. The output characterization property shows that the

departure process of a flow can be represented using the traffic of the arrival process and the service process provided to the arrival process. This property holds under the combination of the $v.w.d$ SAC and the $i.d$ SSC. However, the derived departure process has a $i.a.t$ SAC while not the $v.w.d$ SAC.

The concatenation property can simplify the analysis and yield tighter performance bounds compared with the node-by-node analysis approach. We indirectly prove that the $i.d$ SSC holds such a property through transforming the $i.d$ into the $\eta-$stochastic service curve. From investigating this property, the reason of defining $\eta-$stochastic service curve is clarified. This property is only relevant to the service process.

Exploring the superposition property attempts to tackle a research problem which has been studied for several decades, i.e., the superposition of renewal processes [32]. The available literature on this research problem mainly focuses on finding the exact stationary distribution of the interval between two events. In order to achieve this objective, most work requires extensive mathematical background and is not readily understandable. In the context of stochastic network calculus, although the superposition property cannot be directly proved from the temporal perspective, it can be indirectly obtained through transforming the time-domain arrival curve into the space-domain arrival curve. This property only replies on the arrival process.

Chapter 5

Concretization of Generic Models

The main work in this chapter is based on the following paper.

Jing Xie and Yuming Jiang. "A Temporal Network Calculus Approach to Service Guarantee Analysis of Stochastic Networks." *In Proceedings of 5th International ICST Conference on Performance Evaluation Methodologies and Tools (ValueTools)*, Paris, France, May 2011. (Regular paper)

In the previous chapters, the time-domain traffic and service models have been defined. Based on the defined models, four fundamental properties have been explored. However, a clear guidance on finding the arrival curve characterization for an arrival process or the service curve characterization for a service process is missing. In this chapter, we concretize the time-domain traffic and service models by linking some well-known stochastic processes to them. In addition, we exemplify the temporal analysis approach by investigating the delay performance of a Gilbert-Elliott channel. The results show that the delay bound can be improved under the independence condition. Furthermore, a comparison between the temporal and the spatial analysis results reveals that the two analytical approaches essentially yield close results.

5.1 Arrival Process Characterization

In this section, we first recall the definitions of the *i.a.t* SAC and the *v.w.d* SAC models and then obtain the stochastic arrival curve characterization of the arrival process.

5.1.1 Moment Generating Function of Inter-arrival Time

When observing the traffic arrival process $\Gamma(m, m+n)$ from the temporal perspective, it is indeed formed from process $\{\tau_{m+1}\}$, where $\tau_{m+1} \equiv a(m+1) - a(m)$. In order to guarantee a certain level of QoS to this arrival process, the packet inter-arrival times should be constrained. For the arrival process $\Gamma(m, m+n)$ formed by identically distributed $\{\tau_{m+1}\}$, suppose there exists $\varphi(\eta, n)$ satisfying

$$\mathbf{E}[e^{\eta\Gamma(m,m+n)}] \geq e^{\eta n\varphi(\eta,n)},$$

which becomes the following expression

$$\frac{1}{\eta n} log\mathbf{E}\left[e^{\eta\Gamma(m,m+n)}\right] \geq \varphi(\eta, n),$$

where $\mathbf{E}[e^{\eta\Gamma(m,m+n)}]$ is the MGF of the arrival process $\Gamma(m, m+n)$. By convention, we adopt $\varphi(\eta, 0) = 0$.

If $\{\tau_{m+1}\}$ are *i.i.d.*, then it is easily verified that

$$\frac{1}{\eta n} log\mathbf{E}\big[e^{\eta\Gamma(m,m+n)}\big] = \frac{1}{\eta} log\mathbf{E}\big[e^{\eta\tau_0}\big],$$

which is independent of n, and we hence adopt:

$$\frac{1}{\eta} log\mathbf{E}\big[e^{\eta\tau_0}\big] \geq \varphi(\eta).$$

5.1.2 i.a.t Stochastic Arrival Curve Characterization

Recall the definition of the *i.a.t* SAC (see Definition 10). If an arrival process has an *i.a.t* SAC $\lambda(n)$ with bounding function $h(x)$, then there holds:

$$P\Big\{\Gamma(m,m+n) < \big[\lambda(n) - x\big]^+\Big\} \leq h(x). \tag{5.1}$$

Assume $\{\tau_{m+1}\}$ are identically distributed. Let

$$\varphi(\eta,n) \leq \frac{1}{\eta n} log\mathbf{E}\big[e^{\eta\Gamma(m,m+n)}\big].$$

Then Inequality (5.1) is rewritten as

$$\begin{aligned}
& P\big\{\varphi(\eta,n)\cdot n - \Gamma(m,m+n) > x\big\} \\
= \;& P\big\{e^{\eta[\varphi(\eta,n)\cdot n - \Gamma(m,m+n)]} > e^{\eta x}\big\} \\
\leq \;& e^{-\eta x}\mathbf{E}\big[e^{\eta[\varphi(\eta,n)\cdot n - \Gamma(m,m+n)]}\big] \qquad (5.2) \\
= \;& e^{-\eta x}\frac{e^{\eta\varphi(\eta,n)\cdot n}}{\mathbf{E}\big[e^{\eta\Gamma(m,m+n)}\big]} \\
\leq \;& e^{-\eta x} \qquad (5.3)
\end{aligned}$$

for $\eta > 0$. Here step (5.2) is known as the Chernoff bound (see Inequality (2.9)), and step (5.3) is obtained due to $\mathbf{E}\big[e^{\eta\Gamma(m,m+n)}\big] \geq e^{\eta n\varphi(\eta,n)}$ by definition. The following lemma summarizes the above result.

Lemma 14. *For an arrival process $\Gamma(m,m+n)$, if there exists $\varphi(\eta,n)$ which satisfies, for $m \geq 0, n > 0$,*

$$\frac{1}{\eta n} log\boldsymbol{E}\big[e^{\eta\Gamma(m,m+n)}\big] \geq \varphi(\eta,n),$$

then this process has an i.a.t. SAC $\lambda(n) = \varphi(\eta, n) \cdot n$ *with bounding function* $h(x) = e^{-\eta x}$ *for* $\eta > 0$.

If $\{\tau_{m+1}\}$ are *i.i.d.*, let $\varphi(\eta) \leq \frac{1}{\eta}\mathbf{E}[e^{\eta\tau_0}]$. Then Inequality (5.1) is rewritten as follows:

$$\begin{aligned} & P\Big\{\varphi(\eta)\cdot n - \Gamma(m, m+n) > x\Big\} \\ \leq\ & e^{-\eta x}\mathbf{E}\big[e^{\eta[\varphi(\eta)\cdot n - \Gamma(m,m+n)]}\big] \\ =\ & e^{-\eta x}\Big(\mathbf{E}\big[e^{\eta[\varphi(\eta)-\tau_0]}\big]\Big)^n \\ \leq\ & e^{-\eta x}\mathbf{E}\big[e^{\eta[\varphi(\eta)-\tau_0]}\big], \end{aligned}$$

from which, we have the following lemma.

Lemma 15. *If the inter-arrival times of arrival process* $\Gamma(m, m+n)$ *are i.i.d., then the arrival process has an i.a.t. SAC* $\lambda(n) = \varphi(\eta) \cdot n$ *with bounding function* $h(x) = e^{-\eta x}\boldsymbol{E}\big[e^{\eta[\varphi(\eta)-\tau_0]}\big]$, *where* $\varphi(\eta) \leq \frac{1}{\eta} log\boldsymbol{E}\big[e^{\eta\tau_0}\big]$ *for* $\eta > 0$.

Remark: Lemma 14 becomes Lemma 15 by taking into consideration the independence condition of inter-arrival times.

5.1.3 v.w.d Stochastic Arrival Curve Characterization

Recall the definition of the *v.w.d* SAC (see Definition 11). If an arrival process $\Gamma(m, n)$ has a *v.w.d* SAC $\lambda(n)$ with bounding function $h(x)$, there holds

$$P\Big\{\sup_{0\leq m<n}\big\{\lambda(n-m) - \Gamma(m,n)\big\} > x\Big\} \leq h(x). \tag{5.4}$$

Remark. Note that the left-hand side of the above inequality does not take $m = n$ in order to ensure the meaning for the following analysis.

The left-hand side of Inequality (5.4) represents an instantaneous property which is generally hard to calculate. To address this difficulty, additional constraint on the bounding function is needed.

Assume $\{\tau_n\}$ are identically distributed. Without loss of generality, assume when m takes m_0, the following holds

$$\begin{aligned} & \sup_{0\le m<n} \big\{\varphi(\eta, n-m)\cdot(n-m) - \Gamma(m,n)\big\} \\ = \;& \varphi(\eta, n-m_0)\cdot(n-m_0) - \Gamma(m_0,n). \end{aligned}$$

Then from Inequality (5.4), we can write, for any $x \ge 0$,

$$\begin{aligned} & P\Big\{ \sup_{0\le m\le n} \big\{\varphi(\eta, n-m)\cdot(n-m) - \Gamma(m,n)\big\} > x\Big\} \\ = \;& P\Big\{\varphi(\eta, n-m_0)\cdot(n-m_0) - \Gamma(m_0,n) > x\Big\} \\ \le \;& e^{-\eta x}\mathbf{E}[e^{\eta[\varphi(\eta,n-m_0)\cdot(n-m_0)-\Gamma(m_0,n)]}] \\ \le \;& e^{-\eta x} \end{aligned}$$

where, the last step is obtained from Inequality (5.3) which is independent of $(n-m_0)$ and holds for any $0 \le m_0 < n$.

Lemma 16. *For an arrival process $\Gamma(m,n)$, if there exists $\varphi(\eta, n-m)$ which satisfies, for any $0 \le m < n$,*

$$\frac{1}{\eta(n-m)} log\boldsymbol{E}\big[e^{\eta\Gamma(m,n)}\big] \ge \varphi(\eta, n-m)$$

then this arrival process has a v.w.d SAC $\lambda(n) = \varphi(\eta,n)\cdot n$ with bounding function $h(x) = e^{-\eta x}$ for $\eta > 0$.

Remark. Lemma 14 and Lemma 16 give the same stochastic arrival curve

$$\varphi(\eta,n) \le \mathbf{E}[e^{\eta[\varphi(\eta,n-m)\cdot(n-m)-\Gamma(m,n)]}]$$

associated with the same bounding function $h(x) = e^{-\eta x}$. Compared with Theorem 10(2), Lemma 16 provides a tighter bound. Such improvement is obtained under the condition that the bounding function in Lemma 14 is independent of the number of packets.

If the arrival process is formed by the *i.i.d.* inter-arrival times, it has a *v.w.d* SAC given as below.

Lemma 17. *If the inter-arrival times of arrival process $\Gamma(m,n)$ are i.i.d., then the arrival process has a v.w.d SAC $\lambda(n) = \varphi(\eta) \cdot n$ with bounding function $h(x) = e^{\eta\varphi(\eta)}\boldsymbol{E}[e^{-\eta\tau_0}]e^{-\eta x}$ for $\eta > 0$, where $\varphi(\eta) \leq \frac{1}{\eta} log \boldsymbol{E}\big[e^{\eta\tau_0}\big]$.*

Proof. In order to prove this lemma, we need to construct a martingale.

Consider a sequence of non-negative random variables $\{V_m\}$, $m = 1, 2, ..., n-1$, formed by

$$V_m = e^{\eta\varphi(\eta)\cdot m - \eta\Gamma(n-m,n)} = e^{\eta\varphi(\eta)\cdot m - \eta\sum_{k=n-m+1}^{n}\tau_k}.$$

Since $\{\tau_k\}$ are *i.i.d.*, we then have

$$\begin{aligned} & V_{m+1} = e^{\eta\varphi(\eta)\cdot(m+1) - \eta\Gamma(n-m-1,n)} \\ = \;& e^{\eta\varphi(\eta)\cdot(m+1) - \eta\sum_{k=n-m}^{n}\tau_k} \\ = \;& e^{\eta\varphi(\eta)\cdot m - \eta\sum_{k=n-m+1}^{n}\tau_k} \cdot e^{\eta\varphi(\eta) - \eta\tau_{n-m}} \\ = \;& V_m \cdot e^{\eta\varphi(\eta) - \eta\tau_{n-m}}. \end{aligned}$$

In addition, there holds:

$$\begin{aligned} & \mathbf{E}\big[V_{m+1}|V_1, ..., V_m\big] = \mathbf{E}\big[V_{m+1}|\tau_n, \tau_{n-1}, ..., \tau_{n-m+1}\big] \\ = \;& \mathbf{E}\big[V_m \cdot e^{\eta\varphi(\eta) - \eta\tau_{n-m}}|\tau_n, ..., \tau_{n-m+1}\big] \\ = \;& \mathbf{E}\big[V_m|\tau_n, ..., \tau_{n-m+1}\big] \cdot \mathbf{E}\big[e^{\eta\varphi(\eta) - \eta\tau_{n-m}}\big] && (5.5) \\ = \;& V_m \cdot \frac{e^{\eta\varphi(\eta)}}{\mathbf{E}\big[e^{\eta\tau_0}\big]} && (5.6) \\ \leq \;& V_m && (5.7) \end{aligned}$$

where, step (5.5) is due to that τ_{n-m} is independent of $\{\tau_n, \tau_{n-1}, ..., \tau_{n-m+1}\}$, step (5.6) is because $\{\tau_1, \tau_2, ...\}$ are identically distributed and

$$\mathbf{E}\big[V_m(\tau_n, \tau_{n-1}, ..., \tau_{n-m+1})|\tau_n, \tau_{n-1}, ..., \tau_{n-m+1}\big] = V_m,$$

and step (5.7) holds since $\mathbf{E}\left[e^{\eta\tau_0}\right] \geq e^{\eta\varphi(\eta)}$ by definition.

Hence $V_1, V_2, ..., V_n$ form a non-negative supermartingale. From Lemma 4, there holds

$$\begin{aligned} & P\Big\{\sup_{1\leq m<n}\left\{\varphi(\eta)\cdot(n-m)-\Gamma(m,n)\right\} > x\Big\} \\ = \; & P\Big\{\sup_{1\leq m<n}\left\{e^{\varphi(\eta)\cdot(n-m)-\Gamma(m,n)}\right\} > e^{x}\Big\} \\ = \; & P\Big\{\sup_{1\leq m<n} V_m > e^{x}\Big\} \\ \leq \; & \mathbf{E}[V_1]e^{-\eta x} \\ = \; & e^{\eta\varphi(\eta)}\mathbf{E}[e^{-\eta\tau_0}]e^{-\eta x} \end{aligned}$$

which ends the proof. □

Remark. Lemma 16 becomes Lemma 17 by taking into account the independence condition. The bounding function in Lemma 17 contains a scaling factor $e^{\eta\varphi(\eta)}\mathbf{E}[e^{-\eta\tau_0}]$ with regard to $\varphi(\eta)$. This scaling factor will yield tighter arrival curves if it is smaller than 1.

Example 9. *Exponential inter-arrival time distribution.*

Consider an arrival process of packets generated at times $\{a(n)\}$. Suppose the inter-arrival times $\{\tau_n\}$ are *i.i.d.* exponentially distributed random variables with mean $\frac{1}{\mu}$. Then the arrival process $\Gamma(m,n)$ follows gamma distribution with parameters $n-m$ and μ. We thus have

$$\begin{aligned} \mathbf{E}\left[e^{\eta\Gamma(m,n)}\right] &= \left(\frac{\mu}{\mu-\eta}\right)^{n-m} \\ \Rightarrow \quad \frac{1}{\eta(n-m)} log\mathbf{E}\left[e^{\eta\Gamma(m,n)}\right] &= \frac{1}{\eta} log\frac{\mu}{\mu-\eta} \end{aligned}$$

Let $\varphi(\eta) = \frac{1}{\eta} log\frac{\mu}{\mu-\eta}$. By applying Lemma 15 and Lemma 17, the *i.a.t* SAC and the *v.w.d* SAC of the arrival process can be obtained. Specifically, this arrival process has a *v.w.d* SAC $\lambda(n)$ with bounding function $h(x)$, where

$$\begin{aligned} \lambda(n) &= \varphi(\eta)\cdot n = \frac{n}{\eta} log\frac{\mu}{\mu-\eta} \\ h(x) &= e^{\eta\varphi(\eta)}\mathbf{E}[e^{-\eta\tau_0}]e^{-\eta x} = e^{-\eta x}. \end{aligned}$$

5.2 Service Process Characterization

This section first reviews the *i.d* SSC model, from which we define a stochastic strict service curve model to facilitate obtaining the *i.d* SSC. Moreover, a time-domain *impairment process* is introduced to decouple the characterization of the cumulative impaired service time from the real service process.

5.2.1 Concretizing Stochastic Strict Service Curve

Consider a network system having the stochastic nature. The following definition describes the service process of this system by comparing the packet actual departure time $d(n)$ with a virtual departure time $a\bar{\otimes}\gamma(n)$, i.e., the *i.d.* SAC.

If a system provides an *i.d* SAC $\gamma(n)$ with bounding function $j(x)$, then there holds for $x \geq 0$,

$$P\Big\{d(n) - a\bar{\otimes}\gamma(n) > x\Big\} \leq j(x). \tag{5.8}$$

Since the above inequality does not explicitly characterize the service process, the stochastic strict service curve (see Definition 15) is introduced to help find the *i.d* SSC. According to Theorem 13 (1), there exists the following relationship:

$$d(n) - a\bar{\otimes}\gamma(n) \leq \Delta(m,n) - \gamma(n-m+1).$$

If the service times $\{\delta_n\}$ are identically distributed, we define

$$\nu(\eta, n-m+1) \geq \frac{1}{\eta(n-m+1)} log\mathbf{E}\big[e^{\eta\Delta(m,n)}\big].$$

And let $\gamma(n) = (\nu(\eta,n) + \eta_\gamma)\cdot n$ for $\eta_\gamma \geq 0$. According to this, The

left-hand side of Inequality (5.8) is rewritten as:

$$\begin{aligned} & P\Big\{d(n) - a\bar{\otimes}\gamma(n) > x\Big\} \\ \leq\ & P\Big\{\Delta(m,n) - \gamma(n-m+1) > x\Big\} \\ \leq\ & P\Big\{e^{\eta[\Delta(m,n)-\gamma(n-m+1)]} > e^{\eta x}\Big\} \\ \leq\ & e^{-\eta x}\mathbf{E}\big[e^{\eta[\Delta(m,n)-(\nu(\eta,n-m+1)+\eta_\gamma)\cdot(n-m+1)]}\big] \\ \leq\ & e^{-\eta x}e^{-\eta\eta_\gamma(n-m+1)} \\ \leq\ & e^{-\eta x}e^{-\eta\eta_\gamma} \end{aligned}$$

where the last step holds due to the fact $e^{-\eta\eta_\gamma} \geq e^{-k\cdot\eta\eta_\gamma}$ for any $k \geq 1$.

If the service times $\{\delta_n\}$ are *i.i.d.*, the stochastic strict service curve $\gamma(n-m+1) = (\nu(\eta)+\eta_\gamma)\cdot(n-m+1)$ with $\eta_\gamma \geq 0$ and $\nu(\eta)$ satisfying

$$\nu(\eta) = \frac{1}{\eta} log\mathbf{E}\big[e^{\eta\delta_0}\big].$$

From the above result, we rewrite the left-hand side of Inequality (5.8):

$$\begin{aligned} & P\Big\{d(n) - a\bar{\otimes}\gamma(n) > x\Big\} \\ \leq\ & P\Big\{\Delta(m,n) - (\nu(\eta)+\eta_\gamma)\cdot(n-m+1) > x\Big\} \\ \leq\ & e^{-\eta x}\mathbf{E}\big[e^{\eta[\Delta(m,n)-(\nu(\eta)+\eta_\gamma)\cdot(n-m+1)]}\big] \\ =\ & e^{-\eta x}\mathbf{E}\big[e^{\eta[\delta_0-(\nu(\eta)+\eta_\gamma)]^{n-m+1}}\big] \\ \leq\ & e^{-\eta x}e^{-\eta\eta_\gamma}. \end{aligned}$$

The following lemma summarizes the above two cases.

Lemma 18. *If the system provides stochastic strict service curve $\gamma(n)$ with bounding function $j(x)$ and*

- *the service times $\{\delta_n\}$ are identically distributed, then $\gamma(n-m+1) = (\nu(\eta, n-m+1)+\eta_\gamma)\cdot(n-m+1)$ and $j(x) = e^{-\eta x}e^{-\eta\eta_\gamma}$ for all $\eta > 0, \eta_\gamma \geq 0$, where*

$$\nu(\eta, n-m+1) \geq \frac{1}{\eta(n-m+1)} log\boldsymbol{E}\big[e^{\eta\Delta(m,n)}\big];$$

- *the service times* $\{\delta_n\}$ *are i.i.d., then* $\gamma(n-m+1) = (\nu(\eta) + \eta_\gamma) \cdot (n-m+1)$ *and* $j(x) = e^{-\eta x} e^{-\eta\eta_\gamma}$ *for* $\eta > 0, \eta_\gamma \geq 0$, *where*

$$\nu(\eta) = \frac{1}{\eta} log \boldsymbol{E}\left[e^{\eta\delta_0}\right].$$

5.2.2 Time-domain Impairment Process Characterization

Networks that are stochastic in nature may be modelled as a stochastic server which consists of an ideal service process and an impairment process. The former describes the service that the server would have delivered in an interval if there have been no service impairment in the interval, and the latter characterizes the service that cannot be delivered in the interval due to some impairment to the server. The impairment process was originally proposed to describe the impaired service from the spatial perspective [61]. In this section, we present characterizing the impaired service from the temporal perspective through investigating a typical example of the stochastic server.

Consider a wireless channel providing service with stochastic nature which is due to some random impairment process. The impairment degrades the network performance because packets may not be transmitted successfully when the channel condition is 'bad' or packets are queued in the buffer until the channel condition becomes 'good'. Either dropping/re-transmitting the unsuccessfully delivered packets or holding the queued packets longer can be counted as impairment in the service time. The cumulative impairment in service time can be explicitly described by an impairment process.

If the channel is always in 'good' condition, the service time of packet $P(n)$ equals the packet transmission time denoted by $\hat{\delta}_n$. However, the varying link condition may cause packet $P(n)$ to suffer additional delay denoted by ε_n. Then the actual service time equals the packet transmission time plus the additional delay, i.e., $\delta_n = \hat{\delta}_n + \varepsilon_n$.

Let

- $\hat{\Delta}(m,n) = \sum_{k=m}^{n} \hat{\delta}_k$ represent the ideal service process without errors;
- $\mathbb{I}(m,n) = \sum_{k=m}^{n} \varepsilon_k$ represent the error process;

- $\Delta(m,n)$ represent the actual service process.

The cumulative actual service time satisfies, for any $0 \leq m \leq n$,

$$\Delta(m,n) = \hat{\Delta}(m,n) + \mathbb{I}(m,n). \tag{5.9}$$

If the ideal service process $\hat{\Delta}(m,n)$ has a (deterministic) strict service curve[1] $\hat{\gamma}(n)$, i.e., $\hat{\Delta}(m,n) \leq \hat{\gamma}(n-m+1)$ for any $0 \leq m \leq n$. Then from (5.9), there holds

$$\Delta(m,n) \leq \hat{\gamma}(n-m+1) + \mathbb{I}(m,n).$$

Furthermore, if the impairment process $\mathbb{I}(m,n)$ has a stochastic strict service curve $\gamma_{\mathbb{I}}(n)$, we get a further result

$$\begin{aligned} & \Delta(m,n) - \hat{\gamma}(n-m+1) - \gamma_{\mathbb{I}}(n-m+1) \\ \leq \; & \mathbb{I}(m,n) - \gamma_{\mathbb{I}}(n-m+1). \end{aligned}$$

The following lemma illustrates that the above-mentioned stochastic server with impairment process provides a stochastic strict service curve.

Lemma 19. *Consider that a stochastic server consists of an ideal service process $\hat{\Delta}(m,n)$ having a (deterministic) strict service curve $\hat{\gamma}(n)$ and an impairment process $\mathbb{I}(m,n)$ having a stochastic strict service curve $\gamma_{\mathbb{I}}(n)$ with bounding function $j_{\mathbb{I}}(x)$. Then, the stochastic server provides a stochastic strict service curve $\gamma(n)$ with bounding function $j_{\mathbb{I}}(x)$, where*

$$\gamma(n) = \hat{\gamma}(n) + \gamma_{\mathbb{I}}(n).$$

Remark. According to Lemma 18, if $\{\varepsilon_n\}$ are identically distributed, then $\gamma_{\mathbb{I}}(n) = (\nu_{\mathbb{I}}(\eta,n)+\eta_\gamma)\cdot n$ with bounding function $j_{\mathbb{I}}(x) = e^{-\eta x}e^{-\eta\eta_\gamma}$, where

$$\nu_{\mathbb{I}}(\eta, n-m+1) \;\geq\; \frac{1}{\eta(n-m+1)} log\mathbf{E}\left[e^{\eta\mathbb{I}(m,n)}\right].$$

[1]The deterministic strict service curve is a special case of the stochastic strict service curve with bounding function $j(x) = 0$.

If $\{\varepsilon_n\}$ are *i.i.d.*, then $\gamma_{\mathbb{I}}(n) = (\nu_{\mathbb{I}}(\eta) + \eta_\gamma) \cdot n$ with the same bounding function $j_{\mathbb{I}}(x)$, where $\nu_{\mathbb{I}}(\eta) = \frac{1}{\eta} log \mathbf{E}\left[e^{\eta \varepsilon_0}\right]$.

Discussion. Recall Definition 7 which characterizes the service process of the space-domain stochastic server with impairment process. Interestingly, in the space-domain, the impairment process is characterized using a stochastic arrival curve $\alpha_I(t)$. That is the impairment process being considered as a 'cross-traffic' which competes the server resource with the concerned arrival process. Whereas, in the time-domain, it is more intuitively to consider the impairment process as a 'delay server' which introduces additional delay between the packet reaching the HOL and the beginning of successfully transmitting the packet. Thus, the impairment process is represented using a stochastic service curve $\gamma_{\mathbb{I}}(n)$.

5.3 Service Curve Example

This section gives an example to demonstrate how to obtain the stochastic service curve characterization of a Gilbert-Elliott channel. We first analyze the constant rate server which can be considered as the ideal service process of a stochastic server. Then we investigate the Gilbert-Elliott channel in detail.

5.3.1 Constant Rate Server

Consider a server with the constant service rate C. Let L_n denote the packet length of packet $P(n)$.

If all packets of the arrival process have the fixed-length L, then the server provides a deterministic strict service curve

$$\gamma(n) = \frac{L}{C} \cdot n.$$

If the packet lengths of the arrival process are *i.i.d.* random variables with the MGF $M_L(\eta) = \mathbf{E}\left[e^{\eta L_0}\right]$, according to Lemma 18 (2), the service process provided to the arrival process is characterized as follows.

Lemma 20. *Consider a server with constant service rate C. If the packet lengths are i.i.d., then the server provides a stochastic strict service curve $\gamma(n) = (\nu(\eta) + \eta_\gamma) \cdot n$ with bounding function $j(x) = e^{-\eta x} e^{-\eta\eta_\gamma}$ for $\eta > 0, \eta_\gamma \geq 0$, where*

$$\nu(\eta) = \frac{1}{\eta C} log \boldsymbol{E}\left[e^{\eta L_0}\right].$$

Example 10.

If the packet lengths follow the exponential distribution with parameter μ, then the server provides the stochastic strict service curve $(\nu(\eta) + \eta_\gamma) \cdot n$ with bounding function $j(x) = e^{-\eta x} e^{-\eta\eta_\gamma}$, where

$$\nu(\eta) = \frac{1}{\eta C} log \frac{\mu}{\mu - \eta}.$$

Example 11.

If the packet lengths are (discrete) uniformly distributed over the range $[A, B]$, then the server provides the stochastic strict service curve $(\nu(\eta) + \eta_\gamma) \cdot n$ with bounding function $j(x) = e^{-\eta x} e^{-\eta\eta_\gamma}$, where

$$\nu(\eta) = \frac{1}{\eta C} log \frac{e^{\eta B} - e^{\eta A}}{\eta(B - A)}.$$

5.3.2 Gilbert-Elliott Channel: Markov Chain Modeling

Consider a time-slotted[2] Gilbert-Elliott ON-OFF communication channel [40] [50] which is modeled by a two-state homogeneous Markov chain. The time (number of time slots) between state transitions is a random variable with a memory-less distribution[3].

In state ON, the channel transmits packets with a constant rate C. In state OFF, the channel does not transmit any packet and thus has

[2] As the slot length approaches *zero*, the service process is approximately continuous.

[3] Strictly speaking, the intervals between state transition are conditionally independent and follow geometric distribution.

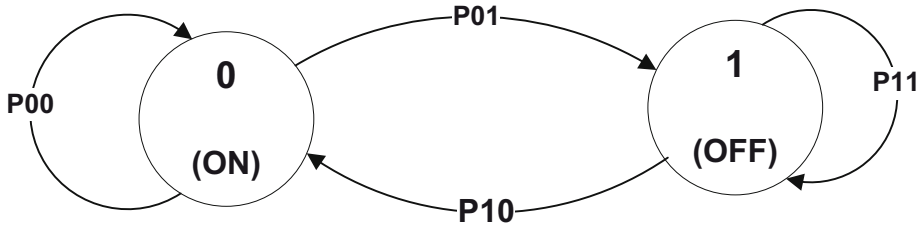

Figure 5.1: Gilbert-Elliott channel model

the transmission rate 0. Here, we assume that when the packet is being transmitted, the channel is always in ON state and does not change to OFF state, i.e., the packet transmission will not be interrupted due to the change of the channel state.

As shown in Figure 5.1, the transition probability from state i to j is denoted by p_{ij}, $i, j = 0, 1$ where 0 represents the 'ON' state and 1 represents the 'OFF' state. The transition probability matrix $\mathbb{P}$ is as follows:

$$\mathbb{P} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}.$$

For this ON-OFF service process, we define the *packet service time* as the interval between the time when a packet reaches HOL and the time when the last bit of this packet has been successfully transmitted. Let $v(n)$ denote the time when a packet reaches the HOL, and be called the *virtual start time* defined by:

$$v(n) = \max[a(n), d(n-1)]. \tag{5.10}$$

The service time δ_n is computed by

$$\delta_n = d(n) - v(n).$$

If the packet reaches the HOL when the channel is in ON state, the packet is transmitted immediately. Otherwise, the packet has to wait until the channel state becomes ON. Let T_n^{Off} denote the OFF interval before packet $P(n)$ can be successfully transmitted. The service time is computed by

$$\delta_n = T_n^{Off} + t_n^{tx}, \tag{5.11}$$

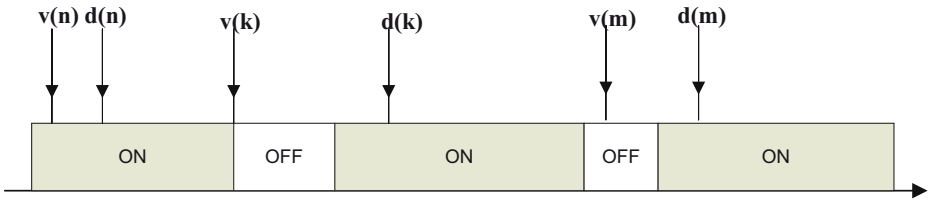

Figure 5.2: On-off model

where t_n^{tx} is the time (number of time slots) of transmitting a packet with length L_n. Assume that the packet lengths are *i.i.d.* and follow some general distribution.

Figure 5.2 shows three scenarios of the packet service time.

- Scenario I: $v(n)$ is within the ON interval, thus

$$\delta_n = t_n^{tx}.$$

- Scenario II: $v(k)$ is the boundary between the ON interval and the OFF interval, thus

$$\delta_k = T_k^{Off} + t_k^{tx}.$$

- Scenario III: $v(m)$ is within the OFF interval, thus

$$\delta_m = T_m^{Off} + t_m^{tx}.$$

Note that T_k^{Off} represents a complete OFF interval while T_m^{Off} denotes the residual of the OFF interval.

For any packet $P(n)$ served in a backlogged period, Scenario III will not happen because $v(n) = d(n-1) + \epsilon$ ($\epsilon \to 0$) and $d(n-1)$ is always in the ON interval. Thus, the service time of packets which are served in backlogged periods equals either t_n^{tx} or $T_n^{Off} + t_n^{tx}$. Since Scenario III takes on value between Scenario I and Scenario II, it is suffice to analyze only Scenario I and Scenario II.

The cumulative service process $\Delta(m, n)$ (in any backlogged period) is formed from process $\{\delta_n^Y\}$ that takes on values $Y \cdot T^{off} + \frac{L_n}{C}$, where

- $Y = 1$ if $v(n)$ is exactly the boundary between the ON interval and the OFF interval;

- $Y = 0$ if $v(n)$ is within the ON interval.

The moment generating function of the *i.i.d.* random variables δ_n^Y are $M_{\delta^Y}(\eta) = \mathbf{E}\big[e^{\eta\delta^Y}\big]$. Let $\mathbb{M}$ be the diagonal matrix

$$\mathbb{M} = \begin{bmatrix} M_{\delta^0}(\eta) & 0 \\ 0 & M_{\delta^1}(\eta) \end{bmatrix}.$$

Given the initial condition $v(0) = i$ $(i = 0, 1)$, from Kolmogorov backward equation, we have

$$\begin{aligned}
& \mathbf{E}\big[e^{\eta\Delta(0,n)}|v(0)=i\big] \\
= \;& \mathbf{E}\big[e^{\eta\delta^Y}|v(0)=i\big]\mathbf{E}\big[e^{\eta(\Delta(0,n)-\delta^Y)}|v(0)=i\big] \\
= \;& M_{\delta^i}(\eta)\sum_{j=0}^{1}\mathbf{E}\big[e^{\eta(\Delta(1,n)}|v(1)=j, v(0)=i\big]\cdot P(v(1)=j|v(0)=i) \\
= \;& M_{\delta^i}(\eta)\sum_{j=0}^{1}\mathbf{E}\big[e^{\eta(\Delta(1,n)-\delta_0)}|v(1)=j\big]p_{ij} \\
= \;& M_{\delta^i}(\eta)\sum_{j=0}^{1}\mathbf{E}\big[e^{\eta\Delta(1,n)}|v(1)=j\big]p_{ij} \qquad (5.12)
\end{aligned}$$

Let

$$\Phi(\eta, n) = \big(\mathbf{E}\big[e^{\eta\Delta(0,n)}|v(0)=0\big], \mathbf{E}\big[e^{\eta\Delta(0,n)}|v(0)=1\big]\big)$$

and $\Phi(\eta, n)^T$ be its transpose. We then rewrite (5.12) in matrix form:

$$\Phi(\eta, n)^T = \mathbb{M}\mathbb{P}\Phi(\eta, n-1)^T. \qquad (5.13)$$

Applying (5.13) to its right-hand side iteratively results in

$$\Phi(\eta, n)^T = \big(\mathbb{M}\mathbb{P}\big)^n\Phi(\eta, 0)^T. \qquad (5.14)$$

The initial condition can be obtained by

$$\Phi(\eta, 0)^T = \mathbb{M}\mathbf{1}^T$$

where $\mathbf{1} = [1 \;\; 1]$ is a vector with two entries being one.

Let π_i be the steady probability at state i and $\psi = [\pi_0 \;\; \pi_1]$. The steady probability at state i are computed by

$$\pi_0 = \frac{p_{10}}{2 - p_{00} - p_{11}}, \quad \pi_1 = \frac{p_{01}}{2 - p_{00} - p_{11}}.$$

Then we have

$$\begin{aligned}\mathbf{E}\big[e^{\eta\Delta(0,n)}\big] &= \psi\Phi(\eta,n)^T \\ &= \psi\big(\mathbb{M}\mathbb{P}\big)^n\mathbb{M}\mathbf{1}^T. \end{aligned} \tag{5.15}$$

Let $\rho(\cdot)$ denote the spectral radius of a matrix:

$$\rho(\cdot) = \sup\{|\alpha| : \alpha \in \sigma(\cdot)\}$$

where $|\cdot|$ denotes the absolute value of α, and $\sigma(\cdot)$ represents the set of all eigenvalues of a matrix. Then the spectral radius of matrix $\mathbb{M}\mathbb{P}$ is denoted by $\rho\big(\mathbb{M}\mathbb{P}\big)$. Note that $\mathbb{M}\mathbb{P}$ is a non-negative matrix. Having known the transition probability matrix $\mathbb{P}$, the spectral radius of $\mathbb{M}\mathbb{P}$ is

$$\rho(\mathbb{M}\mathbb{P}) = \frac{p_{00}M_{\delta^0}(\eta) + p_{11}M_{\delta^1}(\eta) + \sqrt{Z}}{2} \tag{5.16}$$

where

$$Z = \big(p_{00}M_{\delta^0}(\eta) - p_{11}M_{\delta^1}(\eta)\big)^2 + 4p_{01}p_{10}M_{\delta^0}(\eta)M_{\delta^1}(\eta).$$

A useful corollary (see Corollary 5.6.13 [57]) is introduced here to facilitate the following analysis.

Corollary 1. *Let $\mathbb{A}$ be an $k \times k$ matrix and $\epsilon > 0$ be given. There is a constant σ_ϵ[4] such that*

$$|(\mathbb{A}^m)_{ij}| \leq \sigma_\epsilon(\rho(\mathbb{A}) + \epsilon)^m \tag{5.17}$$

for all $m = 1, 2, 3, ...$ and $i, j = 1, 2, ..., k$.

From Corollary 1, we know that the every entry of matrix $\big(\mathbb{M}\mathbb{P}\big)^n$ is bounded above by $\sigma_\epsilon\big(\rho(\mathbb{M}\mathbb{P}) + \epsilon\big)^n$ for any $\epsilon > 0$ and some constant $\sigma_\epsilon > 0$.

Then (5.15) is bounded by

$$\begin{aligned}\mathbf{E}\big[e^{\eta\Delta(0,n)}\big] &\leq \psi\sigma_\epsilon\big(\rho(\mathbb{M}\mathbb{P}) + \epsilon\big)^n\mathbb{M}\mathbf{1}^T \\ &= \sigma_\epsilon\big(\rho(\mathbb{M}\mathbb{P}) + \epsilon\big)^n\psi\mathbb{M}\mathbf{1}^T \\ &= \sigma_\epsilon\big(\rho(\mathbb{M}\mathbb{P}) + \epsilon\big)^n \\ &\leq \big(\rho(\mathbb{M}\mathbb{P}) + \epsilon\big)^{n+1} \end{aligned} \tag{5.18}$$

[4]This parameter is relevant to ϵ.

for $\sigma_\epsilon \leq \rho(\mathbb{MP}) + \epsilon$. Here step (5.18) is obtained because $\psi\mathbb{M} = \psi$ and then $\psi\mathbf{1}^T = 1$. Hence, we have

$$\frac{1}{\eta(n+1)} log\mathbf{E}\left[e^{\eta\Delta(0,n)}\right] \quad \leq \quad \frac{1}{\eta} log(\rho(\mathbb{MP}) + \epsilon).$$

As ϵ is arbitrary, letting $\epsilon \to 0$ results in the stochastic strict service curve $(\nu(\eta) + \eta_\gamma) \cdot (n - m + 1)$ for the service process $\Delta(m, n)$ with the bounding function $j(x) = e^{-\eta x} e^{-\eta\eta_\gamma}$, where

$$\nu(\eta) = \frac{1}{\eta} log\rho(\mathbb{MP}). \tag{5.19}$$

Example 12.

Consider a flow of variable-length packets. Suppose the packet lengths are *i.i.d.* variables with moment generating function $M_L(\eta)$. The OFF intervals follow geometric distribution with parameter π_0[5]. Then

$$M_{\delta^0}(\eta) = \big(M_L(\eta)\big)^{\frac{1}{C}}, \quad M_{\delta^1}(\eta) = M_{\delta^0}(\eta)\frac{\pi_0 e^\eta}{1 - \pi_1 e^\eta}$$

for $\pi_1 e^\eta < 1$. Inserting $M_{\delta^0}(\eta)$ and $M_{\delta^1}(\eta)$ into (5.16), we obtain the stochastic service curve $(\nu(\eta) + \eta_\gamma) \cdot n$ with bounding function $j(x) = e^{-\eta x} e^{-\eta\eta_\gamma}$ for the flow, where

$$\nu(\eta) = \frac{1}{\eta} log \frac{M_{\delta^0}(\eta)\left[p_{00} + \frac{p_{11}\pi_0 e^\eta}{1-\pi_1 e^\eta} + \Upsilon\right]}{2} \tag{5.20}$$

with

$$\Upsilon = \sqrt{\left(p_{00} - \frac{p_{11}\pi_0 e^\eta}{1 - \pi_1 e^\eta}\right)^2 + \frac{4p_{10}p_{01}\pi_0 e^\eta}{1 - \pi_1 e^\eta}}. \tag{5.21}$$

Example 13.

Consider a flow consisting of fixed-length packets. The OFF intervals follow geometric distribution with parameter π_0. Let the packet

[5]Suppose that the state of time slots are independent trail which are performed until the first slot being in ON state (i.e., success). Each slot has probability π_0 of being in ON state. Let X be the number of time slots needed until the first slot being in ON state. Then X is said to be a *geometric* random variable with parameter π_0 [91].

transmission time be T time slots. Similar to Example 12, we have the moment generating functions of δ^i, $i = 0, 1$, as below:

$$M_{\delta^0}(\eta) = e^{\eta T}, \quad M_{\delta^1}(\eta) = e^{\eta T}\frac{\pi_0 e^{\eta}}{1-\pi_1 e^{\eta}}$$

for $\pi_1 e^{\eta} < 1$.

Inserting $M_{\delta^0}(\eta)$ and $M_{\delta^1}(\eta)$ into (5.16), we have the stochastic service curve $(\nu(\eta) + \eta_\gamma) \cdot n$ with bounding function $j(x) = e^{-\eta x}e^{-\eta\eta_\gamma}$ for such flow, where

$$\begin{aligned} \nu(\eta) &= \frac{1}{\eta} log \frac{e^{\eta T}\left[p_{00} + \frac{p_{11}\pi_0 e^{\eta}}{1-\pi_1 e^{\eta}} + \Upsilon\right]}{2} \\ &= T + \frac{1}{\eta} log \frac{p_{00} + \frac{p_{11}\pi_0 e^{\eta}}{1-\pi_1 e^{\eta}} + \Upsilon}{2} \end{aligned} \tag{5.22}$$

where Υ is given in (5.21).

5.3.3 Gilbert-Elliott Channel: Impairment Process Analysis

The previous subsection is based on directly analyzing backlogged periods. In this subsection, we adopt an intuitive way which models the channel OFF intervals as an impairment process. The channel is treated as a stochastic server consisting of an ideal service process and an impairment process.

The ideal service process provides service at a constant rate C. The impairment process is described as an ON-OFF process. As shown in Figure 5.3, the channel ON state corresponds to the impairment OFF state and the channel OFF state corresponds to the impairment ON state. When the channel is in impairment ON state, the impairment process provides service time ε_n following geometric distribution with parameter π_0. When the channel is in impairment OFF state, the impairment process does not provide service, i.e., the service time is zero.

The transition probability matrix is the same as $\mathbb{P}$. However, the diagonal matrix $\mathbb{M}_{\mathbb{I}}$ is $diag\{M_{\mathbb{I},0}(\eta), M_{\mathbb{I},1}(\eta)\}$ where $M_{\mathbb{I},0}(\eta) = 1$ and $M_{\mathbb{I},1}(\eta) = \frac{\pi_0 e^{\eta}}{1-\pi_1 e^{\eta}}$, i.e.,

$$\mathbb{M}_{\mathbb{I}} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{\pi_0 e^{\eta}}{1-\pi_1 e^{\eta}} \end{bmatrix}.$$

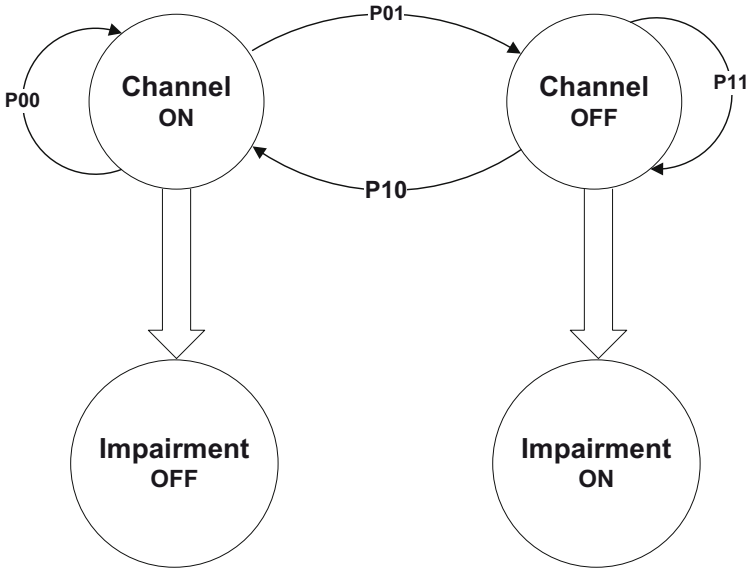

Figure 5.3: Impairment Process On-off model

The impairment process provides a stochastic strict service curve

$$\gamma_{\mathbb{I}}(n) = (\frac{1}{\eta} log\rho(\mathbb{M}_{\mathbb{I}}\mathbb{P}) + \eta_{\gamma}) \cdot n$$

with bounding function $j_{\mathbb{I}}(x) = e^{-\eta x} e^{-\eta\eta_{\gamma}}$, where $\rho(\mathbb{M}_{\mathbb{I}}\mathbb{P})$ is obtained from (5.16):

$$\rho(\mathbb{M}_{\mathbb{I}}\mathbb{P}) = \frac{p_{00} + \frac{p_{11}\pi_0 e^{\eta}}{1-\pi_1 e^{\eta}} + \Upsilon}{2} \tag{5.23}$$

with Υ given in (5.21).

By taking into account the ideal service process $\hat{\gamma}(n)$, the stochastic server thus provides a stochastic strict service curve $\gamma(n) = \hat{\gamma}(n) + \gamma_{\mathbb{I}}(n)$.

Example 14.

Consider the same flow given in Example 12. The packet lengths are *i.i.d.* random variables with moment generating function $M_L(\eta)$. Then the ideal service process provides a (deterministic) strict service curve $\hat{\gamma}(n)$:

$$\hat{\gamma}(n) = \frac{1}{\eta} log\mathbf{E}\big[e^{\eta\frac{L_0}{C}}\big].$$

Combining the stochastic strict service curve provided by the impairment process, i.e., Eq.(5.23), then the Gilbert-Elliott channel essentially provides a stochastic strict service curve $(\nu(\eta)+\eta_\gamma)\cdot n$ with bounding function $j(x)=e^{-\eta x}e^{-\eta\eta_\gamma}$, where

$$\begin{aligned}\nu(\eta) &= \frac{1}{\eta}log\mathbf{E}\left[e^{\eta\frac{L_0}{C}}\right]+\frac{1}{\eta}log\rho(\mathbb{M}_{\mathbb{I}}\mathbb{P})\\ &= \frac{1}{\eta}log\frac{\mathbf{E}\left[e^{\eta\frac{L_0}{C}}\right]\left(p_{00}+\frac{p_{11}\pi_0e^{\eta}}{1-\pi_1e^{\eta}}+\Upsilon\right)}{2}\end{aligned}$$

which matches the result of Example 12.

Example 15.

Consider the same flow given in Example 13. The packets have the same length L and the corresponding transmission time is T slots. Then the Gilbert-Elliott channel provides a stochastic strict service curve $(\nu(\eta)+\eta_\gamma)\cdot n$ with bounding function $j(x)=e^{-\eta x}e^{-\eta\eta_\gamma}$, where

$$\nu(\eta) = T+\frac{1}{\eta}log\frac{p_{00}+\frac{p_{11}\pi_0e^{\eta}}{1-\pi_1e^{\eta}}+\Upsilon}{2} \tag{5.24}$$

which matches the result of Example 13.

Remark. By comparing the above two examples to Example 12 and Example 13, we notice that the backlogged period analysis and the impairment process analysis yield the same results for analyzing the two-state Gilbert-Elliott channel. However, these two analysis methods may not yield the same results in general.

5.4 Stochastic Delay Bound

In the previous sections, we have introduced applying the logarithmic moment generating function to finding the stochastic arrival curve and the stochastic service curve for several arrival processes and service processes. If the stochastic arrival curve of an arrival process is known and the service process provided to the arrival process can be characterized by a stochastic service curve, we readily obtain the delay bound based on the results of Section 4.1.1.

5.4.1 Delay Bound Analysis

Recall two delay bounds derived under the condition that the stochastic arrival curve $\lambda(n)$ with bounding function $h(x)$ and the stochastic service curve $\gamma(n)$ with bounding function $j(x)$ are known.

From Theorem 14, the system delay is bounded by:

$$P\{D(n) > x\} \leq j \otimes h\big([x - \gamma \oslash \lambda(1)]^+\big). \tag{5.25}$$

If the arrival process $\Gamma(m, n)$ is independent of the service process $\Delta(m, n)$, from Lemma 8, the stochastic delay bound is given below

$$P\{D(n) > x\} \leq 1 - \int_0^{x^*} \big(1 - j(x^* - z)\big) d(1 - h(z)) \tag{5.26}$$

where $x^* = [x - \gamma \oslash \lambda(1)]^+$.

Remark. Inequality (5.25) holds with the requirement that the arrival process has a $v.w.d$ stochastic arrival curve and the service process has an $i.d$ stochastic service curve. Moreover, Inequality (5.26) implies that at least one of $h(x)$ and $j(x)$ should be integrable.

Example 16.

Consider a flow of fixed-length packets, of which the inter-arrival times follow the exponential distribution with mean $\frac{1}{\mu}$. Packets of this flow arrive to a wireless node and are queued in the buffer before they are transmitted over a Gilbert-Elliott On-Off channel as given in Example 15. Considering the assumption that when a packet is being transmitted, the channel state will not change to OFF, we set the time slot length to one packet transmission time.

From Example 9, the arrival process has a $v.w.d$ SAC:

$$\lambda(n) = \frac{n}{\eta} log \frac{\mu}{\mu - \eta}, \quad h(x) = e^{-\eta x}.$$

From Example 15, the service process has a stochastic strict service curve:

$$\begin{aligned} \gamma(n) &= (1 + \frac{1}{\eta} log \frac{p_{00} + \frac{p_{11}\pi_0 e^{\eta}}{1 - \pi_1 e^{\eta}} + \Upsilon}{2} + \eta_\gamma) \cdot n, \\ j(x) &= e^{-\eta x} e^{-\eta \eta_\gamma}. \end{aligned}$$

In order to ensure system stability, we know

$$\gamma \oslash \lambda(1) = \gamma(1).$$

Then according to Inequality (5.25), the delay that a packet experiences in this system is stochastically bounded by

$$P\{D(n) > x\} \leq \inf_{0<\eta\leq\eta^*} \inf_{0\leq z\leq x^*} \left[e^{-\eta\eta_\gamma} e^{-\eta z} + e^{-\eta(x^*-z)}\right] \tag{5.27}$$

where $x^* = [x - \gamma(1)]^+$ and η^* is the maximal value that η can take under $\pi_1 e^{\eta} < 1$.

If the arrival process is independent of the service process, according to Inequality (5.26), the stochastic delay bound is given by

$$P\{D(n) > x\} \leq \inf_{0<\eta\leq\eta^*} \left[e^{-\eta x^*} + x^*\eta e^{-\eta(\eta_\gamma + x^*)})\right], \tag{5.28}$$

where x^* is the same as that in (5.27).

5.4.2 Comparison between Spatial and Temporal Analysis

As introduced in Section 2.4.3, the available stochastic network calculus literature on performance guarantee analysis focuses on the spatial perspective. In this subsection, we derive the system delay bound using the spatial analysis approach. For ease of exposition, the arrival process given in Example 9 and the network system given in Example 15 are adopted here.

The spatial approach characterizes the arrival process based on the cumulative amount of arrival traffic (in number of arrival packets) and the service process based on the cumulative amount of service (in number of served packets). Accordingly, the *space-domain* stochastic arrival curve and the *space-domain* stochastic service curve are the bounds on the cumulative amount of traffic and service, respectively.

Recall that $\mathcal{A}(t)$ and $\alpha(t)$ respectively denote the space-domain arrival process and its arrival curve which is associated with the bounding function $f(y)$. And $\mathcal{S}(t)$ and $\beta(t)$ respectively denote the space-domain service process and its service curve which is associated with the bounding function $g(y)$.

Example 17.

The arrival process given in Example 9 is a compound Poisson process from the spatial perspective. Let L be the packet length for all packets. The arrival rate of the Poisson process is μ. This compound Poisson arrival process has the stochastic arrival curve as below [65], for $t \geq 0$ and $\theta > 0$

$$\frac{1}{\theta t} log\mathbf{E}\big[e^{\theta \mathcal{A}(t)}\big] = \frac{\mu}{\theta}(e^{\theta L} - 1), \tag{5.29}$$

from which, the compound Poisson process has a space-domain $v.b.c$ stochastic arrival curve $\alpha(t)$:

$$\alpha(t) = \Big(\frac{\mu}{\theta}(e^{\theta L} - 1) + \theta_\alpha\Big) \cdot t, \quad f(y) = e^{-\theta\theta_\alpha} e^{-\theta y},$$

for $\theta_\alpha \geq 0$.

For the Gilbert-Elliott ON-OFF channel, its space-domain stochastic service curve is the variation of the ON-OFF model's envelop process (see [21]), for all $t \geq 0$ and $\theta > 0$,

$$\beta(t) = \frac{t}{\theta} log\Big(\frac{p_{11} + p_{00} e^{\frac{1}{T}\theta} + \mathbf{Y}}{2}\Big), \quad g(y) = e^{-\theta y},$$

where

$$\mathbf{Y} = \sqrt{(p_{11} + p_{00} e^{\frac{1}{T}\theta})^2 - 4(p_{11} + p_{00} - 1) e^{\frac{1}{T}\theta}}.$$

Here $1/T$ represents the channel transmission rate (number of packets) in the 'ON' state due to the time-slotted channel with the slot length T.

Recall that the maximum horizontal distance between functions $\alpha(t)$ and $\beta(t)$ is $h(\alpha, \beta)$ (see Definition 8):

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

which can be considered as the maximal system delay of a virtual system, where the arrival process is $\alpha(t)$ and the service process is $\beta(t)$.

The system delay of the traffic arriving at time $t \geq 0$ is bounded by (see Theorem 4):

$$\begin{aligned} & P\Big\{\mathcal{D}(t) > h(\alpha + y, \beta)\Big\} \\ \leq \; & f \otimes g(y) \\ = \; & \inf_{0 < \theta \leq \theta^*} \inf_{0 \leq z \leq y} \big[e^{-\theta\theta_\alpha} e^{-\theta z} + e^{-\theta(y-z)}\big], \end{aligned} \tag{5.30}$$

where θ^* is the maximum meaningful value of θ.

Similar to Inequality (5.26), if the arrival process $\mathcal{A}(t)$ is independent of the service process $\mathcal{S}(t)$, the system delay is stochastically bounded by the Stieltjes convolution of $f(y)$ and $g(y)$:

$$\begin{aligned} & P\Big\{\mathcal{D}(t) > h(\alpha+y,\beta)\Big\} \\ \leq\ & 1-\int_0^y (1-g(y-z))d(1-f(z)) \\ =\ & \inf_{0<\theta\leq\theta^*} e^{-\theta y}+\theta y e^{-\theta\theta_\alpha}e^{-\theta y}. \end{aligned} \tag{5.31}$$

Remark. Although the bounding functions obtained by applying the spatial approach (see Inequalities (5.30) and (5.31)) look very similar as those obtained from the temporal analysis (see Inequalities (5.27) and (5.28)), their arguments have different meanings. To compute the delay bound, the spatial approach uses the amount of traffic denoted by y as the argument, while the temporal approach uses the time denoted by x (or x^*) as the argument. Since the bounding functions of both approaches are negative exponential functions, a larger argument yields a smaller result and vice versa.

5.4.3 Numerical Results

The available literature [67] provides a simple example to illustrate that by considering the independence condition, the tightness of the delay bound may be improved, i.e., Inequality (5.26) may provide a tighter bound compared to Inequlity (5.25). In order to intuitively illustrate these two bounds, we use Matlab to numerically compute the two bounds derived in Example 16 (see Inequalities (5.27) and (5.28)). Then, we investigate how the optimal parameter η_γ impact the delay bound. Moreover, the system delay bounds obtained by the temporal and the spatial approaches are compared.

The Gilbert-Elliott channel provides $C = 2Mbps$ capacity when it is in the ON state. All packets have the same length 250 bytes. Hence the packet transmission time $T = 1msec$ which is the time slot length. The transition probabilities between ON and OFF states hold such relationship $p_{10}/p_{01} = 3$, from which, we calculate

$$\pi_0 = 0.75, \quad \pi_1 = 0.25, \quad p_{00} = (2+p_{11})/3.$$

If we set $p_{00} = 0.95$, the corresponding transition probabilities are

$$p_{01} = 0.05, \quad p_{10} = 0.15, \quad p_{11} = 0.85.$$

Moreover, according to $\pi_1 = 0.25$, we obtain the maximal value of η, $\eta^* = 1.386$ in terms of $\pi_1 e^{\eta} < 1$.

In the following figures, we use **Bound 1** and **Bound 2** to represent the bounds given in Inequalities (5.27) and (5.28), respectively.

As shown in Figure 5.4, Bound 1 is looser than Bound 2 under the same condition, i.e., the same arrival process and the same network system. This result implies that by considering the independence condition, the bound may be improved. The same phenomenon has been discussed in [67].

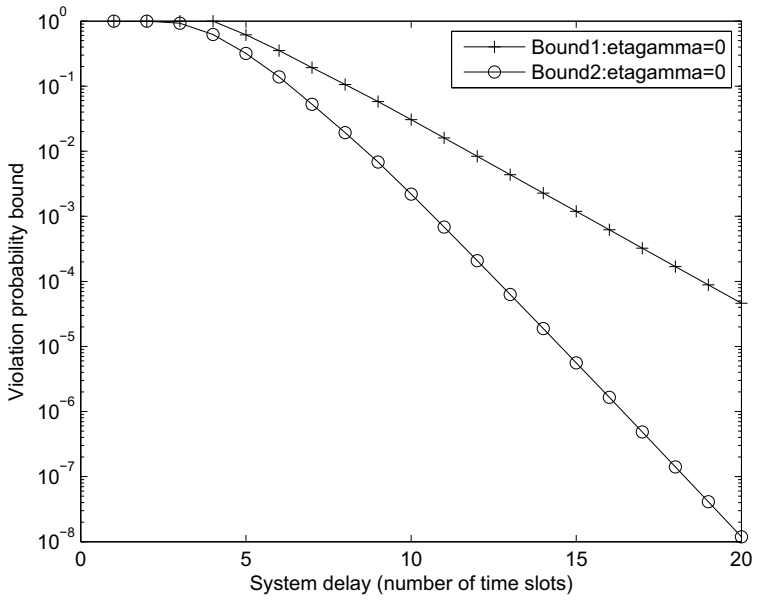

Figure 5.4: Bound Comparison: $\eta_\gamma = 0$

Since both Inequalities (5.27) and (5.28) contain the scaling factor, $e^{-\eta\eta_\gamma}$, how does this *scaling parameter* impact the delay bound?

From the definition of η_γ, it should be set in terms of $\nu(\eta)$. Figure 5.5 shows Bound 1 when η_γ takes 0, $0.5\nu(\eta)$ and $0.6\nu(\eta)$. The bound obtained by setting $\eta_\gamma = 0.5\nu(\eta)$ is tighter than that obtained by setting $\eta_\gamma = 0$. However, when η_γ exceeds $0.5\nu(\eta)$ such as $0.6\nu(\eta)$, the

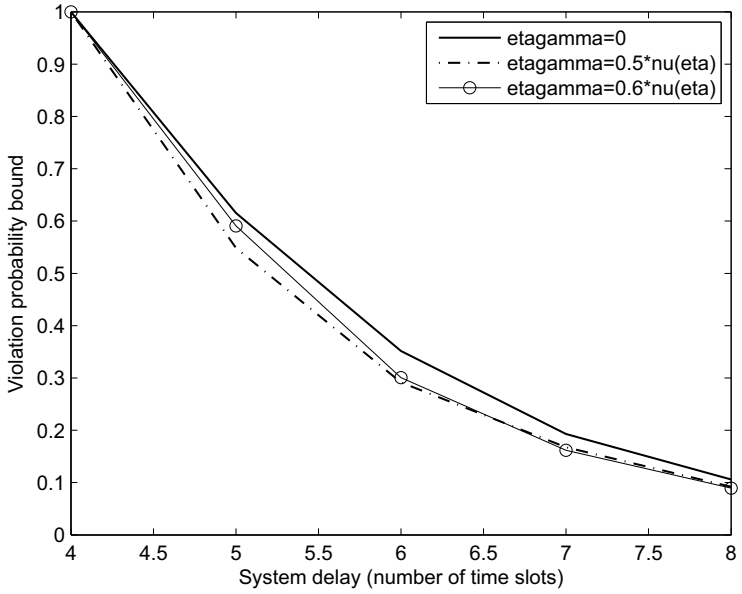

Figure 5.5: Bound 1 vs. Varying η_γ

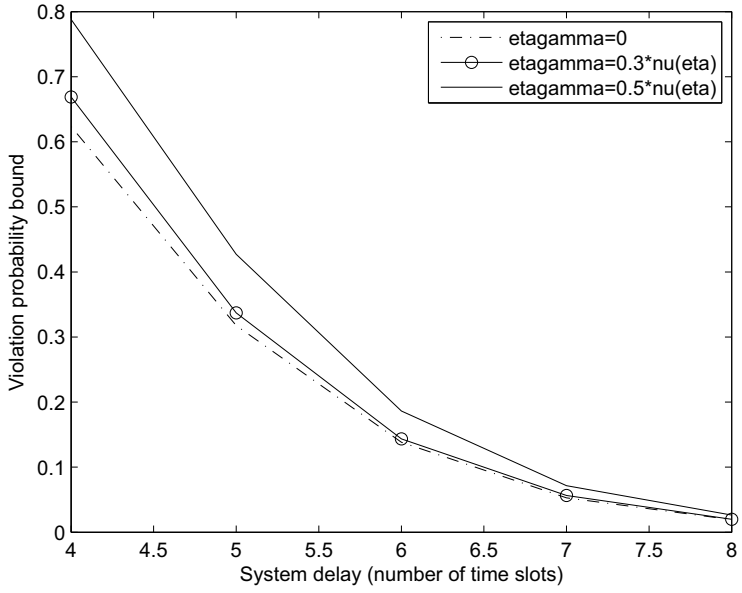

Figure 5.6: Bound 2 vs. Varying η_γ

bound becomes more loose than that obtained by setting $\eta_\gamma = 0.5\nu(\eta)$. Thus, $\eta_\gamma = 0.5\nu(\eta)$ is the optimal value.

Similarly, Figure 5.6 shows Bound 2 against various η_γ. As η_γ increases from 0 to $0.3\nu(\eta)$ or $0.5\nu(\eta)$, the bound becomes looser. This is because the increment of η_γ results in the decrement of y in Inequality (5.28) for a fixed x. Moreover, taking the infimum impacts the final result as well.

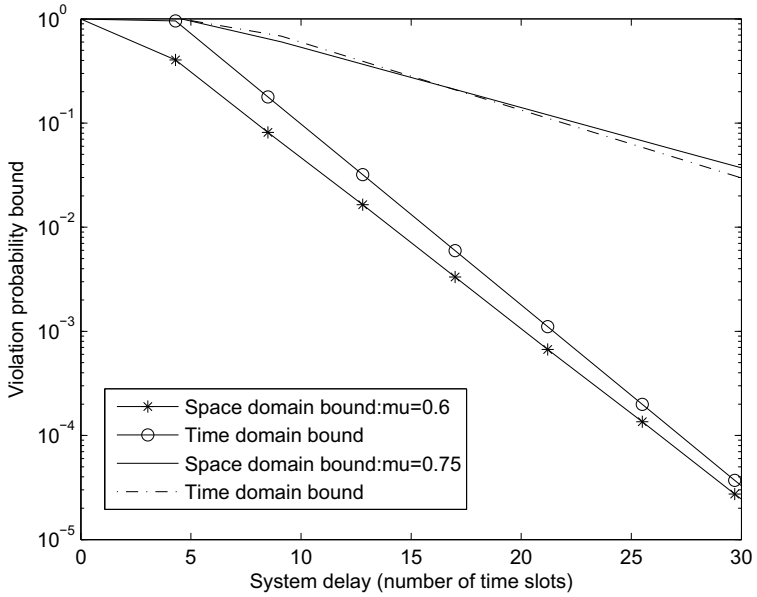

Figure 5.7: Space and time domain bound comparison vs. varying μ ($\theta^* = \eta$, $\eta_\gamma = 0$ and $\theta_\alpha = 0$)

Figure 5.7 shows the bounds obtained from the spatial and the temporal approaches. When computing the space-domain delay bound according to Inequality (5.30), we need to first compute the space-domain arrival curve $\alpha(t)$ and the service curve $\beta(t)$ under an implicit relation, $\alpha(t) \leq \beta(t)$, which ensures system stability. With this condition and varying the arrival rate μ of the Poisson process, we can determine the maximal meaningful value of θ, θ^*. As the amount of traffic y varies, the bound on the probability that the system delay exceeds $h(\alpha + y, \beta)$ can be obtained with θ^*.

Let $\eta = \theta^*$ and $x = h(\alpha + y, \beta)$ when computing the time-domain delay bound according to Inequality (5.27). In Figure 5.7, $\mu = 0.6$

or $\mu = 0.75$ means the arrival rate per time slot. As μ increases, more packets arrive and then the system delay becomes longer. Thus, fixing a certain time and observing the bounds obtained by setting $\mu = 0.6$ and $\mu = 0.75$, the trend is that a smaller μ causes a tighter bound. When $\mu = 0.75$, the space-domain bound and the time-domain bound are very close. However, when $\mu = 0.6$, the space-domain bound is tighter than the time-domain bound. The reason is that under the current parameter setting, the amount of traffic y used to compute the space-domain bound is larger than the time $h(\alpha+y, \beta) - \gamma(1)$ used to compute the time-domain bound. A larger argument yields a smaller result for the negative exponential functions as we have discussed in Section 5.4.2.

From Figure 5.7, we notice that the spatial and the temporal approaches give close results. However, how the individual parameters of bounding functions influence the final result still needs more investigation.

5.5 Conclusion

This chapter concretizes the temporal network calculus approach for performance guarantee analysis of stochastic networks. A key technique used in linking an arrival process or a service process to the time-domain stochastic arrival curve characterization or stochastic service curve characterization is Moment Generating Function. Based on the arrival process characterization and the service process characterization, performance bounds such as delay bound can be further derived from the temporal stochastic network calculus.

Moreover, the Gilbert-Elliott channel is particularly investigated to demonstrate how to obtain the MGF of the service process. The numerical results show that the delay bound is improved by taking into consideration the independence between the arrival process and the service process. Finally, we illustrate that the temporal and the spatial analysis approaches give close performance bounds under the appropriate match between the arguments used in both approaches.

Chapter 6

Application Case: IEEE 802.11 Delay Evaluation

The material in this chapter has been partially published as follows:

Jing Xie and Yuming Jiang. "A Network Calculus Approach to Delay Evaluation of IEEE 802.11 DCF." In *Proceedings of the* 35^{th} *IEEE Conference on Local Computer Networks (LCN)*, Denver, USA, October 2010.

6.1 Introduction

The time-domain stochastic service curves are defined to characterize the service process of networks which contain inherently stochastic factors. Wireless networks are one typical representative of such networks. Some work on applying stochastic network calculus to service analysis of wireless networks from the spatial perspective has been reported [16] [46] [66]. In this chapter, the temporal approach will be the analytical basis. The central idea is to compare the service process with a virtual time function [67] as used by the deterministic GR server model [53]. GR server model is an important model in defining the two Internet QoS architectures: Integrated Services (IntServ) and Differentiated Services (DiffServ) architectures [67].

The IEEE 802.11 wireless local area networks (WLANs) [1] [25] [26] have been widely deployed to provide low-cost broadband wireless Internet access [48]. To understand the potential of IEEE 802.11 for supporting real-time applications, it is important to evaluate the delay characteristics at the MAC layer.

The fundamental mechanism to access the medium in IEEE 802.11 networks is the distributed coordination function (DCF). The available literature on delay evaluation of the DCF mainly focuses on investigating the packet service time under either the saturated condition [92] or the non-saturated condition [104]. The impact of the arrival pattern on the delay performance is still open. However, the various arrival patterns may significantly influence delay performance and should be taken into account when investigating the delay performance. The *objective* of this chapter is to take one step towards evaluating delay performance for various arrival processes through stochastic network calculus.

This chapter is based on modeling a saturated, single cell network with an **ideal channel condition** (without capture, fading or frame error), where a finite number of homogeneous stations (STAs) contend for a shared wireless channel. Each STA always has packets available for transmission[1]. Assume that **all packets have the same length**.

The per-packet service time of the DCF is defined as the interval between the instant when the packet reaches the HOL and the instant when the packet is successfully received at the destination STA. The

[1]This implies that the analysis in this chapter applies to backlog period analysis.

stochastic service curve essentially characterizes the service received by a concerned STA[2] in its backlogged periods. The stochastic service curve derived under the saturated condition implies a probabilistically guaranteed service time for packets transmitted from the concerned STA under the maximal load that can be carried by the network.

6.2 IEEE 802.11 Distributed Coordination Function

The DCF incorporates the CSMA/CA protocol and the **truncated binary exponential backoff** (BEB) scheme. The DCF includes two implementation mechanisms, the default *basic access* and the optional *request-to-send/clear-to-send* (RTS/CTS). The RTS/CTS mechanism can reduce the collision duration and the system degradation due to the hidden terminal problem. However, this mechanism increases overhead for transmitting short data packets and should not be used for every data packet transmission [1]. To comprehensively study the DCF, we investigate both the basic access and the RTS/CTS mechanisms.

In a network employing the CSMA/CA protocol, each STA having a packet to transmit should sense the channel to determine if another STA is transmitting. If the channel is sensed idle for an interval greater than the *distributed interframe space* (DIFS), the STA proceeds to transmission. If the channel is sensed busy, the STA defers transmission and keeps sensing until the channel is sensed idle for a DIFS. Then the STA generates a random backoff interval for an additional deferred time before transmitting.

Backoff intervals are slotted. SATs are only allowed to transmit at the beginning of a slot. When the backoff timer is initiated, a backoff interval (in slots) is uniformly chosen in the range $\{0, 1, ..., CW_k - 1\}$ where CW_k is the contention window of the kth backoff stage. At the first transmission attempt, CW is set to the minimum contention window (CW), CW_{min}. The backoff counter is decremented by 1 after one idle slot elapses. When the channel becomes busy during the countdown, the counter is *frozen* and reactivated when the channel is sensed idle more than one DIFS again. Such intervals when the channel becomes busy are called *inter-transmissions* [11].

[2]In the homogeneous network, all STAs equally receive the service.

The STA attempts to transmit packet when its countdown finishes. A collision occurs when the counter of two or more STAs reach zero in the same slot. Then all the involved STAs have to wait another backoff interval until the next attempt. The new contention window doubles the previous contention window. Doubling the contention window stops when the current contention window CW_k reaches the maximal value $2^M CW_{min}$, where M is the doubling limit. If the packet has not been transmitted successfully after M retransmissions, the contention window is kept at $2^M CW_{min}$ for the following attempts until the packet is transmitted successfully, or until the retransmissions reach the maximum number K specified in [1]. If the packet is still not transmitted successfully after K retransmissions, it is discarded. After the packet is successfully received, the destination STA waits for a short inter-frame space (SIFS) interval and then immediately transmits an acknowledgement (ACK) to the source STA.

The RTS/CTS mechanism transmits the RTS and CTS control packets prior to data packet transmissions. An STA senses the channel before transmitting RTS packet. If the channel is idle during the DIFS interval, this STA starts to transmit the RTS packet; otherwise, the STA retains the RTS packet and keeps sensing until the channel becomes idle. This procedure is the same as the basic access mechanism. A successful exchange of RTS and CTS packets reserves the channel for transmitting data packets. Such reservation can reduce the bandwidth loss since collisions occur only when transmitting RTS packets.

6.2.1 Decoupling Approximation

In this chapter, we only consider the case that all the STAs have the same backoff parameters, i.e., the **homogeneous case**. The **decoupling approximation** [75] effectively facilitates the backoff process analysis and thus is adopted here to conduct the following analysis.

Let N denote the number of contending STAs. For a concerned STA, the decoupling approximation assumes that the aggregate attempt process of the other $N-1$ STAs is independent of the backoff process of the concerned STA.

The key approximation introduced by Bianchi [13] is that the collision probability for each packet is constant and independent regardless of the number of retransmissions. Let p_c denote the collision proba-

bility that at least one of the $N-1$ non-concerned STAs transmit in the same slot. The attempt probability of an STA transmitting in a random slot, denoted by p_a, is constant and independent of the backoff stage [13]. Given the number of STAs, p_c can be expressed in terms of p_a and μ_k which denotes the mean backoff interval at the kth backoff stage [28]:

$$\begin{aligned} p_c &= 1 - e^{-(N-1)p_a}, \\ p_a &= \frac{\sum_{k=0}^{K} p_c^k}{\sum_{k=0}^{K} \mu_k p_c^k}. \end{aligned} \tag{6.1}$$

6.3 Stochastic Characteristics of The Service Time

This section reviews related results on per-packet service time [92], which will be used in the later analysis.

The concerned STA can be considered as an FIFO scheduler and the shared wireless channel can be modeled as a stochastic delay server as shown in Figure 6.1. The packet first passes through the FIFO queue and then enters the stochastic server, where the packet suffers a random delay in the delay process before it is finally served in the ideal service process. The length of time that a packet stays in the stochastic delay server is called *per-packet service time*.

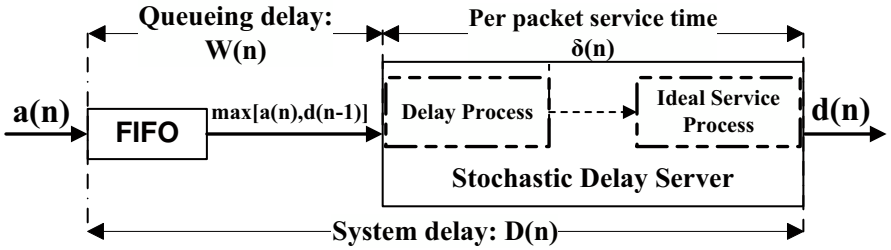

Figure 6.1: System Model

We decouple the delay process from the complete service process and study the delay process separately. The delay process represents the interval between the time when a packet reaches the HOL and the beginning of the successful transmission of the packet.

The homogeneous case implies that the transmission of any STA is interfered by other STAs with the same probability. The per-packet service time of all packets follows the same distribution due to the fixed packet size and the ideal channel assumptions. Thus, it suffices to analyze the per-packet service time for a concerned STA instead of the whole network.

6.3.1 Per-Packet Service Time Analysis

In the considered IEEE 802.11 network, the service time of any packet $P(n)$ is related to the number of collisions κ it experiences. Let C_n denote the sum of κ collisions. Then C_n equals $\kappa \cdot t_c$, where t_c is the duration of a collision. We denote the backoff interval at the kth backoff stage by b_k and the sum of backoff intervals by

$$B_n = \sum_{k=0}^{\kappa} b_k.$$

Let I_n represent the sum of inter-transmissions of non-concerned STAs. The service time of packet $P(n)$ can be expressed as below [92]:

$$\delta_n = C_n + B_n \cdot \sigma + I_n + t_s, \tag{6.2}$$

where σ denotes one slot time and t_s is the interval when the channel is occupied because of a successful transmission.

Since the service time of all packets have the same distribution, for ease of exposition, we remove the index n from the notations of (6.2) in the following analysis.

Let $H = L_{PHY} + L_{MAC}$ be the packet header and t_P be the packet transmission time. Then t_s is given below [13]:

$$\begin{cases} t_s^b = H + t_P + SIFS + ACK + DIFS \\ t_s^r = H + t_P + RTS + CTS + 3SIFS + ACK + DIFS, \end{cases}$$

where 'b' and 'r' are used to distinguish the basic access mechanism and the RTS/CTS mechanism, respectively. Note that the propagation delay is not considered in this paper.

6.3.2 Probability Distribution

Define P_k as the probability that a packet is successfully transmitted at the kth retransmission. Then we have

$$\begin{cases} P_k = (1-p_c)p_c^k, & k \in \{0,1,...,K-1\} \\ P_K = p_c^K. \end{cases}$$

Since a packet is either successfully transmitted within K retransmissions or dropped after K failed retransmissions, there holds

$$\sum_{k=0}^{K} P_k = 1.$$

The probability mass function (PMF) of the sum of collisions is given by

$$P\{C = \kappa \cdot t_c\} = P_\kappa, \qquad \kappa \in \{0,...,K\}, \tag{6.3}$$

from which, we compute the first and second moments of C

$$\begin{aligned} M_C^1 &= t_c \sum_{\kappa=1}^{K} p_c^\kappa, \\ M_C^2 &= (t_c)^2 \sum_{\kappa=1}^{K} (2\kappa - 1) p_c^\kappa. \end{aligned}$$

For the basic access mechanism, the collision occurs in data packet transmission. For the RTS/CTS mechanism, the collision occurs only in RTS packet transmission. The collision duration of the two mechanisms are given by [13]:

$$\begin{cases} t_c^b = H + t_P + DIFS \\ t_c^r = RTS + DIFS. \end{cases}$$

where, t_c^r is always shorter than t_c^b [13] according to the numerical values provided by the standard [1].

The distribution of the cumulative inter-transmission I relies on the cumulative backoff interval B since the inter-transmission can occur at any slot when the channel is in 'idle' state. When the concerned STA is in backoff state, its backoff counter decrements according to the perceived channel state. We adopt the assumption [92] that if the concerned STA detects the channel being occupied, the current slot is

considered to pass. Two mutually exclusive events can cause the channel being occupied: either a successful transmission or an RTS packet collision. Each slot will be sensed idle with the probability P_{idle} [92] if no non-concerned STAs transmit, where

$$P_{idle} = (1 - p_a)^{N-1}.$$

The channel occupied by a successful transmission implies that only one non-concerned STA transmits with the probability

$$P_{suc} = (N - 1)p_a(1 - p_a)^{N-2}.$$

The probability that the channel is occupied due to the collision involving only non-concerned STAs is

$$1 - P_{idle} - P_{suc}.$$

Let X_i denote the length of time that the backoff counter decrements by 1, where $i = 1, 2, ...$ The PMF of X_i is given by

$$f_X(x) = \begin{cases} P\{X_i = \sigma\} = (1 - p_a)^{N-1} = 1 - p_c \\ P\{X_i = t_s\} = P_{suc} \\ P\{X_i = t_c\} = p_c - P_{suc} \end{cases}$$

from which, the first and second moments of X_i are obtained

$$M_X^1 = (1 - p_c) \cdot \sigma + P_{suc} \cdot (t_s - t_c) + p_c \cdot t_c$$
$$M_X^2 = (1 - p_c) \cdot \sigma^2 + P_{suc} \cdot (t_s^2 - t_c^2) + p_c \cdot t_c^2.$$

Each slot is interrupted with the equal probability, thus

$$B \cdot \sigma + I = \sum_{i=1}^{B} X_i. \tag{6.4}$$

We denote the PMF of b_k by $f_k(\cdot)$ with mean μ_k and variance σ_k^2. The PMF of B can be expressed by conventional convolution of $\kappa + 1$ functions [28]:

$$f_B(x) = \sum_{\kappa=0}^{K} (f_0 * \cdots * f_\kappa)(x) \cdot P_\kappa.$$

Obviously, the computation of $f_B(x)$ involves multiple convolutions which can make the computation of (6.4) so complicated that it is difficult to characterize the exact distribution of per-packet service time. In order to ease the expression to help our understanding, the next section makes use of stochastic network calculus, which enjoys the advantage of flexibly analytical solutions and has the essence of finding some bounds for the tail probability of the interested distribution.

6.4 Probabilistic Bounds

In this section, we analyze and derive bounds for the tail probability of the per-packet service time, the cumulative service time and the system delay. The key technique is time-domain stochastic network calculus.

6.4.1 Per-Packet Service Time Bound

The time-domain service curve characterizes the stochastic nature of the cumulative per-packet service time. Section 6.3.2 reveals the difficulty of finding the exact probability distribution of the per-packet service time. To solve this difficulty, this subsection presents two approaches for finding the bounds on the tail probability of the per-packet service time.

We first adopt the moment bound [88]. Let M_Z^Q denote the Qth moment of Z. Then, the moment bound tells that the tail probability of Z is bounded by

$$P\{Z > z\} \leq \inf_{Q \geq 0} \frac{M_Z^Q}{z^Q}. \tag{6.5}$$

Recall that b_k and B denote the backoff interval at the kth backoff stage and the sum of backoff intervals, respectively. Given the mean and variance of b_k, the first and second moments of B are obtained [28]:

$$\begin{aligned} M_B^1 &= \sum_{k=0}^{K} \mu_k p_c^k, \\ M_B^2 &= \sum_{k=0}^{K} (\mu_k^2 + \sigma_k^2) p_c^k + 2\Big[\sum_{k=1}^{K} \mu_k p_c^k \cdot \sum_{j=0}^{k-1} \mu_j\Big]. \end{aligned}$$

The right-hand side of Eq.(6.4) is a compound random variable of which the first and second moments are:

$$\begin{aligned} M_{B+I}^1 &= M_B^1 M_X^1, \\ M_{B+I}^2 &= M_B^1 M_X^2 + (M_B^2 - M_B^1)(M_X^1)^2. \end{aligned}$$

Combining the above results together with the per-packet service time analysis in the previous section, the first moment (mean) of the service time δ is calculated by:

$$M_\delta^1 = M_C^1 + M_{B+I}^1 + t_s. \tag{6.6}$$

Let $\Omega = C + B \cdot \sigma + I$, and correspondingly

$$\begin{aligned} M_\Omega^1 &= M_C^1 + M_{B+I}^1, \\ M_\Omega^2 &= M_C^2 + M_{B+I}^2 + 2M_C^1 M_{B+I}^1. \end{aligned}$$

By applying Inequality (6.5), we get a bound on the per-packet service time:

$$P\{\delta > x\} = P\{\Omega > x - t_s\} \leq \inf\Big[\frac{M_\Omega^1}{x - t_s}, \frac{M_\Omega^2}{(x - t_s)^2}\Big]. \tag{6.7}$$

The moment bound (6.7) is simple while not a closed-form expression. In order to derive the delay bound, the bounding function associated with the service curve should be closed-form (see Theorem 14). The desired form should be integrable if the arrival process and the service process are independent of each other (see Lemma 8).

To solve the above difficulty, the other bound is obtained as follows. Note that for any $\theta > 0$ and $0 \leq Y \leq 1$, there always holds (e.g. see Lemma 2.2 in [85]):

$$e^{\theta Y} \leq 1 - Y + Ye^{\theta}.$$

If Y is a bounded random variable between 0 and 1, letting $q \equiv \mathbf{E}(Y)$, we have

$$\mathbf{E}[e^{\theta Y}] \leq 1 - q + qe^{\theta}. \tag{6.8}$$

In order to apply Inequality (6.8) for bounding δ, we need to normalize δ. Two extreme events determine the range of δ.

- No collision occurs and the backoff interval equals 0. Hence δ takes the minimum value t_s.

- δ reaches the maximum value $K \cdot t_c + B_{max} \cdot t_s + t_s$ because the number of collisions reaches K, the sum of backoff intervals attains the maximum value B_{max}, and every slot is interrupted by a non-concerned STA's successful transmission, where

$$B_{max} = \sum_{k=0}^{K} (CW_k - 1).$$

Let us now define the following notations:

$$\begin{aligned} Y &= \frac{\delta - t_s}{K \cdot t_c + B_{max} \cdot t_s}, \\ q &= \frac{M_\delta^1 - t_s}{K \cdot t_c + B_{max} \cdot t_s}, \\ y &= \frac{x - t_s}{K \cdot t_c + B_{max} \cdot t_s}. \end{aligned}$$

Applying first Chernoff bound to the per-packet service time and then following Inequality (6.8) yield, for any $\theta > 0$,

$$\begin{aligned} P\{\delta > x\} &= P\{Y > y\} \\ &\leq e^{-\theta y}\mathbf{E}[e^{\theta Y}] \\ &\leq e^{-\theta y}(1 - q + qe^{\theta}). \end{aligned}$$

By setting $e^{\theta} = \frac{y(1-q)}{q(1-y)}$, we have the following per-packet service time bound:

$$P\{\delta > x\} \leq \left(\frac{q}{y}\right)^y \left(\frac{1-q}{1-y}\right)^{1-y}.$$

The following lemma summarizes the above two bounds on the per-packet service time.

Lemma 21. *For a homogeneous single cell IEEE 802.11 network where all contending STAs employ the DCF scheme*[3]*, the per-packet service time can be bounded by the following bounds.*

Bound 1*. For $x > t_s$, there holds*[4]

$$P\{\delta > x\} \leq \inf\left[\frac{M_\Omega^1}{x - t_s}, \frac{M_\Omega^2}{(x - t_s)^2}\right]_1. \tag{6.9}$$

where $[z]_1 \equiv \min\{z, 1\}$.

Bound 2*. For $0 < x - t_s < K \cdot t_c + B_{max} \cdot t_s$, there holds*

$$P\{\delta > x\} \leq \left(\frac{q}{y}\right)^y \left(\frac{1-q}{1-y}\right)^{1-y}. \tag{6.10}$$

[3]Assume that all contending nodes employ the same access mechanism.

[4]The left-hand side of Inequality (6.9) has an upper bounded $\inf\left[\frac{M_\Omega^1}{x-t_s}, \frac{M_\Omega^2}{(x-t_s)^2}\right]$. It is a probability and should not be greater than 1. Thus, we take the minimum between 1 and the upper bound.

Remark. For Bound 1, Inequality (6.9) still holds mathematically even under $x - t_s > K \cdot t_c + B_{max} \cdot t_s$. However, for Bound 2, if $x - t_s \geq K \cdot t_c + B_{max} \cdot t_s$ which causes $1 - y \leq 0$, then Inequality (6.10) loses the mathematical meaning.

6.4.2 Network Calculus Approach

The (weak) law of large numbers states that the sample average converges in probability towards the expected value (first moment), i.e., for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\Big\{\Big|\frac{\sum_{k=0}^{n}\delta_k}{n+1} - M_\delta^1\Big| < \epsilon\Big\} = 1,$$

which implies the average of the cumulative service time will approach the first moment of the per-packet service time as 'n' becomes sufficiently large. Then we get a stochastic service curve for the concerned STA.

Lemma 22. *In a homogeneous single cell IEEE 802.11 network, for a concerned STA, the DCF access scheme provides a (time-domain) stochastic service curve* $\gamma_\eta(n) = M_\delta^1 \cdot n + \eta \cdot n$ *with bounding function* $j(x)$ *for* $\eta \geq 0$*, where for any* $x \geq 0$*,*

$$j(x) = \Big(\frac{q}{q+y+\bar{\eta}}\Big)^{q+y+\bar{\eta}}\Big(\frac{1-q}{1-q-y-\bar{\eta}}\Big)^{1-q-y-\bar{\eta}} \tag{6.11}$$

where

$$\begin{aligned}\bar{\eta} &= \frac{\eta}{K \cdot t_c + B_{max} \cdot t_s},\\ y &= \frac{x}{K \cdot t_c + B_{max} \cdot t_s + t_s}.\end{aligned}$$

Proof. We shall prove that $P\{d(n) - a\bar{\otimes}\gamma(n) > x\} \leq j(x)$.

Expanding $d(n)$ and $a\bar{\otimes}\gamma(n)$ yield

$$\begin{aligned} & d(n) - a\bar{\otimes}\gamma(n) \\ = & \sup_{0\le m\le n}\big\{a(m) + \sum_{k=m}^{n}\delta_k\big\} - \sup_{0\le m\le n}\big\{a(m) + \gamma_\eta(n-m+1)\big\} \\ \le & \sup_{0\le m\le n}\big\{\sum_{k=m}^{n}\delta_k - \gamma_\eta(n-m+1)\big\} \\ = & \sup_{0\le m\le n}\big\{\sum_{k=m}^{n}(\delta_k - M_\delta^1 - \eta)\big\}. \end{aligned}$$

Let $U_l \equiv \sum_{k=n-l}^{n}(\delta_k - M_\delta^1 - \eta)$ and $V_l \equiv e^{\theta U_l}$ with $\theta > 0$ and $0 \le l < n$. Note that δ_k, $k = 1, \ldots, n$, are independent of each other and follow the same distribution.

Let $\mathcal{M}_l$ denote the σ-algebra generated from the process $\mathbf{V} = \{V_l : 0 \le l < n\}$. If $\mathbf{E}[e^{\theta(\delta_k - M_\delta^1 - \eta)}] \le 1$, we then have:

$$\begin{aligned} \mathbf{E}\big[V_{l+1}|\mathcal{M}_l\big] &= V_l\mathbf{E}\big[e^{\theta[\delta_{n-1-(l+1)} - M_\delta^1 - \eta]}\big] \\ &= V_l\mathbf{E}\big[e^{\theta(\delta_0 - M_\delta^1 - \eta)}\big] \\ &\le V_l, \end{aligned}$$

from which, we know that $V_0, \ldots, V_{n-1}$ form a supermartingale.

Similarly, if $\mathbf{E}[e^{\theta(\delta_k - M_\delta^1 - \eta)}] \ge 1$, $V_0, \ldots, V_{n-1}$ form a submartingale. Combining both cases, $\mathbf{V}$ is a martingale, i.e., $\mathbf{E}[V_{l+1}|\mathcal{M}_l] = V_l$. With Doob's martingale inequality (see Lemma 3), we have

$$\begin{aligned} & P\{d(n) - a\bar{\otimes}\gamma(n) > x\} \\ \le & P\{\sup_{0\le l<n} U_l > x\} \\ = & P\{\sup_{0\le l<n} V_l > e^{\theta x}\} \\ \le & e^{-\theta x}E[V_0] \\ = & e^{-\theta x}E\big[e^{\theta(\delta_n - M_\delta^1 - \eta)}\big]. \end{aligned} \tag{6.12}$$

If the distribution of δ_n is known, Inequality (6.12) usually gives a good bound. However, as discussed earlier, for the studied network scenario, an accurate estimation on the distribution is difficult. In the following, we shall make use of the fact that in the considered network, δ_n is formed by the sum of bounded random variables, as implied by Eq.(6.2).

Let

$$\begin{aligned} Y_k &= \frac{\delta_k}{K \cdot t_c + B_{max} \cdot t_s + t_s}, \\ \bar{\eta} &= \frac{\eta}{K \cdot t_c + B_{max} \cdot t_s}, \\ y &= \frac{x}{K \cdot t_c + B_{max} \cdot t_s + t_s}. \end{aligned}$$

Applying Y_k to Inequality (6.12) and following the same principle gives:

$$\begin{aligned} & P\{d(n) - a\bar{\otimes}\gamma(n) > x\} \\ \leq\ & P\Big\{ \sup_{0\leq m\leq n} \sum_{k=m}^{n} (Y_k - q - \bar{\eta}) > y\Big\} \\ \leq\ & e^{-\theta y} E[e^{\theta(Y_n - q - \bar{\eta})}] \\ \leq\ & e^{-\theta y} e^{-\theta(q+\bar{\eta})}(1 - q + qe^{\theta}), \end{aligned} \tag{6.13}$$

which holds for any $\theta > 0$. Setting $e^{\theta} = \frac{(q+\bar{\eta})(1-q)}{q(1-q-\bar{\eta})}$ results in

$$e^{-\theta y} e^{-\theta(q+\bar{\eta})}(1 - q + qe^{\theta}) = \Big(\frac{q}{q+\bar{\eta}}\Big)^{y+q+\bar{\eta}} \Big(\frac{1-q}{1-q-\bar{\eta}}\Big)^{1-y-q-\bar{\eta}}$$

from which, (6.11) is proved. □

Remark. If a packet has not yet been transmitted successfully when the retransmission limitation has reached, the packet is discarded. However, the discarded packets have consumed the service provided by the server. Thus, the service curve represents a stochastically guaranteed service time for all arrival packets regardless of whether the

packets are successfully transmitted or discarded. In addition, Lemma 22 also holds for other non-concerned STAs in the homogeneous IEEE 802.11 network.

Lemma 22 presents a stochastic service curve for the concerned node. If we know the arrival curve of the arrival process, the system delay that a packet experiences in the concerned STA has an upper bound by applying Theorem 14.

In the following, we discuss two types of arrival, denoted by F_1 and F_2 respectively. The packets of F_1 arrive at constant intervals T. This arrival process has a (deterministic) arrival curve $\lambda^{cnt}(n) = T \cdot n$ with bounding function $h^{cnt}(x) = 0$. For F_2, the packet inter-arrival times are exponentially distributed with mean $\frac{1}{\Lambda}$. This arrival process has a $v.w.d$ stochastic arrival curve $\lambda^{exp}(n) = \hbar \cdot n$ ($\hbar < \frac{1}{\Lambda}$) with bounding function $h^{exp}(x)$ (see Example 4):

$$h^{exp}(x) = 1 - (1-\rho) \sum_{i=0}^{\lfloor \frac{x}{\hbar} \rfloor} e^{-\Lambda(i\hbar - x)} \frac{[\Lambda(i\hbar - x)]^i}{i!} \tag{6.14}$$

where $\rho = \Lambda \cdot \hbar$. To ensure system stability, we require $\lim_{n\to\infty} \frac{1}{n}[\gamma(n) - \lambda(n)] \leq 0$ for both F_1 and F_2.

With Theorem 14, we readily obtain the system delay bounds with two arrival processes, accordingly.

- The system delay of any packet belonging to flow F_1 is bounded by
$$P\{D^{cnt}(n) > x\} \leq j(x - M_\delta^1 - \eta). \tag{6.15}$$
- The system delay of any packet belonging to flow F_2 is bounded by
$$P\{D^{exp}(n) > x\} \leq j \otimes h^{exp}(x - M_\delta^1 - \eta). \tag{6.16}$$

Remark. Recall that the system delay can be expressed as

$$D(n) = \Big[a(m_0) + \sum_{k=m_0}^{n} \delta_k\Big] - a(n),$$

where $a(m_0)$ is the beginning of the latest backlogged period. The saturated condition implies $a(m_0) = a(0)$ for all packets. The stochastic service curve given by Lemma 22 implies a probabilistically guaranteed service time under the maximal load that can be carried by the

network. Thus, the system delay derived from this stochastic service curve is essentially the probabilistic upper bound of the system delay which an arrival process actually experiences in the network.

6.5 System Delay Bounds under Finite Buffer

The system delay bounds derived from Theorem 14 rely on an important assumption that the buffer space is sufficient to store all incoming packets, i.e., a lossless system. This assumption has been commonly used in the available literature on network calculus. However, the realistic situation is not like this since the physical buffer must be finite. How does the finite buffer space impact the system delay bound is thus investigated in this section.

The saturation condition implies that every packet $P(n)$ arrives to the system before the previous packet $P(n-1)$ leaves the system, i.e., $a(n) \leq d(n-1)$. We discuss '$<$' and '$=$' separately. Recall the system delay

$$D(n) = d(n) - a(n). \tag{6.17}$$

Replacing $d(n)$ with Eq.(3.13) yields

$$D(n) = \sup_{0 \leq m \leq n} [a(m) + \sum_{k=m}^{n} \delta_k] - a(n). \tag{6.18}$$

Scenario I.

The first scenario is that $a(n) = d(n-1)$ for all $n \geq 1$, i.e., packet $P(n)$ arrives to the system as packet $P(n-1)$ departs from the system. As shown in Figure 6.2, there is no queueing delay for any packet and thus Eq.(6.17) returns $D(n) = \delta_n$. Then the per-packet service time bound given by Lemma 21 is also the system delay bound. Lemma 21 is also applicable when $P(n)$ arrives to the system, $P(n-1)$ has departed from the system, i.e., $d(n-1) < a(n)$, where, there holds $a(n+1) - a(n) \geq \delta_n$[5].

[5]However, this does not satisfy the saturation condition.

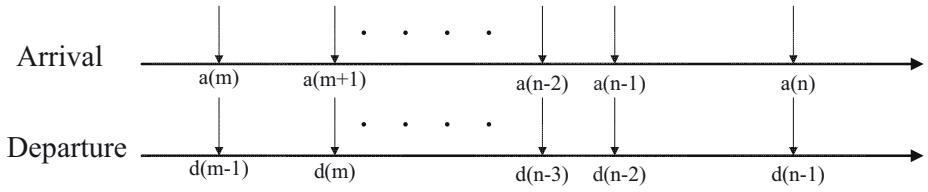

Figure 6.2: No queueing delay

Scenario II.

The second scenario considers the queueing delay. The saturation condition implies that all contending STAs are always backlogged. However, an infinite queue is not desired and realistic. For uplink transmission in IEEE 802.11 networks, the per-packet service time may possess long range dependence [92] which can cause excessive queueing delay. In the following, we assume a finite buffer of capacity Φ (number of packets[6]). For packets allowed to enter the buffer, their system delay should be upper bounded in terms of Φ.

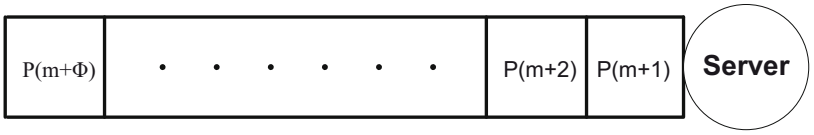

Figure 6.3: Finite buffer is full

We investigate the maximum system delay when the buffer is full. Without loss of generality, suppose packet $P(m+1)$ is the first packet being placed into the buffer at some instant when the buffer is empty while there is one packet being transmitted. More packets arrive to the system until the buffer is full. When packet $P(m+\Phi)$ enters the buffer, the first packet $P(m+1)$ is still in the system. As shown in Figure 6.3, such scenario implies $a(m+\Phi) \leq d(m+1)$. Then Eq. (6.18) returns

$$D(m+\Phi) \leq \sum_{k=1}^{\Phi} \delta_{m+k}. \tag{6.19}$$

[6]Assume that packets have the same length. Otherwise, the buffer capacity should be measured using the number of bits.

Lemma 21 has given two bounds on the service time of a single packet. In order to bound the cumulative service time presented in Inequality (6.19), we generalize Lemma 21 into Lemma 23.

Lemma 23. *Consider a concerned STA with a finite buffer size* Φ *in a homogeneous single cell IEEE 802.11 network. The DCF guarantees a system delay bound for packets transmitted from the concerned STA as follows.*

Bound 1. *Recall* $\Omega = C + B \cdot \sigma + I$*. According to Inequality (6.5), we have for any* $n \geq 0$ *and* $x - t_s > 0$

$$P\{D(n) > x\} \leq \inf \Big[\frac{\Phi M_\Omega^1}{x - \Phi t_s}, \frac{\Phi M_\Omega^2 + (\Phi^2 - \Phi)(M_\Omega^1)^2}{(x - \Phi t_s)^2} \Big]_1. \qquad (6.20)$$

Bound 2. *Based on Inequality (6.10), we have the following bound on the cumulative service time:*

$$P\{D(n) > x\} \leq \Big\{ \Big(\frac{q}{y}\Big)^y \Big(\frac{1-q}{1-y}\Big)^{1-y} \Big\}^\Phi \qquad (6.21)$$

where

$$\begin{aligned} y &= \frac{x}{\Phi \cdot (K \cdot t_c + B_{max} \cdot t_s + t_s)}, \\ q &= \frac{M_\delta^1}{K \cdot t_c + B_{max} \cdot t_s + t_s}. \end{aligned}$$

Proof. Inequality (6.20) is obtained by applying the moment bound to $\Phi \cdot \Omega$. Having known M_Ω^1 and M_Ω^2, then the first and second moments of $\Phi \cdot \Omega$ are given below respectively:

$$\begin{aligned} M_{\Phi \cdot \Omega}^1 &= \Phi M_\Omega^1, \\ M_{\Phi \cdot \Omega}^2 &= \Phi M_\Omega^2 + (\Phi^2 - \Phi)(M_\Omega^1)^2, \end{aligned}$$

with which, by applying the moment bound (6.5) to $\Phi \cdot \Omega$, we get the moment bound (6.20) on the cumulative service time of Φ packets.

Inequality (6.21) is the extension of Inequality (6.10). Here, q and Y_k have the same definition as those in Inequality (6.10). Let

$$y = \frac{x}{\Phi(K \cdot t_c + B_{max} \cdot t_s + t_s)}.$$

Then we have

$$\begin{aligned} P\{\sum_{k=1}^{\Phi} \delta_k > x\} &= P\{\sum_{i=1}^{\Phi} Y_k > \Phi \cdot y\} \\ &\leq e^{-\theta(\Phi \cdot y)} \mathbf{E}[e^{\theta \sum_{i=1}^{\Phi} Y_i}] \\ &\leq \left(e^{-\theta y}(1 - q + qe^{\theta})\right)^{\Phi}, \end{aligned}$$

from which we get Inequality (6.21) by letting $e^{\theta} = \frac{y(1-q)}{q(1-q-y)}$. □

Remark. When $\Phi = 1$, Inequality (6.21) becomes Inequality (6.10). In addition, Φ also implies the burst tolerance of the system.

6.6 Numerical Evaluation and Discussion

To better understand the analytical bounds derived in Section 6.4, we use numerical analysis to extensively examine the relevant parameters. We adopt the parameter setting of the IEEE 802.11b as listed in Table 6.1. Two data packet sizes $L_1 = 1500$ *bytes* and $L_2 = 150$ *bytes* are considered here.

We first examine the average per-packet service time calculated using Eq.(6.6) and shown in Figure 6.4. For the short packet size $L_2 = 150$ bytes, the basic access always outperforms the RTS/CTS access against the varying number of STAS. This result is consistent with the clarification of the RTS/CTS [1] that the RTS/CTS is not suitable for short data packet transmission. For the long packet size $L_1 = 1500$ bytes, we notice that although the RTS/CTS access performs better when the number of STAs exceeds 25, the advantage is not apparent. The reason is that the physical header and all control packets are always transmitted at the rate of $2Mbps$ while the data packets are

Table 6.1: IEEE 802.11b Parameters (for DSSS system)

Parameter	Value
Control bit rate	2 Mbps
Data bit rate	11 Mbps
PHY header	192 bits
MAC header	224 bits
ACK packet	112 bits + PHY header
RTS packet	160 bits + PHY header
CTS packet	112 bits + PHY header
SlotTime (σ)	20 μs
SIFS	10 μs
DIFS	50 μs
Min CW (CW_{min})	32
Max retransmission (K)	6
Doubling limit (M)	5

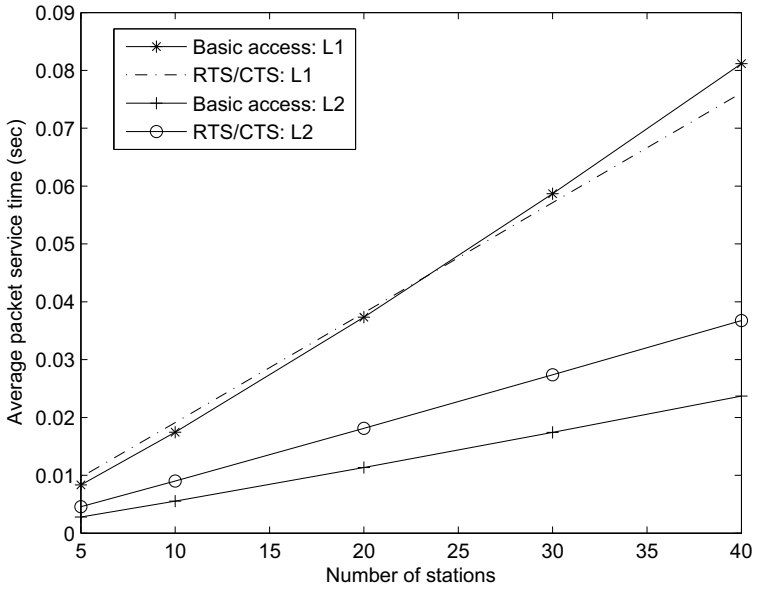

Figure 6.4: Average Packet Service Time

transmitted at the rate of $11Mbps$. The influence of data packet size on performance is beyond scope of this chapter. More discussion about the influence of data packet size can be found [24].

To ease presentation, in Figure 6.5 - Figure 6.8, the x-axis represents the **normalized** time.

Lemma 21 provides two bounds on the per-packet service time. In Figure 6.5, we compare these two bounds by configuring the scenario that all contending STAs (10 and 20 respectively) employ the basic access mechanism and the packet size is L_1. For both 10-STA and 20-STA cases, the bound given by Inequality (6.9) (Bound 1 in Figure 6.5) are tighter than the bound given by Inequality (6.10) (Bound 2 in Figure 6.5). An implicit reason is that while both the first and second moments are used to find Inequality (6.9), the derivation of Inequality (6.10) only involves the first moment.

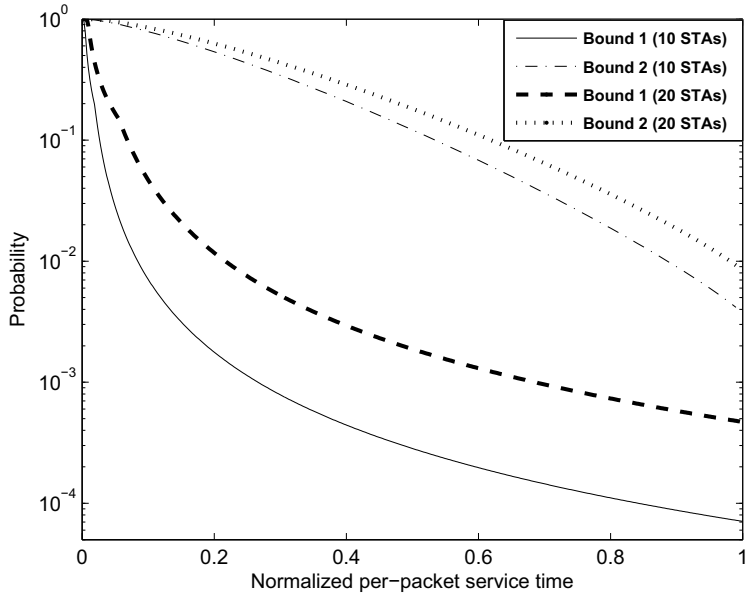

Figure 6.5: Two bounds given in Lemma 21 (basic access and packet size $L1$)

The stochastic service curve $\gamma_\eta(n)$ defined in Lemma 22 presents an upper bound on the exceedance probability that the cumulative actual service time exceeds the cumulative guaranteed service time. As shown in Inequality (6.11), η is an adjustable parameter. According

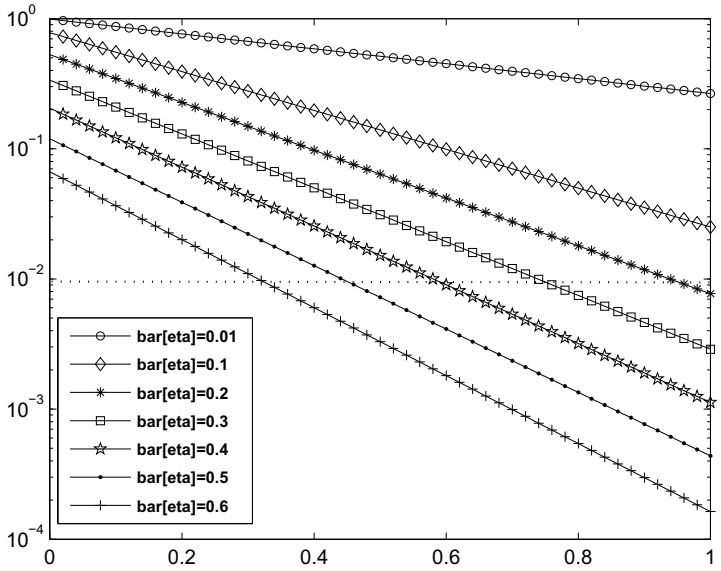

Figure 6.6: Stochastic service curve for basic access (packet size L_1 and 10 STAs)

to the analytical formula (6.11), Figure 6.6 shows the service curve as the normalized η, denoted by **bar[eta]** in Figure 6.6, varies from 0.01 to 0.6. The considered scenario is that the network consists of 10 contending STAs and all STAs employ the basic access mechanism. All packets have the same size L_1. As $\bar{\eta}$ increases, the guaranteed per-packet service time becomes larger and accordingly the exceedance probability decreases. If $\bar{\eta}$ becomes so large that the normalized $M_\delta^1+\eta$ approaches '1', the guaranteed per-packet service time approaches the maximal service time and then the service curve is not much meaningful. We are interested in that for a specific criteria, how to choose the service curve. For example, the dotted line in Figure 6.6 represents the probability of 1%. The intersection of this dotted line and a service curves represents the service time than which less than[7] 1% of packets will receive a larger service time.

We now discuss the system delay bounds derived using the network calculus approach and the finite buffer size. The investigated scenario is that 10 contending STAs employ the RTS/CTS access mechanism

[7]Note that the service curve is an upper bound on the violation probability.

and the packet size is L_1.

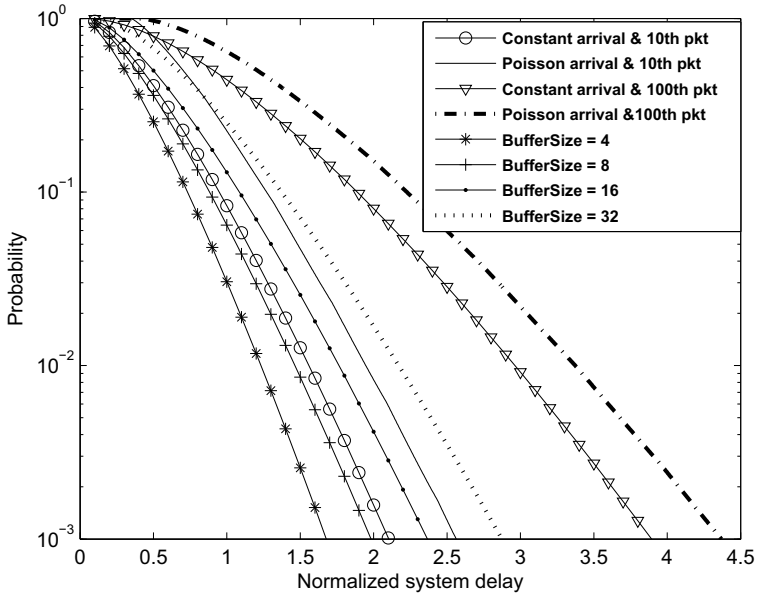

Figure 6.7: System delay bound for RTS/CTS (packet size L_1 and 10 STAs)

For the network calculus approach, we evaluate the analytical bounds for the Poisson and the constant arrival processes. The parameters of Poisson arrival used in Inequality (6.14) are set to $\hbar = 0.04sec$ and $\Lambda = 12.5$. The parameters of the constant arrival do not impact the system delay bound. The system delay bound of two arrival flows are computed using Inequality (6.15) and Inequality (6.16) and shown in Figure 6.7. The system delay bound of the constant arrival flow is tighter than that of the Poisson arrival flow. This result is consistent with the analytical bounds.

For the same flow, the $100th$ packet may suffer longer system delay than the $10th$ packet because the service curve is defined for the situation that all contending STAs are always backlogged. Both the cumulative waiting delay and cumulative service time of previously transmitted packets contribute to the system delay of the current packet. The packet arriving later will hence wait longer before it can be served.

Note that Inequality (4.1) (Theorem 14) gives the system delay bound under the assumption of an infinite buffer size, which implies

that all packets in the buffer will receive service regardless of their waiting delay. However, the buffer size is limited in real networks. Thus, bound 2 (see Inequality (6.21)) in Lemma 23 gives a maximal system delay bound for a finite buffer. Figure 6.7 also shows such system delay bounds against various buffer sizes, Φ, which can be understood as either the physical buffer capacity or a threshold of buffer occupancy. Φ can be dynamically set for differentiating the QoS of different applications. For example, for the 10th packet of the constant arrival process, when $\Phi = 4$ and 8, Inequality (6.21) provides a tighter bound than Inequality (6.15); if $\Phi > 10$ such as 16 and 32, Inequality (6.21) becomes loose compared with Inequality (6.15). Such trend is understandable since Inequality (6.15) and Inequality (6.21) are close when the parameter n of Inequality (6.15) equals the parameter Φ of Inequality (6.21). However, for the Poisson arrival process, the impact of the arrival curve causes that such relationship between n and Φ does not hold in general. For instance, for the 10th packet of the Poisson arrival, Inequality (6.21) is tighter than Inequality (6.16) even when $\Phi = 16$.

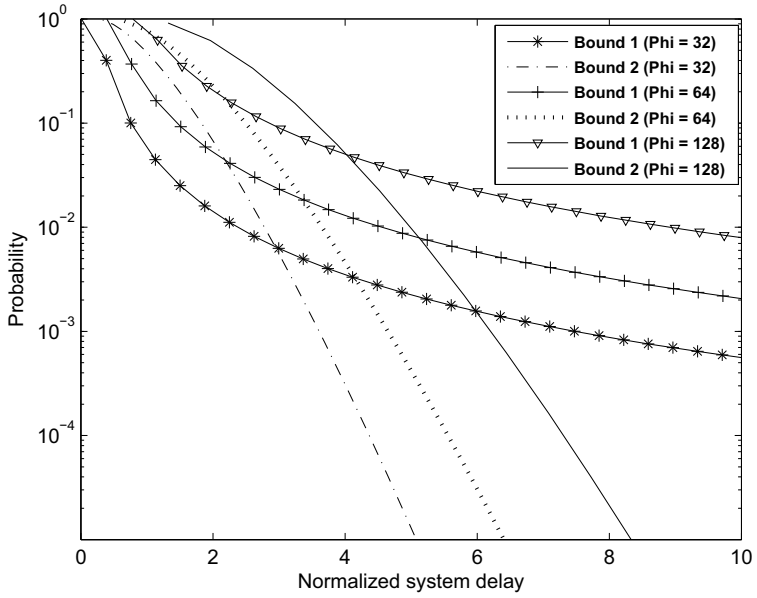

Figure 6.8: Two bounds given in Lemma 23 for RTC/CTS access (packet size L_1 and 10 STAs)

For a finite buffer, Lemma 23 also gives another bound (see Inequality (6.20)) obtained using the moment bound. Figure 6.8 compares two bounds given in Lemma 23. Interestingly, neither Bound 1 nor Bound 2 shows absolutely advantage under various buffer sizes. If the system delay is lower than some threshold, such as 3 when $\Phi = 32$, Bound 1 outperforms Bound 2. Once the system delay exceeds this threshold, Bound 2 decays much faster than Bound 1 and thus provides a tighter bound. As the buffer size Φ increases, such threshold occurs earlier compared with the maximal system delay. Comparing to Bound 2, Bound 1 shows a relatively slow decay. To obtain an optimal bound, we may take the minimal one between these two bounds.

6.7 Conclusion

This chapter demonstrates an application of stochastic network calculus to delay evaluation of IEEE 802.11 DCF. The DCF behavior is characterized using the $i.d$ stochastic service curve model. The actual per-packet service time is described by comparing with the guaranteed service time. We present two approaches for bounding the tail probability of the per-packet service time. Comparing these two bounds, the moment bound (the first approach) shows an apparent advantage but does not provide a closed-form expression. Particularly, for defining the stochastic service curve, we expect an integrable bounding function which is applicable for getting further results, such as delay bound. According to the property of martingale, we obtain a stochastic service curve associated with an integrable bounding function.

Based on the stochastic service curve, system delay bounds for the constant and Poisson arrival processes are derived. Moreover, system delay bound under finite buffer is investigated through extending the per-packet service time bounds to bounds on the cumulative service time. For this case, the moment bound outperforms the other bound only when the cumulative service time is lower than some threshold, which becomes relatively low as the number of served packets increases.

It should be stressed that stochastic network calculus focuses on characterizing the tail probability of the concerned performance metrics. For many cases, we can only find a bound on the tail probability instead of the exact probability distribution. Particularly, we show

that different mathematical approaches should be adopted appropriately for bounding various performance metrics. To obtain a tighter bound, a suitable approach should be applied and the stochastic nature of the studied performance metrics should be further considered.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Many computer networks such as wireless and multi-access networks are stochastic in nature. In order to conduct performance guarantee analysis of such networks, appropriate analytical tools and/or models are required. Stochastic network calculus is a relatively new theory evolved to deal with performance analysis issues over several years. This thesis focuses on extending the application of stochastic network calculus to analytical time-domain models and temporal behavior analysis of networks.

This thesis consists of two main parts: the first part is developing a generic time-domain framework for modeling network behavior and analyzing network performance; the second part mainly presents two applications to concretize the generic models and exemplify the whole procedure of applying the developed framework to network performance analysis.

Central to this thesis is the model definition which is the focus of Chapter 3 and exploration of fundamental properties presented in Chapter 4. The time-domain traffic models and service models may be considered as a generalization of the models in the classical queueing theory. More specifically, the traffic models in queueing theory mainly focus on describing the inter-arrival distribution between two consecutive customers while the time-domain traffic models defined in this thesis describe the cumulative inter-arrival time between two arbitrary customers (packets in this thesis). Similarly, the service models in queueing theory mainly focus on characterizing the per customer service time while the time-domain service models characterizes the cumulative service time of multiple packets.

Interestingly, the underlying connections between the time-domain models and the queueing models help to both intuitively understand the meaning of the time-domain models and explore the fundamental properties. For instance, the virtual Single Server Queue is introduced to explore the waiting delay in a virtual system which has been often used in this thesis. Both the time-domain traffic models and service models can be illustrated as a virtual system. Accordingly, some available queueing theory results about waiting delay distribution can be used to derive the bounding function for some specific examples.

Defining models has to compromise when simple models that may not be applicable for exploring the fundamental properties and the more constrained models that may be difficult to build. One way to

solve this difficulty is constructing transformation between the simple model and the constrained model. However, such transformation may sacrifice some of the precision in obtaining the bounding function, i.e., the bound may become loose.

The Chernoff bound is often applied to the arrival process characterization or the service process characterization. To develop the Chernoff bound, we need to compute the Moment Generating Function. However, if the packet inter-arrival time or service time does not follow the well-known distributions of which the MGF are known, it may be hard to find the MGF of such general distributions. Moreover, the worse case is that not every random variable has a moment generating function.

The $GI/GI/1$ queueing system can be easily represented using the time-domain models. For this special class of queueing systems, the available results in martingale play an important role in concretizing the generic time-domain models. Particularly, if a stochastic process is proven to be a martingale (or supermartingale/submartingale), the supremum of this stochastic process is stochastically bounded from above by a specific random variable in this stochastic process. This technique nicely overcomes the difficulty in analyzing the supremum of a stochastic process.

Some simple examples are given to help illustrate the newly defined models. Although these examples look intuitively simple, they still convey the potential application fields. For example, the constant rate server discussed in Section 5.3.1 is readily applicable for modeling the wired networks.

The ON-OFF model is revisited in Chapter 5. This model has been extensively used to characterize both the arrival process and the service process from the spatial perspective [21] [46] [65] [72]. It can also be applied to modeling the service process of error-prone wireless channels in the time-domain. Particularly, the channel 'good' state corresponds to the 'ON' state of the model, and the channel 'bad' state corresponds to the 'OFF' state of the model. Moreover, the impairment process [61] is readily applicable for modeling the error-prone wireless channel. Specifically, the impairment process of the error-prone wireless channel can be represented using an ON-OFF model as well. However, the channel 'good' state and 'bad' state are mapped into the 'OFF' state and 'ON' state of the impairment process, respectively. Interestingly, the two ON-OFF models above yield the same analytical performance bounds. However, this conclusion may

not hold in general.

Another insight gained from investigating the ON-OFF model is that the time-domain analysis and the space-domain analysis may yield close results under appropriately mapping the model arguments.

The performance analysis of IEEE 802.11 is an important research issue related to investigating the potential whether WLANs could support the emerging real-time multimedia applications. The Distributed Coordination Function is the basic medium access mechanism employed in WLANs while essentially difficult to analyze theoretically. In Chapter 6, how to formulate the DCF mechanism from the temporal perspective and then apply the time-domain stochastic network calculus model to the formulated system is presented in detail. One crucial strength of stochastic network calculus is to find the bounds on the probability distribution instead of computing the exact probability distribution. Such strength significantly alleviates the difficulty of the DCF performance analysis. Moreover, selecting the appropriate mathematical tools is part of exemplifying the temporal network calculus approach as well.

The preliminary work presented in Annex A mainly focused on studying the impaired service caused by the bit-level transmission errors. The corresponding stochastic process is defined as an error process which is a concretization of the impairment process. The concatenation property is particularly investigated to reveal how does the order of placing the ideal service process and the error process impact the analysis. Moreover, the various error handling schemes influence the network performance differently, for example, prolonging the delay or degrading the throughput.

7.2 Open Research Issues

This thesis aims to develop a general framework for stochastic network calculus to formulate queueing systems in the complicated computer networks, derive performance bounds and provide some applications to demonstrate how to apply the developed framework to the real network analysis.

However, there are several issues which have been considered or attempted to tackle but still open. In the following sections, we discuss them respectively.

7.2.1 A More Realistic IEEE 802.11 DCF Scenario

In this thesis, the IEEE 802.11 DCF is analyzed under the homogeneous and saturated conditions. The saturated condition implies that the analysis is conducted based on the worst case network behavior. For these scenarios, the Poisson traffic is not really applicable because it cannot guarantee the saturated condition. Thus, a more realistic scenario should be considered. It is worthing highlighting that the network load will impact the collision probability and the attempt probability.

First, the network load is *finite*, i.e., non-saturated condition. However, in order to formulate the non-saturated IEEE 802.11 DCF network, two important questions have to be answered first.

- How to determine the probability that the buffer is empty/nonempty according to the network load?
- How to characterize the relation between the network load and the average attempt rate which decides the collision probability based on the Bianchi model [13]?

Some available work on modeling the non-saturated condition has attempted to solve the above questions [18] [82] [106]. Particularly, some assumptions or conditions used in [106] match the considerations of Chapter 6 well, including a small buffer model and an infinite buffer model, which are analyzed separately, and the Poisson arrival traffic.

7.2.2 Leftover Service Characterization

The leftover service analysis from the spatial perspective is intuitive and readily obtained. In the space-domain, the leftover service characterization is represented using the surplus of the aggregate flow's stochastic service curve minus the cross traffic's stochastic arrival curve. However, exploring this property from the temporal perspective is much difficult and is yet lacked.

Paper C attempted to explore this property under the combination that the arrival process has a deterministic SAC and the service process provides an $i.d$ SSC. However, the difficulty of decoupling the constituent flow's arrival process from the aggregate arrival process

is still not solved. Indeed, the difficulty of exploring the superposition property indirectly invokes the difficulty of studying the leftover service characterization, i.e., the superposition of multiple independent renewal processes is generally not a renewal process. Moreover, another difficulty is to find the connection between the stochastic arrival curve of the cross traffic and the service process provided to the aggregate arrival process.

As discussed in Section 4.4.4, the Poisson process is an exceptional case of renewal processes. It may be easier to prove the leftover service characterization property under the condition that all constituent arrival processes are Poisson process and independent of each other. For such condition, the decomposition of a Poisson process [8] may provide some useful results for studying the leftover service characterization.

If the constituent arrival processes follow some general distributions, one optional approach is to transform the time-domain models into the space-domain models, based on which, the leftover service characterization of the concerned flow can be obtained. Then, the space-domain stochastic service curve of the concerned flow needs to be transformed into the time-domain stochastic service curve. However, in order to apply this approach, the transformation between the time-domain stochastic service curve and the space-domain stochastic service curve has to be established. Such transformation is missed yet.

7.2.3 Finite-State Markov Channel Analysis

The two-state Markov wireless channel is analyzed as an example of concretizing the generic time-domain models in Chapter 5. However, the two-state Markov channel may not be applicable for modeling the channel characteristics which may vary dramatically [98] [105]. The finite-state Markov channel (FSMC) has been extensively applied for investigating wireless network performance, such as computing the channel capacity [51] [56] or deriving the expected performance metrics in steady states [79]. Moreover, a closed-form expression of the effective bandwidth is derived from the FSMC subject to packet loss and delay constraints [55]. The effective bandwidth can be readily mapped into the stochastic arrival curve model. Accordingly, in order to conduct the performance analysis, the stochastic service curve of the wireless channel which can be modeled using the FSMC is needed.

7.3 Future Work

This thesis develops a temporal analysis approach under the umbrella of stochastic network calculus. The future work may cover several aspects.

- Concretization of the time-domain stochastic arrival curve models

  Many well-known types of traffic can be represented using the space-domain stochastic arrival curve models. It is worth investigating whether these traffic types can also be represented using the time-domain stochastic arrival curve models. Moreover, there may exist many traffic types which are not readily characterized using the space-domain stochastic arrival curve while more suitable for being modeled as the time-domain stochastic arrival curve.

- Multi-hop wireless network analysis

  The concatenation property is explored to facilitate the end-to-end performance analysis. Since the single hop IEEE 802.11 network has been investigated in this thesis, it is natural to consider analyzing the performance of multi-hop 802.11 networks. Two challenges may be faced in analyzing multi-hop wireless LANs [54].

  1. Each node hears different events on the channel. There is no common view of the wireless channel.
  2. With a general channel model, the possible channel states in the multi-hop wireless LANs are more complicated than those in the single hop case.

  Moreover, for 802.11 networks which may have a diameter of about 2 or 3 hops, the intra-flow contention may severely degrade the network performance [101].

- Wireless scheduling discipline analysis

  In wireless networks, the available bandwidth depends on the channel state. Hence, the wireless scheduling with QoS guarantees is *channel state dependent*. The wireless scheduler has a very important characteristic which is to utilize asynchronous channel

variations or multi-user diversity [99]. Many wireless scheduling disciplines are derived from the well-known GR servers. For example, wireless fluid fair queueing (WFFQ) and idealized wireless fair queueing (IWFQ) [81] and channel-condition independent packet fair queue (CIF-Q) [43] are based on GPS and WFQ. The common characteristic among these wireless scheduling disciplines is to compare the received service of a flow with an ideal error-free service which is defined as the weighted fair queuing (WFQ).

Appendix A

Service Model with Impairment Process: A Concrete Example

The material in this part has been partially published as follows:

Jing Xie and Yuming Jiang. "An Analysis on Error Servers for Stochastic Network Calculus." In *Proceedings of the* 33^{rd} *IEEE Conference on Local Computer Networks (LCN)*, Montreal, Canada, October 2008.

Chapter A. Service Model with Impairment Process: A Concrete Example

The purpose of this chapter is to propose a service model by taking into account the impaired service which is caused by the bit-level transmission errors. Typically, the transmission error issue is not severe in wired networks whereas wireless networks often suffer higher transmission errors [10]. The performance of wireless network is impacted, accordingly. In order to investigate the performance of wireless networks, the appropriate service model which can explicitly characterize the random transmission error is needed.

We propose a service model under the umbrella of stochastic network calculus, which makes a step forward towards addressing the above issue. Central to this issue is errors in transmission. Particularly, in a network with error-prone links, errors are inherent in the random quality nature of these links. The network may react accordingly to the error information, such as re-send at the sender and/or drop at the receiver. However, in the current network calculus literature, errors are either not considered or are considered only implicitly. The amount of service corresponding to errors is simply treated as *impaired service* which is characterized using an *impairment process* and deduced in service guarantee analysis [61]. This simple way of treating errors in the analysis makes it difficult to apply existing network calculus results to investigate error-prone networks where some error handling methods are adopted to adapt service provision based on the error information.

The key idea is to introduce an error process in the service model. The error process essentially concretizes the impairment process. We use an *ideal service process* and an *error process* to model a server. The ideal service process characterizes the amount of service when there would be no transmission errors. The error process characterizes the transmission errors in the service. The idea of introducing an error process in the service model is intuitively simple. In addition, the proposed service model may look similar to the *space-domain stochastic strict server* model introduced in [61] where an *impairment process* is used together with the ideal service process to model the server. Essentially, the error process concretizes the impairment process in the context of error-prone wireless networks.

The contributions of this chapter are several-fold. First, instant error processes and cumulative error processes are introduced, based on which the proposed error server model is described. Second, the concatenation property (P3.) of the proposed service model is investigated. Third, a simple network is studied to demonstrate how to apply

the introduced concepts and to show the impact of error handling on system performance.

The next section defines the service model with the impairment process. Section A.2 investigates the concatenation property of the proposed service model. In Section A.3, the proposed model is generalized, the concept of stochastic error curve is introduced, and its concatenation property is presented. Section A.4 considers a simple network to demonstrate the use of the proposed service model. Performance bounds are derived and compared for two simplest error handling methods.

A.1 Concretization of Impairment Process: Error Process

This section defines the service model with the error process which concretizes the impairment process. To explain the idea, we assume bit-stream traffic in this section. The intuition of the service model is based on the fact that if a received bit is different from the corresponding bit that has been sent, a transmission error has happened.

A.1.1 Error Processes

To describe errors, two stochastic processes $\hat{E}(t)$ and $E(t)$ are considered. They are an *instant bit error process* and a *cumulative bit error process*.

Definition 16.

1. The instant bit error process $\hat{E}(t)$ *is a collection of random variables* $\{\epsilon(t), t = 1, 2, ...\}$ *with* $\epsilon(0) = 0$*, where* $\epsilon(t)$ *is a Bernoulli random variable:* $\epsilon(t) = 1$ *if the error happens, and* $\epsilon(t) = 0$ *if the error does not happen.*

Table A.1: XOR Truth Table

$\oplus$		u(t) 0	u(t) 1
$\hat{u}(t)$	0	0	1
	1	1	0

2. The cumulative bit error process $E(t)$ *is a collection of random variables* $\{e(t) = \sum_{s=0}^{t} \epsilon(s), t = 1, 2, ...\}$ *with* $e(0) = 0$*, where* $e(t)$ *represents the cumulative number of errors in interval* $(0, t]$.

Consider a system with instant bit error process $\hat{E}(t)$, which has input $\mathcal{A}(t)$ and output $\hat{\mathcal{A}}(t)$. Assume no delay in the system. If $u(t)$ is the input bit at time t, then the corresponding output bit $\hat{u}(t)$ is:

$$\hat{u}(t) = u(t) \oplus \epsilon(t) \tag{A.1}$$

where, "$\oplus$" denotes the bitwise XOR operation.

According to the truth table of XOR operation shown in Table A.1, the equation below holds.

$$\epsilon(t) = u(t) \oplus \hat{u}(t). \tag{A.2}$$

Then, the cumulative error by time t equals

$$e(t) = \sum_{s=0}^{t} \epsilon(s) = \sum_{s=0}^{t} u(s) \oplus \hat{u}(s). \tag{A.3}$$

Let us say the cumulative error process is in state k when $e(t) = k$, i.e. the cumulative number of error bits equals k at time t. It is then clear that the cumulative error process can be described by a Markov process, since there holds

$$P\{e(t+1) = k+1\} = P\{\epsilon(t+1) = 1 | e(t) = k\} + P\{\epsilon(t+1) = 0 | e(t) = k+1\}.$$

This implies that all properties of Markov process apply to it. Specifically, $E(t)$ is a pure birth process.

Assume that $\epsilon(t), t = 1,\ 2, \ldots,$ are i.i.d. random variables. Then, it is easy to verify that the cumulative error process has stationary

increments, i.e. $e(s, s+t) =_{st} e(t)$ for all $s, t \geq 0$, where $e(s, s+t) \equiv e(s+t) - e(t)$, and '$=_{st}$' denotes *stochastically equal*[1].

A.1.2 Service Model Definition

Having introduced the error processes, we now use them to model a server. Specifically, we model the server with two processes: an ideal service process $\mathcal{S}$ and an error process $\hat{E}$ as shown in Fig.A.1. The ideal service process denotes the cumulative amount of service (in bits) that the server provides no matter whether there is transmission error occurring in delivering service. The error process represents the number of errors (in bits) in the service. We call this model the *service model with error process*.

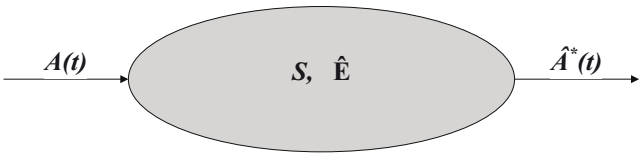

Figure A.1: Service model with error process

Consider a communication link as a simple example. Suppose the link capacity is C bps. In this case, the ideal service process $\mathcal{S}$ has a (space-domain) strict service curve [37]: $\mathcal{S}(t) = C \cdot t$. While traffic transmitted over this link may be received with errors, the damaged transmission or service is not excluded from C. To take errors into account in the model, we use the error process $\hat{E}(t)$. While on high-quality wired links (e.g. optical fiber links), errors may rarely happen and $\hat{E}(t)$ can be ignored, they can happen frequently on wireless links and other wired links (e.g. DSL links).

The model views the system as a black-box because we are mainly concerned about the difference between the initial input $\mathcal{A}(t)$ and the final output $\hat{\mathcal{A}}^*(t)$.

Note that when the instant error process is considered in the service model, the system delay must be taken into account when comparing the output bit with its corresponding input bit, because they must refer to the same information bit on the flow. Suppose bit $u(t)$ is

[1]The knowledge of stochastic ordering has been introduced in Section 2.3.3.

input to the system at time t and is delayed until time $t + \mathcal{D}_{u(t)}$ to leave the system, where $\mathcal{D}_{u(t)}$ is the system delay for $u(t)$. Then, the corresponding output bit $\hat{u}^*(t + \mathcal{D}_{u(t)})$ is obtained by

$$\hat{u}^*(t + \mathcal{D}_{u(t)}) = u(t) \oplus \epsilon(t + \mathcal{D}_{u(t)}). \tag{A.4}$$

Accordingly, the equation below holds

$$\epsilon(t + \mathcal{D}_{u(t)}) = u(t) \oplus \hat{u}^*(t + \mathcal{D}_{u(t)}). \tag{A.5}$$

In addition, the cumulative error is

$$e(t + \mathcal{D}_{u(t)}) = \sum_{s=0}^{t} \epsilon(s + \mathcal{D}_{u(s)}). \tag{A.6}$$

Due to delay, some care is needed in studying the service model with error process, which will be discussed in the next section where the focus is on investigating the concatenation property of such service models.

A.2 Concatenation Property

This section examines the concatenation property for the proposed service model. As discussed in the literature, the concatenation property is both useful and important for network service guarantee analysis, since it can result in much improved results [17] [29] [61] [66] [67] [78]. As in the previous section, we also assume bit-stream traffic for the explanation in this section.

We first focus on the simplest single server case. Then, we study two error processes in tandem. With the obtained results, we thirdly investigate the concatenation of multiple servers and present the concatenation property.

A.2.1 Single Server: Ideal Service Process + Error Process

Consider the single server shown in Fig.A.1. We are now interested in decoupling the ideal service process and the error process, and use two

virtual servers to represent the real server. One virtual server represents the ideal service process, and the other virtual server represents the error process with no delay. We consider two scenarios where the two virtual servers are arranged differently as shown in Fig.A.2.

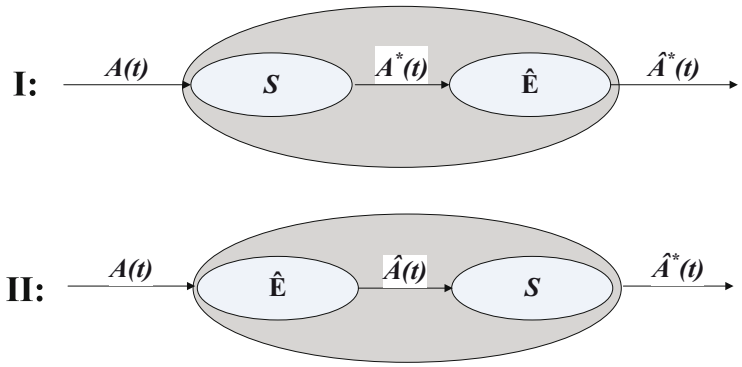

Figure A.2: Concatenation of Service Process and Error Process

Scenario I

As Fig.A.2 (I) illustrates, the initial input $\mathcal{A}$(t) first flows into a server with the ideal service process $\mathcal{S}$ and then goes through a server that has no delay but with error process $\hat{E}$. Let $\hat{\mathcal{A}}^*(t)$ be the actual output process from the error server. Denote by $\mathcal{A}^*(t)$ the output process from $\mathcal{S}$, which is also the input process to $\hat{E}$. Denote by u^* (in $\mathcal{A}^*$) the corresponding output bit of u in $\mathcal{A}$ from $\mathcal{S}$.

For bit $u^*(t)$ in $\mathcal{A}^*(t)$, its corresponding output from $\hat{E}(t)$, denoted by $\hat{u}^*(t)$, satisfies:

$$\hat{u}^*(t) = u^*(t) \oplus \epsilon(t).$$

Since $\mathcal{S}$ is the ideal service process that only delays traffic but does not introduce errors, we have

$$\hat{u}_I^*(t + \mathcal{D}_{u(t)}) = u^*(t + \mathcal{D}_{u(t)}) \oplus \epsilon(t + \mathcal{D}_{u(t)}), \tag{A.7}$$

where $\mathcal{D}_{u(t)}$ is the delay for bit $u(t)$ passing through the ideal server with service process $\mathcal{S}$.

With the above investigation, the output cumulative error can be expressed as:

$$e_I(t + \mathcal{D}_{u(t)}) = \sum_{s=0}^{t} \epsilon(s + \mathcal{D}_{u(s)}). \tag{A.8}$$

Scenario II

Fig.A.2 (II) shows an alternative to represent the real server with two virtual servers. In this scenario, the flow first traverses the error process before flowing into the service process. Let $\hat{\mathcal{A}}(t)$ denote the output from the error process $\hat{E}(t)$, which is also the input to the ideal service process $\mathcal{S}$. Denote by $\hat{u}$ (in $\hat{\mathcal{A}}(t)$) the corresponding output bit of u in $\mathcal{A}$ from $\hat{E}(t)$.

For bit $u(t)$ in $\mathcal{A}(t)$, its corresponding output from $\hat{E}(t)$ is:

$$\hat{u}(t) = u(t) \oplus \epsilon(t).$$

The cumulative error in $\hat{\mathcal{A}}(t)$ by time t is

$$e_{\hat{E}}(t) = \sum_{s=0}^{t} \epsilon(s).$$

Since the error process does not introduce delay and the ideal service process $\mathcal{S}$ does not introduce error, the output of $u(t)$ from the error server now becomes:

$$\hat{u}^*_{II}(t + \mathcal{D}_{u(t)}) = \hat{u}(t) = u(t) \oplus \epsilon(t). \tag{A.9}$$

Then, under this scenario, the output cumulative error can be expressed as:

$$e_{II}(t + \mathcal{D}_{u(t)}) = \sum_{s=0}^{t} \epsilon(s). \tag{A.10}$$

Comparison

Interestingly, the right side of Eq.(A.7) and that of Eq.(A.9) are different, even though both scenarios are used to represent the same error server.

Specifically, for the right side of Eq.(A.7) and that of Eq.(A.9), we have

$$u^*(t + \mathcal{D}_{u(t)}) = u(t).$$

However, some care is needed to treat $\epsilon(t + \mathcal{D}_{u(t)})$ and $\epsilon(t)$.

In the rest, we assume the service process and the error process are independent of each other, and the instant error process is comprised of i.i.d. random variables. Under this assumption, it can be verified

$$\epsilon(t + \mathcal{D}_{u(t)}) =_{st} \epsilon(t) \tag{A.11}$$

and hence $u^*(t+\mathcal{D}_{u(t)}) \oplus \epsilon(t+\mathcal{D}_{u(t)}) =_{st} u(t) \oplus \epsilon(t)$. Or, in other words, there holds

$$\hat{u}_I^*(t + \mathcal{D}_{u(t)}) =_{st} \hat{u}_{II}^*(t + \mathcal{D}_{u(t)}).$$

In addition, for the cumulative error,

$$e_I(t + \mathcal{D}_{u(t)}) =_{st} e_{II}(t + \mathcal{D}_{u(t)}). \tag{A.12}$$

Formally, we have the following result:

Theorem 19. ***(Concatenation of a Service Process and an Error Process).***

Consider a flow traversing a system that consists of an ideal service process $\mathcal{S}$ and an error process $\hat{E}(t)$. Assume $\mathcal{S}(t)$ and $\hat{E}(t)$ are independent of each other and $\hat{E}(t)$ is comprised of i.i.d. random variables. Then, for the flow, the instant error introduced by the system remains stochastically unchanged no matter how the error process and the ideal service process are ordered, so does the cumulative error.

Note that without the assumption in Theorem 19, the stochastic equivalence between the two alternatives of ordering these two processes may not hold in general.

A.2.2 Two Error Processes

The concatenation of two pure error processes may be difficult to match to real network scenarios. However, if the concatenation property also holds for error processes connected in tandem, we may separate the concatenation of servers into two process groups, the ideal service process group and the error process group, and then analyze these two process groups accordingly.

Consider the concatenation of two error processes, $\hat{E}^1(t)$ and $\hat{E}^2(t)$, and treat the concatenated system as a blackbox. Let $\mathcal{A}(t)$ and $\hat{\mathcal{A}}(t)$ denote the input process to the system and the output process of the system, respectively. Since we have assumed no delay in any error process, it follows immediately that there is no delay in the concatenation of multiple error processes.

Then, by time t, the cumulative error introduced by the concatenation system is given by

$$e(t) = \sum_{s=0}^{t} u(t) \oplus \hat{u}(t)$$

where $u(t)$ and $\hat{u}(t)$ respectively denote the input bit and output bit of the system at time t.

Similarly as discussed in Section A.2.1, there are two alternatives to order the two error processes in the blackbox.

Scenario I: $\hat{E}^1(t)$ followed by $\hat{E}^2(t)$

In this case, the input to $\hat{E}^1(t)$ is $\mathcal{A}(t)$, which is the same as the input to the black-box, and the output of $\hat{E}^2(t)$ is $\hat{\mathcal{A}}(t)$, which is the same as the output of the black-box. Denote by $\hat{\mathcal{A}}^1(t)$ the output of $\hat{E}^1(t)$, which is also the input to $\hat{E}^2(t)$. The final output from $\hat{E}^2(t)$ for bit $u(t)$ of $\mathcal{A}$ is

$$\hat{u}_I(t) = (u(t) \oplus \epsilon^1(t)) \oplus \epsilon^2(t).$$

Based on the associativity of "$\oplus$", $\hat{u}_I(t)$ can be written as

$$\hat{u}_I(t) = u(t) \oplus (\epsilon^1(t) \oplus \epsilon^2(t)) \tag{A.13}$$

with which, the instant error of bit $u(t)$ is $\epsilon^1(t) \oplus \epsilon^2(t)$.

Then, we obtain the cumulative error as

$$e_I(t) = \sum_{s=0}^{t} \epsilon^1(s) \oplus \epsilon^2(s). \tag{A.14}$$

Scenario II: $\hat{E}^2(t)$ followed by $\hat{E}^1(t)$

Following the same discussion as above, the system output of unit $u(t)$ is

$$\hat{u}_{II}(t) = u(t) \oplus (\epsilon^2(t) \oplus \epsilon^1(t)) \tag{A.15}$$

and the cumulative error process of the system is

$$e_{II}(t) = \sum_{s=0}^{t} \epsilon^2(s) \oplus \epsilon^1(s). \tag{A.16}$$

Comparison

Comparing the right side of Eq.(A.13) with that of Eq.(A.15), we know

$$e_I(t) = e_{II}(t).$$

Consequently, we also have

$$\hat{u}_I(t) = \hat{u}_{II}(t).$$

Based on the above discussion, we conclude:

Theorem 20. ***(Concatenation of two Error Processes).***

Consider a flow traversing a system that consists of two error processes, $\hat{E}^1(t)$ and $\hat{E}^2(t)$. For this system, both the instant system error process and the cumulative system error process do not change no matter how the two error processes constituting the system are ordered.

Theorem 20 can be directly extended to the concatenation of multiple error processes.

A.2.3 Multiple Servers

Based on the discussion in the previous two subsections, we study a system of multiple servers in tandem. We focus on the simple case of concatenating two servers, based on which, the results are generalized to multiple servers.

The concatenation of two servers can be viewed as four individual processes connected in tandem. Since we have discussed the concatenation of two processes in Sections A.2.1 and A.2.2, we shall only consider three scenarios as illustrated in Fig.A.3 for the concatenation of two servers.

As discussed in Section A.2.1, the delay in the service process should be considered. The delay of a bit $u(t)$ in the system is the

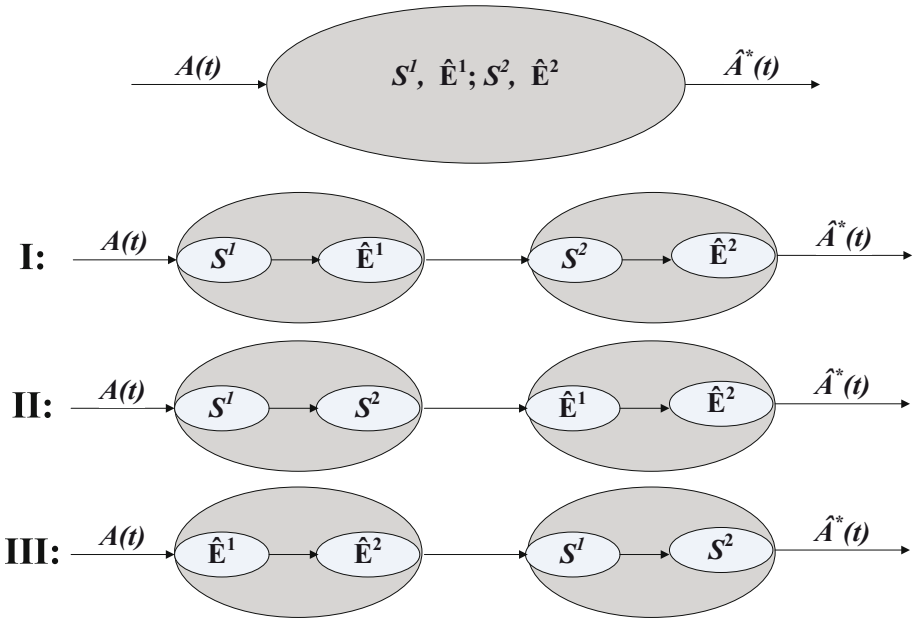

Figure A.3: Concatenation of Two Service Processes and Two Error Processes

summation of the corresponding delay in each of the two individual service processes. View this system as a blackbox with input $\mathcal{A}(t)$ and output $\mathcal{A}^*(t)$. The bit $u(t)$ in $\mathcal{A}(t)$ injected into this system at time t departures from the system at time $t + \mathcal{D}_{u(t)}$, where $\mathcal{D}_{u(t)}$ denotes the system delay of $u(t)$ and equals the summation of its delay in $\mathcal{S}^1(t)$, denoted by $\mathcal{D}^1_{u(t)}$, and that in $S^2(t)$, denoted by $\mathcal{D}^2_{u(t)}$. The cumulative error introduced by this system for $\mathcal{A}(t)$ is then

$$e(t + \mathcal{D}_{u(t)}) = \sum_{s=0}^{t} u(s) \oplus \hat{u}^*(s + \mathcal{D}_{u(s)}).$$

The procedure of analyzing four processes separately is similar to that of analyzing the single server and the two error processes. Let $u^{1s,*}$ denote the output of bit $u(t)$ from $\mathcal{S}^1$. Similarly, after bit $u(t)$ goes through $\mathcal{S}^1$ and $\mathcal{S}^2$ consecutively, its output is denoted by $u^{12s,*}$.

We omit the details and directly write the system output bit cor-

responding to $u(t)$ under each scenario as follows.

$$\begin{aligned}
\hat{u}_I^*(t+\mathcal{D}_{u(t)}) &= u^{1s,*}(t+\mathcal{D}^1_{u(t)}) \oplus (\epsilon^1(t+\mathcal{D}^1_{u(t)}) \oplus \epsilon^2(t+\mathcal{D}_{u(t)})) \\
\hat{u}_{II}^*(t+\mathcal{D}_{u(t)}) &= u^{12s,*}(t+\mathcal{D}_{u(t)}) \oplus (\epsilon^1(t+\mathcal{D}_{u(t)}) \oplus \epsilon^2(t+\mathcal{D}_{u(t)})) \\
\hat{u}_{III}^*(t+\mathcal{D}_{u(t)}) &= u(t) \oplus (\epsilon^1(t) \oplus \epsilon^2(t)).
\end{aligned}$$

Since an error process does not bring delay, it can then be verified that

$$u^{1s,*}(t+\mathcal{D}^1_{u(t)}) = u^{12s,*}(t+\mathcal{D}^1_{u(t)}) = u(t).$$

Assume that all service processes and error processes are independent of each other, and each instant bit error process is comprised of i.i.d. random variables. Then, with simple analysis, we can further conclude that

$$\hat{u}_I^*(t+\mathcal{D}_{u(t)}) =_{st} \hat{u}_{II}^*(t+\mathcal{D}_{u(t)}) =_{st} \hat{u}_{III}^*(t+\mathcal{D}_{u(t)}). \tag{A.17}$$

The following result summarizes the above discussion and extends to the case of multiple servers.

Theorem 21. ***(Concatenation of Multiple Error Servers).***

Consider a flow traversing a system $\mathcal{S}$ that consists of M ($M \geq 1$) ideal service processes, $\mathcal{S}^i$ (i=1...M), and N ($N \geq 1$) error processes, $\hat{E}^j$ ($j = 1...N$), where M and N are not necessarily equal. Assume all these processes are independent, and for each bit error process $\hat{E}^j$, it is comprised of i.i.d. random variables. Then, for this flow, the instant error introduced by the system is stochastically equal no matter how these processes are ordered, so is the cumulative error introduced by the system. Particularly, for the instant error process introduced by the system, there holds:

$$\epsilon(\tau) =_{st} \epsilon^1(\tau) \oplus \epsilon^2(\tau) \cdots \oplus \epsilon^N(\tau), \tag{A.18}$$

and for the cumulative error process, there holds:

$$e(t) =_{st} \sum_{s=0}^{t} \left[\epsilon^1(s) \oplus \epsilon^2(s) \cdots \oplus \epsilon^N(s)\right]. \tag{A.19}$$

A.3 Stochastic Error Curve

In the previous Sections A.1 and A.2, we have assumed bit-stream traffic and focused on bit error processes. However, in real networks, if there are bit errors in a received transmission unit, e.g. a packet, the whole unit may be counted as error unit. To address this, we generalize the definitions of error processes and the service model to transmission-unit-level. In the rest of this chapter, we adopt the **discrete time model**[2].

A.3.1 Error Processes and Service Model

We again use two stochastic processes $\hat{E}(t)$ and $E(t)$ to describe errors. However, an error here should be interpreted as an **error unit**, so when the error is counted, it is the total number of bits of the error unit that should be counted. These two error processes are respectively called the *instant error process* and the *cumulative error process*. With these generalized error process definitions, the service model is defined the same as in Section A.1.2.

Definition 17.

1. The instant error process $\hat{E}(t)$ *is a collection of random variables* $\{\epsilon(t), t = 1, 2, ...\}$ *with* $\epsilon(0) = 0$*, where* $\epsilon(t)$ *is the length of error unit at time* t*.*

2. The cumulative error process $E(t)$ *is a collection of random variables* $\{e(t) = \sum_{s=0}^{t} \epsilon(s), t = 1, 2, ...\}$ *with* $e(0) = 0$*, where* $e(t)$ *represents the cumulative amount (in bits) of error units in interval* $(0, t]$*.*

It is worth highlighting that the above error process definitions are different from Definition 16, in which, the definitions are respectively

[2]A transmission unit is considered out of a server when and only when its last bit has been served by this server. A transmission unit can be served only when its last bit has arrived.

special case of the above generalized definitions by considering bit as the unit. However, with the generalized definitions, if $u(t)$ is the input unit at time t and the corresponding output unit is $\hat{u}^*(t + d_{u(t)})$ where $d_{u(t)}$ denotes the delay of the unit in the system, we generally do not have (A.4), or in other words,

$$\hat{u}^*(t + d_{u(t)}) \neq u(t) \oplus \epsilon(t + d_{u(t)}).$$

We can use a different way to express the instant error in terms of the input and the corresponding output:

$$\epsilon(t + d_{u(t)}) = \begin{cases} 0, & u(t) \oplus \hat{u}^*(t + d_{u(t)}) = 0 \\ \text{length of u(t)}, & u(t) \oplus \hat{u}^*(t + d_{u(t)}) \neq 0 \end{cases}$$

Nevertheless, similar concatenation property investigated in Section A.2 holds.

A.3.2 The Concatenation Property

In the rest of this chapter, we assume the instant error process $\hat{E}(t)$ is comprised of i.i.d random variables, i.e. $\epsilon(t), t = 1, 2, ...,$ are i.i.d. Under this assumption, it is easy to verify that the cumulative error process $E(t)$ has stationary increments, i.e. $E(0, t) =_{st} E(s, s + t)$ for all $s, t \geq 0$.

The following result presents the concatenation property for the generalized service model with error process, corresponding to results in Section A.2. Since it can be verified that Theorem 19 and Theorem 20 are only special cases of Theorem 21, we only introduce the corresponding result of Theorem 21 in the following.

Theorem 22. ***(Concatenation of Multiple Servers).***

Consider a flow traversing a system that consists of M ($M \geq 1$) ideal service processes, $\mathcal{S}^i$ (i=1...M), and N ($N \geq 1$) error processes, $\hat{E}^j$ ($j = 1...N$), where M and N are not necessarily equal. Assume all these processes are independent, and for each instant error process $\hat{E}^j$, it is comprised of i.i.d random variables. Then, for this flow, the instant error introduced by the system is stochastically equal no matter

how these processes are ordered, so is the cumulative error introduced by the system. Particularly, for the instant error process, there holds:

$$\epsilon(t) \leq_{st} \epsilon^1(t) + \epsilon^2(t) \cdots + \epsilon^N(t), \tag{A.20}$$

and for the cumulative error process, there holds:

$$e(t) \leq_{st} \sum_{s=0}^{t} \left[\epsilon^1(s) + \epsilon^2(s) \cdots + \epsilon^N(s)\right]. \tag{A.21}$$

Proof. The proof of Theorem 22 follows similar discussion in Section A.2 by exploring the independence assumption of the error processes. We use a simple example to intuitively explain Eq.(A.20). As shown in Figure A.4, suppose we take a snapshot of a system consisting of two error processes at time τ. We capture two units, m and n which are leaving E^1 and E^2 respectively. The instant error process of this system can be expressed as

$$\epsilon(\tau) = \epsilon^1(\tau) + \epsilon^2(\tau), \tag{A.22}$$

where $\epsilon^1(\tau)$ equals either 0 or the length of unit m, and $\epsilon^2(\tau)$ equals either 0 or the length of unit n.

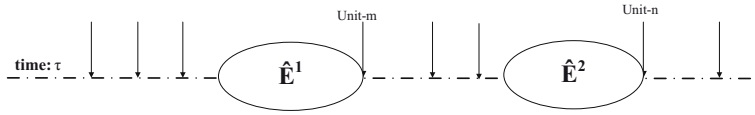

Figure A.4: Concatenation of two error processes

By taking into consideration the *i.i.d.* instant errors, we generalize the example of two error processes to the case of multiple error processes and obtain Eq.(A.20). Then Eq.(A.21) is readily derived from the definition of the cumulative error process. □

A.3.3 Stochastic Error Curve

So far we have introduced the instant error process and the cumulative error process concepts. Another concept that is helpful in service guarantee analysis is related to the cumulative error in a time interval. Specifically, we denote by $E(s,t)$ the cumulative number of errors in time interval $(s,t]$. From the definition of the cumulative error process, it is know that $E(s,t) = E(t) - E(s)$.

We then define a stochastic error curve model as follows.

Definition 18. *A system is said to introduce to a flow a (space-domain) v.b.c stochastic error curve α_e with bounding function f_e, if for all $s, t \geq 0$ and all $x \geq 0$, there holds*

$$P\{\sup_{0 \leq s \leq t} [E(s,t) - \alpha_e(t-s)] > x\} \leq f_e(x).$$

Alert reader may have noticed that Definition 18 is similar to the definition of the space-domain v.b.c stochastic arrival curve in Definition 4[3]. In fact, if we view the error process as a virtual 'error flow', Definition 18 implies that the error process has a v.b.c SAC [61].

Based on Eq.(A.20), under the same assumption as Theorem 22, it can be verified that

$$E(s,t) \leq_{st} E^1(s,t) + E^2(s,t) \cdots + E^N(s,t)$$

with which, we can further have the following representation of the concatenation property. Its proof follows easily from Lemma 1 and is omitted.

Theorem 23. ***(Concatenation Property).***

Consider a flow traversing a system that is a tandem of N servers. Suppose each server introduces an error process $E^i(t), i = 1, \ldots, N$, to its input. Assume all the ideal service processes and error processes are independent of each other. Also assume the corresponding

[3]Following the same idea as in defining the various variations of the stochastic arrival curve [61] [67], similar variations of stochastic error curve can be defined accordingly.

instant error process of each error server is comprised of i.i.d random variables $\epsilon^i(1), \epsilon^i(2), \cdots, (i = 1, \ldots, N)$. *Then, if the error process of each server has a stochastic error curve* α_e^i *with bounding* f_e^i, *the error process of the system has a stochastic error curve* α_e *with bounding function* f_e*:*

$$\alpha_e(t) = \alpha_e^1(t) + \alpha_e^2(t) \cdots + \alpha_e^N(t) \quad \text{(A.23)}$$

$$f_e(x) = 1 - \bar{f_e^1} * \bar{f_e^2} \cdot * \bar{f_e^N}(x), \quad \text{(A.24)}$$

where $\bar{f_e^i} = 1 - [f_e^i]_1$.

Note that Eq.(A.24) is obtained from Lemma 1.

A.4 Error Handling and Performance Bounds

Having introduced the service model with error process and its concatenation property, in this section we consider a simple network. We apply two different error handling methods in this simple network, then analyze the delay and backlog performance under two error handling methods, and compare the obtained performance bounds.

When an error is detected, the network has many ways to handle it. For example, the sender may re-transmit, or the error unit is simply dropped by the receiver and no re-transmission is needed. The former method is important for correctness-critical applications such as file transfer, while the latter can be used for delay-critical applications such as real-time inter-active applications which can tolerate a certain amount of errors or packet loss.

The simple network consists of a single link with one input flow. The link is error-prone. When there is no error in the ideal case, the service rate of the link is C. In other words, the ideal service process of the link has space-domain strict service curve $\beta(t) = C \cdot t$. Suppose the link error can be modeled with an error process $E(t)$ that has a stochastic error curve α_e with bounding function f_e. In addition,

the input flow $\mathcal{A}$ has a space-domain stochastic arrival curve α_a with bounding function f_a. For ease of expression and comparison, we assume $\alpha_a(t) = r_a \cdot t$ and $\alpha_e(t) = r_e \cdot t$, and $r_a + r_e < C$.

A.4.1 Scenario I: Delay Model

In this scenario, we assume that when an error happens, the sender simply re-transmits the corresponding unit. This may also be viewed as if the sender holds the transmission whenever the link is not error-free and sends immediately when the link becomes error-free.

In this case, the error process can indeed be thought of as an impairment process under the (space-domain) stochastic strict server model [61]. As shown in Figure A.5, under this way of error handling, the transmission units can be considered as passing through a *virtual delay process* before they finally reach the ideal service process.

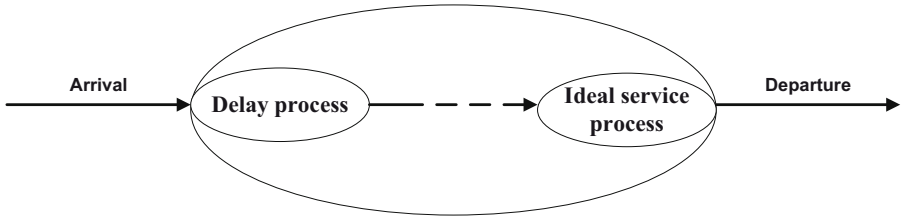

Figure A.5: Delay Model

Therefore, the network can be viewed as a stochastic strict server providing strict service curve $\beta(t) = C \cdot t$ with impairment process $I(t) = E(t)$, which has a *v.b.c* stochastic arrival curve α_e with bounding function f_e. With Definition 7, we obtain the following service guarantees.

- From Theorem 3, the backlog $\mathcal{B}(t)$ at the sender is bounded by:

$$\begin{aligned} P\{\mathcal{B}(t) > x\} &\leq 1 - \bar{f}_a * \bar{f}_e\Big(x - \sup_{s \geq 0}\big[r_a \cdot s - (C - r_e) \cdot s\big]\Big) \\ &= 1 - \bar{f}_a * \bar{f}_e(x); \end{aligned} \tag{A.25}$$

- From Theorem 5, the delay $\mathcal{D}(t)$ is bounded by

$$P\{\mathcal{D}(t) > \frac{x}{C - r_e}\} \leq 1 - \bar{f}_a * \bar{f}_e(x), \tag{A.26}$$

for all $t \geq 0$ and $x \geq 0$, where $\bar{f}_a(x) = 1 - [f_a(x)]_1$ and $\bar{f}_e(x) = 1 - [f_e(x)]_1$.

A.4.2 Scenario II: Loss Model

In this scenario, the sender does not care about the transmission error occurrence and transmits as if the link always operates in the ideal error-free condition. As shown in Figure A.6, the output traffic from the channel may contain errors which make the corresponding transmission units useless.

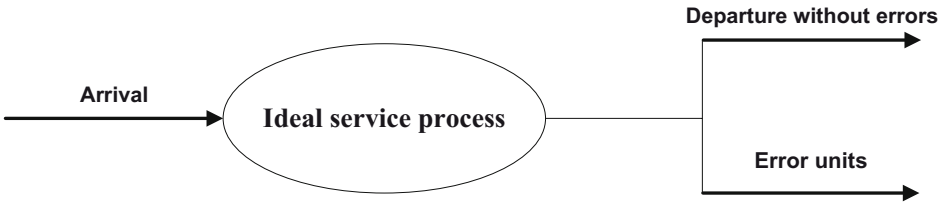

Figure A.6: Loss Model

Then the network can be viewed as a strict server providing strict service curve $\beta(t) = C \cdot t$ to the input, which implies a (weak) service curve $\beta(t) = C \cdot t$ with bounding function $g(x) = 0$. Similarly, with Definition 7, we obtain the following service guarantees.

- From Theorem 2, the backlog $\mathcal{B}(t)$ at the sender is bounded by

$$\begin{aligned} P\{\mathcal{B}(t) > x\} &\leq f_a \otimes g\Big(x - \sup_{s \geq 0}\big[r_a \cdot s - C \cdot s\big]\Big) \\ &= f_a(x); \end{aligned} \tag{A.27}$$

- From Theorem 4, the delay $\mathcal{D}(t)$ is bounded by

$$P\{\mathcal{D}(t) > \frac{x}{C}\} \leq f_a(x), \tag{A.28}$$

  for all $t \geq 0$ and $x \geq 0$.

While in Scenario I, there is no error at the receiver side, in the second scenario, the error rate by time t, defined as $\bar{\epsilon}(t) \equiv \frac{E[t]}{t}$, may be considered as an important performance measure. If all error units

are dropped at the receiver, $\bar{\epsilon}(t)$ can be viewed as the dropping rate. Since the error process has a stochastic error curve $\alpha_e = r_e \cdot t$ with bounding function f_e, it can then be easily verified that the error rate is bounded by:

$$P\{\bar{\epsilon}(t) > r_e\} \leq f_e(0). \tag{A.29}$$

A.4.3 Comparison

Comparing the performance bounds obtained under the two error handling methods, we can see that in terms of backlog bound and delay bound, the second error handling method gives shorter backlog and delay. To give a clearer picture about this, let $f_a(x) = f_e(x) = e^{-x}$. Then Table A.2 presents a comparison of the performance bounds obtained under the two error handling methods.

Handling	Bound on $P\{\mathcal{B}(t) > x\}$	Bound on $P\{\mathcal{D}(t) > d\}$
Method I	$(1+x)e^{-x}$	$[1+(C-r_e)d]e^{-(C-r_e)d}$
Method II	e^{-x}	e^{-Cd}

Table A.2: Comparison of delay bounds

On the other hand, the second method has to sacrifice the error performance and probably also the loss performance at the receiver side as a compromise.

A.5 Conclusion

In this chapter we introduced a service model which concretizes the generic impairment process by defining an error process. The error processes characterize transmission errors. An ideal service process and an error process are used to model the behavior of a server. Much of the study has been devoted to deriving the concatenation property for this service model, which is an important property for network calculus.

We started with assuming bit-stream traffic to exploit the XOR relationship among the input bit, the corresponding output bit and the error bit. We also extended the model to more realistic configuration

by using generalized error processes. For both cases, we proved that under some independence assumption, the error performance of a tandem system is stochastically equal no matter how the error processes in the system are ordered. In addition, we defined the stochastic error curve and derived its concatenation property. Moreover, to demonstrate the use of the proposed service model, we studied the service guarantee performance of a simple network under two error handling methods. While these two error handling methods are intuitively simple, the importance of introducing the service model is mostly revealed: error handling has significant impact on the system service guarantee performance, and the proposed service model can facilitate the analysis.

Many networks provide service stochastically due to inevitable unreliability, such as wireless networks where wireless channels are inherently error-prone. For such networks, only few results have been presented [46] [66] and the progress of their service guarantee analysis is however far-behind the progress of their wide implementation. We believe the analysis in this chapter sheds light on and makes one step forward towards studying these networks where the transmission error is an indispensable feature, which will be our future work.

Bibliography

[1] Wireless LAN medium access control (MAC) and physical layer (PHY) specification, 1999.

[2] *NIST/SEMATECH e-Handbook of Statistical Methods*. 2006.

[3] I. Adan and J. Resing. *Queueing Theory*. Eindhoven Univ. of Technology, 2001.

[4] R. Addie, P. Mannersalo, and I. Norros. Most probable paths and performance formulae for buffers with gaussian input traffic. *European Transactions on Telecommunications*, 13(3):183–196, 2002.

[5] A.Karasaridis and D. Hatzinakos. Network heavey traffic modeling using α-stable self-similar processes. *IEEE Trans. Commun.*, 49(7):1203–1214, July 2001.

[6] S. L. Albin. On poisson approximations for superposition arrival processes in queues. *Management Science*, 28(2):126–137, Feb. 1982.

[7] S. L. Albin. Approximating a point process by a renewal processes, II: Superposition arrival processes to queues. *Operations Research*, 32(5):1133–1162, 1984.

[8] M. A. Arcones. Arcones Manual for exam MLC: Superposition and Decomposition of a Poisson process. *http://www.math.binghamton.edu/arcones/exam-mlc/sect-11-4.pdf*, 2009.

[9] F. Bacceli, G. Cohen, G. J. Olsder, and J.-O. Quadrat. *Synchronization and Linearity: An Algebra for Discrete Event Systems*. Wiley, 1992.

[10] F. Baccelli, K. B. Kim, and D. D. Vleeschauwer. Analysis of the competition between wired, DSL and wireless users in an access network. In *Proc. IEEE INFOCOM*, 2005.

[11] G. Berger-Sabbatel, A. Duda, O. Gaudoin, M. Heusse, and F. Rousseau. Fairness and its impact on delay in 802.11 networks. In *Proc. IEEE GLOBECOM*, 2004.

[12] U. Narayan Bhat. Introduction to queueing theory. *Lecture Notes of EMIS 8372*, 2005.

[13] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE J. Select Areas Commun.*, 18(3):537–547, Mar. 2000.

[14] J.-Y. Le Boudec. Application of network calculus to guaranteed service networks. *IEEE Trans. Infor. Theory*, 44(3):1087–1096, May 1998.

[15] J.-Y. Le Boudec and P. Thiran. *Network Calculus: A Theory of Deterministic Queueing Systems for the Internet*. Springer-Verlag, 2001.

[16] M. Bredel and M. Fidler. Understanding fairness and its impact on quality of service in IEEE 802.11. In *Proc. IEEE INFOCOM*, 2009.

[17] A. Burchard, J. Liebeherr, and S. D. Patek. A min-plus calculus for end-to-end statistical service guarantees. *IEEE Trans. Information Theory*, 52(9):4105–4114, Sept. 2006.

[18] G. R. Cantieni, C. B. Q. Ni, and T. Turletti. Performance analysis under finite load and improvements for multirate 802.11. *Computer Comm.*, 28(10):1095–1109.

[19] C.-S. Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. Auto. Control*, 39(5):913–931, May 1994.

[20] C.-S. Chang. On the exponentiality of stochastic linear systems under the max-plus algebra. *IEEE Trans. Automatic Control*, 41(8):1182–1188, Aug. 1996.

[21] C.-S. Chang. *Performance Guarantees in Communication Networks*. Springer-Verlag, 2000.

[22] C.-S. Chang, R. L. Cruz, J.-Y. Le Boudec, and P. Thiran. A min, + system theory for constrained traffic regulation and dynamic service guarantees. *IEEE/ACM Tran. Networking*, 10(6):805–817, Dec. 2002.

[23] C.-S. Chang and Y. H. Lin. A general framework for deterministic service guarantees in telecommunication networks with variable length packets. *IEEE/ACM Trans. Automatic Control*, 46(2):210–221, Feb. 2001.

[24] P. Chatzimisios, A. C. Boucouvalas, and V. Vitsas. Optimimisation of RTS/CTS handshake in IEEE 802.11 wireless LANs for maximum performance. In *Proc. IEEE GLOBECOM Workshops*, 2004.

[25] J. Chen, A.-C. Pang, S.-T. Sheu, and H.-W. Tseng. High performance wireless switch protocol for IEEE 802.11 wireless networks. *J. Spec. Top. Mob. Netw. Appl.*, 10(5):741–751, 2005.

[26] X. Chen, H. Zhai, X. Tian, and Y. Fang. Supporting QoS in IEEE 802.11e wireless LANs. *IEEE Trans. Wireless Commun.*, 5(8):2217–2227, 2006.

[27] J. Cheo and N. B. Shroff. A central-limit-theorem-based approach for analyzing queue behavior in high-speed networks. *IEEE/ACM Trans. Networking*, 6(5):659–671, Oct. 1998.

[28] J.-W. Cho and Y. Jiang. Basic theorems on the backoff process in 802.11. *ACM SIGMETRICS Performance Evaluation Review*, 37(2), 2009.

[29] F. Ciucu, A. Burchard, and J. Liebeherr. Scaling properties of statistical end-to-end bounds in the network calculus. *IEEE/ACM Trans. Information Theory*, 52(6):2300–2312.

[30] F. Ciucu, A. Burchard, and J. Liebeherr. Scaling properties of statistical end-to-end bounds in the network calculus. *IEEE Trans. Information Theory*, 52(6):2300–2312, June 2006.

[31] J. A. Cobb. Preserving quality of service guarantees in spite of flow aggregation. *IEEE/ACM Trans. Networking*, 10(1):43–53, Feb. 2002.

[32] D. R. Cox and W. L. Smith. On the superposition of renewal processes. *Biometrika*, 41(1-2):91–99, 1954.

[33] R. L. Cruz. A calculus for network delay, part I: Network elements in isolation. *IEEE Trans. Information Theory*, 37(1):114–131, Jan. 1991.

[34] R. L. Cruz. A calculus for network delay, part II: Network analysis. *IEEE Trans. Information Theory*, 37(1):132–141, Jan. 1991.

[35] R. L. Cruz. Quality of service guarantees in virtual circuit switched networks. *IEEE J. Select Areas Commun.*, 13(6):1048–1056, Aug. 1995.

[36] R. L. Cruz. Quality of service management in integrated services networks. In *Proc. 1st Semi-Annual Research Review*, 1996.

[37] R. L. Cruz and C. Okino. Service guarantees for window flow control. In *Proc. 34th Annual Allerton Conference On Communication, Control and Computing*, 1996.

[38] C. Demichelis and P. Chimento. IP packet delay variation metric for IP performance metrics (IPPM). *IETF RFC3393*, 2002.

[39] J. L. Doob. *Stochastic Processes*. Wiley, 1953.

[40] E. O. Elliott. Estimates of error rates on bursty-noise channels. *Bell Systems Technical Journal*, 42(9):1977–1997, Sept. 1963.

[41] K. M. F. Elsayed and H. G. Perros. The superposition of discrete-time markov renewal processes with an application to statistical multiplexing of bursty traffic sources. *Applied Mathematics and Computation*, 15(1):43–62, Oct. 2000.

[42] A. Elwalid and D. Mitra. Effective bandwidth of general markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. Networking*, 1(33):329–343, June 1993.

[43] T. S. Eugene Ng, I. Stoica, and H. Zhang. Packet fair queueing algorithms for wireless networks with location-dependent errors. In *Proc. IEEE INFOCOM*, 1998.

[44] D. Ferrari. Client requirements for real-time communication services. *IEEE Commun. Magazine*, 28(11):65–72, Nov. 1990.

[45] M. Fidler. An end-to-end probabilistic network calculus with moment generating functions. In *Proc. IEEE IWQoS*, 2006.

[46] M. Fidler. A network calculus approach to probabilistic quality of service analysis of fading channels. In *Proc. IEEE GLOBECOM*, 2006.

[47] M. Fidler. A survey of deterministic and stochastic service curve models in the network calculus. *IEEE Commun. Surveys and Turotials*, 12(1):59–86, Feb. 2010.

[48] C.-H. Gan and Y.-B Lin. An effective power conservation scheme for IEEE 802.11 wireless networks. *IEEE Trans. Vehicular Technology*, 58(4):1920–1929, May 2009.

[49] M. Garetto and D. Towsley. Modeling, simulation and measurements of queueing delay under long-tail internet traffic. In *Proc. ACM Sigmetrics*, 2003.

[50] E. N. Gilbert. Capacity of a bursty-noise channel. *Bell Systems Technical Journal*, 39(5):1253–1265, Sept. 1960.

[51] A. J. Goldsmith and P. P. Varaiya. Capacity, mutual information, and coding for finite-state markov channels. *IEEE Trans. Information Theory*, 42(3):868–886, May 1996.

[52] P. Goyal, S. S. Lam, and H. M. Vin. Determining end-to-end delay bounds in heterogeneous networks. *Multimedia System*, 5(3):157–163, May 1997.

[53] P. Goyal and H. M. Vin. Generalized guaranteed rate scheduling algorithms: A framework. *IEEE/ACM Trans. Networking*, 5(4):561–571, Aug. 1997.

[54] N. Gupta and P. R. Kumar. A performance analysis of the 802.11 wireless LAN medium access control. *Communications in Information and Systems*, 3(4):279–304, Sept. 2004.

[55] M. Hassan, M. M. Krunz, and I. Matta. Markov-based channel characterization for tractable performance analysis in wireless packet networks. *IEEE Trans. Wireless Commun.*, 3(3):821–831, May 2004.

[56] T. Holliday, A. Goldsmith, and P. Glynn. Capacity of finite state markov channels with general inputs. In *Proc. IEEE International Symposium on Information Theory*, 2003.

[57] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge, 1987.

[58] ITU-TSS Study Group 13. Recommendation I.371 traffic control and congestion control in B-ISDN. 1995.

[59] R. Jain. *Art of Computer Systems Performance Analysis Techniques For Experimental Design Measurements Simulation and Modeling*. John Wiley & Sons, 1991.

[60] Y. Jiang. Delay bounds for a network of guaranteed rate servers with FIFO aggregation. *Computer Networks*, 40(6):683–694.

[61] Y. Jiang. A basic stochastic network calculus. In *Proc. ACM SIGCOMM 2006*, 2006.

[62] Y. Jiang. Internet quality of service - architectures, approaches and analyses. *http://www.q2s.ntnu.no/ jiang/Notes.pdf*, 2006.

[63] Y. Jiang. Per-domain packet scale rate guarantee for expedited forwarding. *IEEE/ACM Tran. Networking*, 14(3):630–643, June 2006.

[64] Y. Jiang. Network calculus and queueing theory: Two sides of one coin. In *Proc. ICST ValueTools*, 2009.

[65] Y. Jiang. A note on applying stochastic network calculus. *http://q2s.ntnu.no/ jiang/publications/note-on-applying-snetcal-v20100501.PDF*, 2010.

[66] Y. Jiang and P. J. Emstad. Analysis of stochastic service guarantees in communication networks: A server model. In *Proc. IWQoS*, 2005.

[67] Y. Jiang and Y. Liu. *Stochastic Network Calculus*. Springer, 2008.

[68] Y. Jiang, Q. Yin, Y. Liu, and S. Jiang. Fundamental calculus on generalized stochastically bounded bursty traffic for communication networks. *Computer Networks*, 53(12):2011–2021, Mar. 2009.

[69] Wagner Meira Jr. Parallel performance understanding via integration of modeling and diagnosis. *Phd Thesis Proposal at University of Rochester*, 1996.

[70] A. Karasaridis and D. Hatzinakos. A non-Gaussian self-similar process for broadband heavy traffic modeling. In *Proc. IEEE GLOBECOM*, 1998.

[71] O. Kella and W. Stadje. Superposition of renewal processes and an application to multi-server queues. *Statistics & Probability Letters*, 76:1914–1924, 2006.

[72] F. Kelly. Notes on effective bandwidths. *In Stochastic Networks: Theory and Applications, Royal Statistical Society Lecture Notes Series 4, Oxford University Press*, 1996.

[73] H. S. Kim and N. B. Shroff. Loss probability calculations and asymptotic analysis for finite buffer multiplexers. *IEEE/ACM Trans. Networking*, 9(6):755–768, Dec. 2001.

[74] L. Kleinrock. *Queueing Systems Volume I: Theory*. John Wiley & Sons, 1975.

[75] A. Kumar, E. Altman, D. Miorandi, and M. Goyal. New insights from a fixed-point analysis of single cell IEEE 802.11 WLANs. *IEEE/ACM Tran. Networking*, 15(3):588–601, June 2007.

[76] C. Y. T. Lam and J. P. Lehoczky. Superposition of renewal processes. *Advances in Applied Probability*, 23(1):64–85, March 1991.

[77] E. D. Lazowska, J. Zahorjan, G. S. Graham, and K. C. Sevcik. *Quantitative system performance: computer system analysis using queueing network models*. Prentice-Hall, 1984.

[78] C. Li, A. Burchard, and J. Liebeherr. A network calculus with effective bandwidth. *IEEE/ACM Trans. Networking*, 15(6):1442–1453, Dec. 2007.

[79] Q. Liu, S. Zhou, and G. B. Giannakis. Queueing with adaptive modulation and coding over wireless links: Cross-layer analysis and design. *IEEE Trans. Wireless Commun.*, 4(3):1142–1153, May 2005.

[80] Y. Liu, C.-K. Tham, and Y. Jiang. A calculus for stochastic QoS analysis. *Performance Evaluation*, 64(6):547–572, July 2007.

[81] S. Lu, V. Bharghavan, and R. Srikant. Fair scheduling in wireless packet networks. *IEEE/ACM Trans. Networking*, 7(4):473–489.

[82] D. Malone, K. Duffy, and D. Leith. Modeling the 802.11 distributed coordination function in non-saturated heterogeneous conditions. *IEEE/ACM Trans. Networking*, 15(1):159–172.

[83] P. Mannersalo and I. Norros. A most probable path approach to queueing systems with general gaussian input. *Computer Networks*, 40(3):399–412, Oct. 2002.

[84] S. Mao and S. S. Panwar. A survey of envelope processes and their applications in quality of service provisioning. *IEEE Commun. Surveys and Turotials*, 8(3):2–19, 2006.

[85] C. McDiarmid. Concentration. *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 1–46, 1998.

[86] A. K. Parekh and R. G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: the single node case. *IEEE/ACM Trans. Networking*, 1(3):344–357, June 1993.

[87] A. K. Parekh and R. G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: the multiple-node case. *IEEE/ACM Trans. Networking*, 2(2):137–150, April 1994.

[88] T. K. Philips and R. Nelson. The moment bound is tigher than chernoff's bound for positive tail probabilities. *American Statistician*, 49(2):175–178, 1995.

[89] Don R. Rice. An analytical model for computer system performance evaluation. *ACM SIGMETRICS Performance Evaluation Review*, 2(2).

[90] S. M. Ross. *Stochastic Processes*. John Wiley & Sons, 2nd edition, 1996.

[91] S. M. Ross. *Introduction to Probability Models*. Elsevier, 2006.

[92] T. Sakurai and H. Vu. MAC access delay of IEEE 802.11 DCF. *IEEE Tran. Wireless Commun.*, 6(5):1702–1710, May 2007.

[93] J. F. Shortle and P. H. Brill. Analytical distribution of waiting time in the M/iD/1 queue. *Queueing Systems*, 50(2):185–197, 2005.

[94] D. Starobinski and M. Sidi. Stochastically bounded burstiness for communication networks. *IEEE/ACM Trans. Information Theory*, 46(1):206–212, Jan. 2000.

[95] D. Stiliadis and A. Varma. Latency-rate servers: A general model for analysis of traffic scheduling algorithms. *IEEE/ACM Trans. Networking*, 6(5):611–624, Oct. 1998.

[96] P. Torab and E. W. Kamen. On approximate renewal models for the superposition of renewal processes. In *Proc. IEEE ICC*, 2001.

[97] D. Tutsch. *Performance Analysis of Network Architectures*. Springer-Verlag, 2006.

[98] H. S. Wang and N. Moayeri. Finite-state markov channel - a useful model for radio communication channels. *IEEE Trans. Veh. Technol.*, 44(1):163–171, Feb. 1995.

[99] D. Wu and R. Negi. Utilizing multiuser diversity for efficient support of quality of service over a fading channel. In *Proc. IEEE ICC*, 2003.

[100] K. Wu, Y. Jiang, and J. Li. On the model transformation in stochastic network calculus. In *Proc. IEEE IWQoS*, 2010.

[101] J. Xu, Y. Jiang, and A. Perkis. Towards analysis of intra-flow contention in multi-hop wireless networks. In *Proc. International Conference on Mobile Ad-hoc and Sensor Networks*, 2010.

[102] O. Yaron and M. Sidi. Performance and stability of communication network via robust exponential bounds. *IEEE/ACM Trans. Networking*, 1(3):372–385, June 1993.

[103] Q. Yin, Y. Jiang, S. Jiang, and P. Y. Kong. Analysis on generalized stochastically bounded bursty traffic for communication networks. In *Proc. 27th IEEE Local Computer Networks*, 2002.

[104] H. Zhai, Y. Kwon, and Y. Fang. Performance analysis of IEEE 802.11 MAC protocols in wireless LANs. *Wirel. Commun. Mob. Comput.*, 4(8):917–931, Jan. 2004.

[105] Q. Zhang and S. A. Kassam. Finite-state markov channel for rayleigh fading channels. *IEEE Trans. Commun.*, 47(11):1688–1692, Nov. 1999.

[106] Q. Zhao, D. H. K. Tsang, and T. Sakurai. A simple and approximate model for nonsaturated IEEE 802.11 DCF. *IEEE/ACM Trans. Networking*, 15(1):159–172.