A Min-plus System Theory Approach for Performance Evaluation and Quality-of-Service Provisioning in Wireless Data Networks

by

Farshid Agharebparast

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES (Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

October 2007

© Farshid Agharebparast, 2007
Abstract

Recent advances in various wireless data networking technologies and their successful implementations have put the dream of ubiquitous high-quality multimedia communications within our reach. However, achieving reasonable service quality is still a main challenge. It is necessary to evaluate the performance of a new or functional network and to study its service quality factors, such as delay and throughput. This thesis proposes and evaluates a set of analytical techniques based on the modern min-plus system theory that collectively presents an effective methodology for modeling, performance evaluation and quality of service (QoS) provisioning in wireless networks. Min-plus system theory is a successful alternative and complement to queuing theory, whose application in deterministic QoS guarantees is known as Network Calculus (NC) or $(\sigma, \rho)$-calculus.

The contributions of this work can be grouped under three main categories: slope domain modeling, stochastic min-plus system theorems and their applications in wireless networking, and filter-bank-based wireless channel models. Slope domain modeling is a useful technique whose role in min-plus theory is similar in concept to that of the Fourier domain in traditional system theory. It facilitates analysis of communications networks, and for certain problems is the only viable solution. This method is used in numerous applications, such as system identification, optimal smoothing and priority scheduling.

This thesis proposes a stochastic min-plus system theory approach that is used as a practical and efficient methodology for QoS performance evaluations in wireless networks, when the service can be defined by its second-order statistics. The method is then used in optimal traffic shaping and resource allocation for transmission over wireless fading channels. Finally, the
concept of load spectrum is introduced as the slope domain representation of this stochastic min-plus theory.

As an alternative method, a wireless link model based on the filter banks of NC elements is introduced. The model employs modular concatenations of selected NC components to represent the effect of a wireless link on the statistical QoS performance at the link layer. This technique is used in optimal smoothing to successfully solve the problem of playback delay and buffer size calculations for multimedia streaming via a wireless network.
# Table of Contents

Abstract .......................................................................................................................... ii  
Table of Contents ........................................................................................................... iv  
List of Tables ................................................................................................................... vii  
List of Figures .................................................................................................................. viii  
List of Acronyms ............................................................................................................. xii  
List of Symbols ............................................................................................................... xiii  
Acknowledgements .......................................................................................................... xiv  
Dedications ...................................................................................................................... xvi  

## Chapter 1: Introduction .............................................................................................. 1
- 1.1. Motivations and Objectives ................................................................................. 3
- 1.2. Main Contributions ............................................................................................. 5
- 1.3. Organization of the Thesis .................................................................................. 7

## Chapter 2: Background .............................................................................................. 8
- 2.1. A Brief History .................................................................................................... 9
- 2.2. Min-plus System Theory Fundamentals ............................................................. 10
- 2.3. Min-plus Algebra ............................................................................................... 14
- 2.4. Traffic Characterizations ................................................................................... 17
- 2.5. Deterministic Network Calculus ......................................................................... 19
- 2.6. Stochastic Min-plus Approaches ....................................................................... 22
- 2.7. Mathematical Transformations for Min-plus Systems ..................................... 25

## Chapter 3: Slope Domain Modeling and Analysis of Data Communication Networks ................................................................................................................. 26
- 3.1. Introduction ......................................................................................................... 26
- 3.2. The Slope Transform .......................................................................................... 28
- 3.3. Slope Domain Analysis of Networks ................................................................... 33
- 3.4. Non-Preemptive Priority Scheduling Example .................................................. 35
- 3.5. Network Identification Application .................................................................... 39
3.6. Optimal Smoothing for Multimedia Streaming ............................................. 46
3.7. Load Spectrum ....................................................................................... 49
3.8. Further Applications .............................................................................. 49
3.9. Comparing Fourier and Slope Domains ................................................ 50
3.10. Comparing Slope with the Legendre Transform and Fenchel Conjugate ................................................................. 51
3.11. Computational Improvement Using the Legendre Transform and Fenchel Conjugate ................................................................. 53
3.12. Summary ............................................................................................... 54

Chapter 4: A Stochastic Min-plus System Theory Approach to Performance Modeling of Wireless Networks.............................................. 55

4.1. Introduction ............................................................................................ 56
4.2. Background .......................................................................................... 57
4.3. Min-Plus Systems with Stochastic Inputs .............................................. 59
4.4. Min-Plus Stochastic Systems .................................................................. 64
4.5. Wireless Channel Modeling, Performance Evaluation and Cross-layer Design .............................................................................. 66
   4.5.1. System Model .................................................................................. 68
   4.5.2. Service Curve of the Channel ......................................................... 69
   4.5.3. Methodology .................................................................................. 72
4.6. Numerical Results ................................................................................. 75
4.7. Discussion .............................................................................................. 83
4.8. Summary ............................................................................................... 89

Chapter 5: Optimal Leaky-Bucket Shaper and Resource Allocation for the Transmission of Multimedia over Wireless Fading Channels..... 90

5.1. Introduction ............................................................................................ 91
5.2. Optimal Leaky-Bucket Shaper ............................................................... 92
5.3. System Model and Building Blocks ....................................................... 96
5.4. Methodology ........................................................................................ 98
5.5. Numerical Results ............................................................................... 99
5.6. Summary .............................................................................................. 116

Chapter 6: An Alternative Wireless Link Model Using Filter Banks of NC Elements ................................................................. 117

6.1. Introduction ............................................................................................ 117
6.2. Wireless Link Model ............................................................................ 119
6.3. Application in Wireless Video Streaming .............................................. 122
6.4. Numerical Example ............................................................................ 128
6.5. Discussion ........................................................................................... 129
6.6. Summary .............................................................................................. 131
Chapter 7: Summary and Future Work .................................................. 132

7.1. Summary .................................................................................. 132
7.2. Further Work ............................................................................ 134

References .......................................................................................... 136
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Properties of the lower slope transform</td>
<td>32</td>
</tr>
<tr>
<td>3.2</td>
<td>Properties of the upper slope transform</td>
<td>32</td>
</tr>
<tr>
<td>3.3</td>
<td>Lower slope transforms of some useful NC functions</td>
<td>35</td>
</tr>
<tr>
<td>4.1</td>
<td>Parameters used in the example and their values</td>
<td>75</td>
</tr>
<tr>
<td>5.1</td>
<td>Common parameters and their values used in the numerical examples</td>
<td>100</td>
</tr>
</tbody>
</table>
List of Figures

Figure 2.1 Min-plus system input/output representation .................................................. 11
Figure 2.2 NC-based wireless link modeling. ................................................................. 24
Figure 3.1 Modeling a network in NC and slope domains............................................... 31
Figure 3.2 Slope transforms of the peak-rate and rate-latency functions, and their convolution. .......................................................... 37
Figure 3.3 Slope transforms of the affine and rate-latency functions, and their convolution. 38
Figure 3.4 System identification by applying the slope domain analysis: (a) Plot of the service curve and inputs of the form b+at for different values of a. (b) Plot of the observed outputs minus the inputs. ........................................ 41
Figure 3.5 System identification by applying the slope domain analysis: Comparison of the numerical result with the analytical slope response ........................................................................ 42
Figure 3.6 System identification using OPNET simulations. ........................................... 43
Figure 3.7 Video streaming network topology. ............................................................... 46
Figure 4.1 Calculation of the autocorrelation. In the left figure, the two instances of time needed for the calculation of the autocorrelation are two possible instances observed at different times but for the same transmission time; on the right, they are at two different transmission times. ...................................................... 64
Figure 4.2 Statistical characterizations of the received traffic using the stochastic min-plus system theory. ................................................................. 66
Figure 4.3 The transmitted traffic (input) and 50 different instances of the channel output due to the time-varying Rayleigh fading channel. ......................................................... 67
Figure 4.4 System model and building blocks. .............................................................. 68
Figure 4.5 Graphical illustration of calculations of the second-order statistics of the delay. 74
Figure 4.6 An example of (a) the traffic delivered to the link layer, (b) shaped traffic ready to be transmitted through the channel, and (c) cumulative representations of traffic in a and b. ........................................................................... 76
Figure 4.7 3D plots of the autocorrelation of the service (a) and the output (b). The x axis designates the transmission time and the y axis represents 50 different instances of observations (with its mirror towards -50).

Figure 4.8 (a) Traffic and 50 instances of channel service curves. (b) Comparison of analytical and simulation results for mean outputs.

Figure 4.9 Graphical calculations of mean and worst delays from the mean and worst-case outputs.

Figure 4.10 Mean and worst delays for different values of $\mu/a$ with a 95% confidence interval (30 iterations of 50 instances each).

Figure 4.11 (a) Delay versus $b$ (shapers burst tolerance), when $E\{\gamma\}$ and $\mu/a$ are 10 dB and 0.8, respectively, with a 95% confidence interval (30 iterations of 50 instances each).

Figure 4.12 Delay versus average SNR ($E\{\gamma\}$) in dB when $\mu/a$ is 0.8, with a 95% confidence interval (30 iterations of 50 instances each).

Figure 4.13 Delay versus different values of shaping rate with the consideration of convolutional channel codes and bit error dependency.

Figure 5.1 Delay and buffer calculations. Method A: using the input and output traffic (a). Method B: using the arrival and service curves (b).

Figure 5.2 The effect of variations of arrival and service parameters on the maximum delay and buffer requirements for a rate-latency server and leaky-bucket-shaped traffic.

Figure 5.3 Calculation of mean and worst delays and buffer sizes. (a) Method A: using the input and output traffic flows. (b) Method B: using the arrival and service curves.

Figure 5.4 Plots of the mean delays versus $b$ and $a$.

Figure 5.5 Contours of plots in Figure 5.4. The values on the graphs are the corresponding mean delays.

Figure 5.6 Plots of the worst delays versus $b$ and $a$.

Figure 5.7 Contours of plots in Figure 5.6. The values on the graphs are the corresponding worst delays.

Figure 5.8 Plots of the mean buffer sizes versus $b$ and $a$. 
Figure 5.9  Contours of plots in Figure 5.8. The values on the graphs are the corresponding mean buffer sizes. .................................................. 106

Figure 5.10  Plots of the worst buffer sizes versus $b$ and $a$. ........................................ 107

Figure 5.11  Contours of plots in Figure 5.10. The values on the graphs are the corresponding worst buffer sizes. .................................................. 107

Figure 5.12  Plots of the mean delays versus $b$ and $a$. .................................................. 108

Figure 5.13  Contours of plots in Figure 5.12. The values on the graphs are the corresponding mean delays. .................................................. 108

Figure 5.14  Plots of the worst delays versus $b$ and $a$. .................................................. 109

Figure 5.15  Contours of plots in Figure 5.14. The values on the graphs are the corresponding worst delays. .................................................. 109

Figure 5.16  Plots of the mean buffer sizes versus $b$ and $\mu$. ....................................... 110

Figure 5.17  Contours of plots in Figure 5.16. The values on the graphs are the corresponding mean buffer sizes. .................................................. 110

Figure 5.18  Plots of the worst buffer sizes versus $b$ and $\mu$. ....................................... 111

Figure 5.19  Contours of plots in Figure 5.18. The values on the graphs are the corresponding worst buffer sizes. .................................................. 111

Figure 5.20  Mean delays versus $b$ and $a$ - dependent bit error scenario......................... 112

Figure 5.21  Worst delays versus $b$ and $a$ - dependent bit error scenario......................... 112

Figure 5.22  Mean buffer sizes versus $b$ and $a$ - dependent bit error scenario.................. 113

Figure 5.23  Worst buffer sizes versus $b$ and $a$ - dependent bit error scenario................. 113

Figure 5.24  Mean delay versus $b$ and $\mu$ - dependent bit error scenario. ..................... 114

Figure 5.25  Worst delay versus $b$ and $\mu$ - dependent bit error scenario..................... 114

Figure 5.26  Mean buffer sizes versus $b$ and $\mu$ - dependent bit error scenario. ............. 115

Figure 5.27  Worst buffer sizes versus $b$ and $\mu$ - dependent bit error scenario. ............. 115

Figure 6.1  Wireless link modeling. ................................................................................... 119
Figure 6.2  a) Clipper. b) Clipper equivalent with ARQ. ........................................... 120
Figure 6.3  Network schematic in the study of video streaming application. ....................... 123
Figure 6.4  NC modeling of $G(t)$ in Figure 6.3. ............................................................. 124
Figure 6.5  Playback delay and buffer size calculation for the Jurassic Park video trace streaming. ........................................................................................................ 128
### List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATM</td>
<td>Asynchronous Transfer Mode</td>
</tr>
<tr>
<td>ARQ</td>
<td>Automatic Repeat-reQuest</td>
</tr>
<tr>
<td>AWGN</td>
<td>Additive White Gaussian Noise</td>
</tr>
<tr>
<td>BER</td>
<td>Bit Error Rate</td>
</tr>
<tr>
<td>DEDS</td>
<td>Discrete Event Dynamic Systems</td>
</tr>
<tr>
<td>DiffServ</td>
<td>Differentiated Services</td>
</tr>
<tr>
<td>DTI</td>
<td>Dilation Translation-Invariant</td>
</tr>
<tr>
<td>ETI</td>
<td>Erosion Translation-Invariant</td>
</tr>
<tr>
<td>GPRS</td>
<td>General Packet Radio Service</td>
</tr>
<tr>
<td>GPS</td>
<td>Generalized Processor Sharing</td>
</tr>
<tr>
<td>IntServ</td>
<td>Integrated Services</td>
</tr>
<tr>
<td>MAN</td>
<td>Metropolitan Area Network</td>
</tr>
<tr>
<td>MER</td>
<td>Minimum Envelope Rate</td>
</tr>
<tr>
<td>MRSVP</td>
<td>Mobile Resource ReSerVation Protocol</td>
</tr>
<tr>
<td>NC</td>
<td>Network Calculus</td>
</tr>
<tr>
<td>pdf</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>PER</td>
<td>Packet Error Rate</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>QPSK</td>
<td>Quadrature Phase-Shift Keying</td>
</tr>
<tr>
<td>RSVP</td>
<td>Resource ReSerVation Protocol</td>
</tr>
<tr>
<td>SLA</td>
<td>Service Level Agreement</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
<tr>
<td>TSPEC</td>
<td>Traffic Specification</td>
</tr>
<tr>
<td>UMTS</td>
<td>Universal Mobile Telecommunications System</td>
</tr>
<tr>
<td>VBR</td>
<td>Variable Bit Rate</td>
</tr>
<tr>
<td>WFQ</td>
<td>Weighted Fair Queuing</td>
</tr>
<tr>
<td>WiMAX</td>
<td>Worldwide Interoperability for Microwave Access</td>
</tr>
<tr>
<td>WSS</td>
<td>Wide-Sense Stationary</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>$\inf$ or $\wedge$</td>
<td>Infimum</td>
</tr>
<tr>
<td>$\min$</td>
<td>Minimum</td>
</tr>
<tr>
<td>$\sup$</td>
<td>Supremum</td>
</tr>
<tr>
<td>$\otimes$</td>
<td>Convolution, min-plus convolution</td>
</tr>
<tr>
<td>$\oslash$</td>
<td>Min-plus deconvolution</td>
</tr>
<tr>
<td>$\ast$</td>
<td>Max-plus convolution</td>
</tr>
<tr>
<td>$\Box$</td>
<td>Max-plus deconvolution</td>
</tr>
<tr>
<td>$\zeta(.)$</td>
<td>Arrival curve (general notation)</td>
</tr>
<tr>
<td>$x(.)$</td>
<td>System input, input traffic</td>
</tr>
<tr>
<td>$h(.)$</td>
<td>Service curve, system function</td>
</tr>
<tr>
<td>$y(.)$</td>
<td>System output, output traffic</td>
</tr>
<tr>
<td>$S_L{\cdot}(\alpha), X_L(\cdot)$, or $X(\cdot)$</td>
<td>Lower slope transform</td>
</tr>
<tr>
<td>$S_U{\cdot}(\alpha)$ or $X_U(\cdot)$</td>
<td>Upper slope transform</td>
</tr>
<tr>
<td>$H(\alpha)$</td>
<td>Slope response</td>
</tr>
<tr>
<td>$F{\cdot}$</td>
<td>Fourier transform</td>
</tr>
<tr>
<td>$\Psi{\cdot}(\cdot)$</td>
<td>Legendre transform</td>
</tr>
<tr>
<td>$\Im{\cdot}$</td>
<td>Fenchel conjugate</td>
</tr>
<tr>
<td>$L(\cdot)$ or $L^h(\cdot)$</td>
<td>Min-plus system operator</td>
</tr>
<tr>
<td>$L^s(\cdot)$</td>
<td>Traditional system operator</td>
</tr>
<tr>
<td>$E{\cdot}(\cdot)$ or $\eta$</td>
<td>Expectation, statistical average</td>
</tr>
<tr>
<td>$R_x(\cdot), R_x(\cdot, \cdot)$</td>
<td>Correlation function</td>
</tr>
<tr>
<td>$\lambda(\cdot)$</td>
<td>Zero step function</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Signal to noise ratio</td>
</tr>
<tr>
<td>$Q(\cdot)$</td>
<td>Complementary error function</td>
</tr>
</tbody>
</table>
Acknowledgements

"A hundred times every day I remind myself that my inner and outer life are based on the labors of other [people], living and dead, and that I must exert myself in order to give in the same measure as I have received and am still receiving."

-- Albert Einstein, "The World as I See It"

I would like to gratefully thank my academic advisor Prof. Victor C.M. Leung for his support and supervision, as well as his faith in my research abilities. It was a great privilege and invaluable experience to work under the guidance of this world-renowned expert, scholar, and gentleman.

I would also like to express my gratitude to the supervisory committee members (Prof. Vikram Krishnamurthy, Prof. Son T. Vuong and Prof. Vincent Wong,), the university examiners (Prof. Philip Loewen and Prof. Rabab Ward), and the external examiner (Prof. Rene L. Cruz) for their valuable and constructive comments. I would also like to thank my current and former colleagues, and professors of the communications research group, who made my academic life at UBC rewarding and whose knowledge I benefited greatly from; with special thanks to Prof. Cyril Leung, Prof. Matthew Yedlin, Dr. Joo-Han Song, and Dr. Yaser P. Fallah. I would also like to thank the friendly and helpful staff at the ECE office, and Craig Wilson of ICICS for helping me to make this dissertation more readable.

I would like to appreciatively acknowledge the following federal, provincial and institutional organizations that graciously and generously supported my Ph.D. studies through postgraduate scholarships, fellowships or grants: the Natural Sciences and Engineering Research Council of Canada (NSERC PGS-B Postgraduate Scholarship and Grant RGPIN44286-04), the Canadian Wireless Telecommunications Association (CWTA Graduate Scholarship), the Advanced Systems Institute of British Columbia (ASI Graduate Scholarship), and the University of British Columbia (UBC Graduate Fellowship, Top-Up and Tuition Fee awards).

I salute UBC, this great university that has provided me with the opportunity to grow
through exposure to academia, arts, music and dance, sports and fitness, and culture. UBC is one of those rare places where a wealth of natural beauty, supreme-quality academia, and a rich multicultural environment coexist harmoniously. The UBC affiliate, St. John’s Residential Graduate College, fostered a magnificent opportunity for me to better grow towards being a well-rounded individual, with its vibrant international community, numerous seminars, and cultural events. While residing at the College, I met so many bright, talented and kind individuals whom I am proud to be friends with. I would like to pay tribute especially to Dr. Kai Hilpert, Paola Ceruti, Rebecca Klady, and Farzad Moin Afshar.

And last but not least, my utmost gratitude and respect to my parents, without whom I would be nowhere. They are everything to me and everything I am. It is my privilege to have had the love, support and encouragement of my beloved parents, sister and brothers throughout my life.
To my beloved parents.
Chapter 1. Introduction

With the ever-increasing popularity and usefulness of wireless data devices such as data-enabled phones, PDAs (personal digital assistant), and laptops, it is an unquestionable fact that a cost-effective high quality anywhere-anytime data communication capability is greatly desirable. The availability of such a service will benefit the general public with an integrated, ubiquitous means of communication. The objective of such systems is to make a wide range of data applications economically available, such as those provided by the Internet. The increasing availability of GPRS (general packet radio service) and UMTS (universal mobile telecommunications system) services, the great success of the commercialized wireless local area networks (WLAN; in particular the IEEE 802.11b standard, which commercially is referred to as Wi-Fi (wireless fidelity) to assure interoperability), as well as impending widespread implementations of wireless metropolitan area networks (interoperable via WiMAX forum) particularly advocate this proposition. In order to make such services economically available, it is necessary to find a set of technologies and methodologies to provide data and Internet resources, as well as high quality voice/video services, to mobile users through efficient wireless packet-switched networks with acceptable quality of service (QoS). This goal challenges researchers and telecommunications companies to not only devise the required infrastructure but to ensure a high degree of service quality.

The characteristics of multimedia applications are their wide range of delay, bandwidth, and content preservation requirements. The QoS required by each application differs and is specified by these characteristics. For some applications, a certain QoS guarantee is crucial for
their proper operation. For example, a multimedia streaming application needs to meet certain maximum delay and delay jitter criteria. This QoS is negotiated with the user through the service level agreement (SLA) phase. Based on the SLA, the network guarantees some level of service, provided that the user complies with the agreed upon conditions.

In order to implement a network capable of providing and guaranteeing a set of pre-defined QoS throughout a communication session, traffic management techniques are devised and used. Traffic management methods include resource allocation, shaping, scheduling, and buffer management mechanisms [1][2]. Performance evaluation is a critical process in which a number of analytical, simulation or emulation procedures are utilized to ensure that a particular network is capable of providing the required service, in all of the design, implementation and enhancement phases.

Min-plus system theory is a modern approach for the study and performance evaluation of a system where a number of customers compete to obtain the service of a number of available resources, such as queuing networks. The theory is based on a special algebra whose operators differ from those of the algebra that we traditionally deal with. Under this algebra, a traditionally non-linear system becomes linear and can be easily studied. The application of this theory in the deterministic study of data communications networks has been called Network Calculus (NC) or (σ, ρ)-calculus. Due to a book with this title, the term Network Calculus has been widely adopted for the application of deterministic min-plus theory in data networking. NC, a fairly new approach in networking analysis, provides further insight in to queuing theory. Some successful examples of NC applications for traffic management in wired networks are its widespread use in the study of ATM (asynchronous transfer mode), IntServ (integrated services) [3], DiffServ (differentiated services) [4][5], and service-curve-based schedulers such as the generalized processor sharing (GPS) [6].

Min-plus system theory is an alternative and complement to the traditional Markovian queuing theory. The main distinction starts with a different traffic characterization. Based on this traffic characterization, a set of theorems and concepts are introduced that provides a filtering theory for the analysis and design of computer communication networks.

Although NC is at an elaborate stage and has been widely used for the deterministic analysis of wired data networks, methods for considering servers with stochastic behavior are still under development. Wireless networks are perfect examples of such servers. This thesis contrib-
utes to the performance evaluation and QoS provisioning in wireless data networks based on the min-plus system theory and its application in some specific problems. In addition, some of the key methods introduced here complement min-plus system theory and enhance its general applicability in data networking. In the next subsection, the motivations and objectives inspiring this thesis are stated.

1.1. Motivations and Objectives

Min-plus system theory is a sophisticated modern approach to performance modeling and QoS provisioning in computer communication networks and discrete event dynamic systems (DEDS) [7]. NC or \((\sigma, \rho)\)-calculus is a successful framework based on the min-plus algebra for deterministic performance guarantees in queuing systems and has been successfully used in the analysis of networking principles [8][9]. It provides an alternative approach to the study of queuing systems, and facilitates the design and performance evaluation of data communication networks. In some scenarios that are difficult to analyze using traditional queuing theory, it is the preferred method [10][11][12]. An example of the successful application of network calculus is the introduction of service-curve-based schedulers. These schedulers are based on the widely adopted rate-latency service model and show link-sharing and decoupled delay bandwidth guarantees [6][13][14][15].

Inspired by the success of NC and its various applications, we were motivated to contribute in this thesis to the min-plus system theory and its application in wireless networking. In order to ensure complete compatibility with the deterministic NC and to facilitate an end-to-end performance evaluation in a network consisting of wired and wireless sections, throughout this thesis the traffic characterization is based on the concept of arrival curves, which will be explained in detail in Chapter 2, Section 2.4. Following this assumption, this thesis contributes to the general theory of stochastic min-plus systems, the application of the slope domain modeling and then application of these concepts in wireless networking.

The main challenge involved is the random behaviour of a wireless channel; therefore, it practically can only be modeled as a stochastic server. The physical layer characteristics of the channel are usually known, and can be used to define a stochastic service for the channel. Thus, if a stochastic min-plus theory is available, the channel can be modeled.
Two major stochastic treatments of the min-plus system theory exist in the literature:

1) The \((\sigma(\theta), \rho(\theta))\)-calculus extends the application of NC by introducing a stochastic traffic characterization based on the log-moment generating function using effective bandwidth theory [8]. The main advantage of such a traffic model is that it extends the application of NC to include traditional traffic models such as Poisson. However, we would like to work with an arrival-curve traffic characterization and consider a stochastic server representing a wireless link or network. Even the effective capacity method [16], which is based on the role reversal of effective bandwidth, is suitable for a standalone link only.

2) The statistical NC is another method that provides general direction in studying a system when a stochastic service curve with a meaningful probability of occurrence is given [17][18][19]. The stochastically bounded burstiness [20] method is a subdivision of the statistical NC methods. The practical problem with this method is how to actually define such a service curve. Chapter 6 of this thesis provides an alternative channel model based on this method.

Therefore, a stochastic min-plus theory is needed that

- fits within the NC filtering theory
- can model a wireless link accurately from the channel's statistical characteristics
- is applicable independent of the traffic model used
- has useful applications for solving practical problems in wireless data networking
- and can be used in an end-to-end QoS analysis of a network consisting of both wired and wireless sections, and therefore is compatible with the arrival curve traffic characterization.

In summary, the motivations for this work can be expressed as follow:

- Network Calculus is a successful framework for the performance evaluation of data communication networks. How can this framework be extended and modified in order to be utilized in a network that consists of wireless data networks, in a simple yet general framework? We focus on the arrival-curve traffic characterization.
Chapter 1. Introduction

- The min-plus system theory is similar to traditional system theory in many respects. Is there a similar framework for stochastic input/output characterization?
- Following a similar reasoning, is there a relationship between the statistics of the input and output of a min-plus system when one is a stochastic process, similar to the traditional system theory?
- The Fourier transform is a fundamental tool with many applications in traditional system theory. Does a mathematical transform exist for the min-plus theory with a broad range of applications in data networking that is not constrained by the limitations that the Legendre transform is constrained by?
- In turn, can the aforementioned frameworks and techniques provide an effective approach to performance evaluation of wireless networks, and to a variety of networking problems, such as optimal smoothing, optimal shaping, and network identification?
- In summary, the objectives of the thesis are the development of a required stochastic min-plus system theory and useful techniques for its application in performance evaluation and QoS provisioning in wireless data networks.

The next subsection lists the main contributions of the thesis.

1.2. Main Contributions

This thesis contributes to the performance evaluation and QoS provisioning techniques in wireless networks by utilizing the min-plus system theory approach. It introduces several new methods that enhance and extend the application of the min-plus theory in queuing systems, including slope domain modeling and a new stochastic min-plus system theory. Based on these methods, models for a wireless channel and consequently a wireless network are presented. These models are then used in a number of applications, such as performance modeling and evaluation of service quality, optimal traffic shaping and resource allocation, optimal smoothing, and system identification.

The contributions of this dissertation can be classified under the following three main themes:
* **Slope domain modeling and its applications:** The development of slope domain modeling and its application in data communications are contributed by this thesis [21]. The proposed approach is based on the application of the mathematical concept of slope transform in min-plus system theory. Slope domain modeling supersedes and encompasses models based on the Legendre transform, as it is not limited to convex (or concave) or differentiable functions. Similar to the application of the Fourier transform in the conventional system theory, the slope transform maps functions in the min-plus domain into another domain, called the slope domain. Similar to the Fourier's frequency domain, this mapping not only facilitates performance modeling and evaluation of data networks, but also is sometimes the only method available for solving certain problems. Slope domain modeling establishes an analytical framework with numerous applications. In particular, system identification, priority scheduling, and optimal smoothing are studied in this thesis.

* **A stochastic min-plus system theory approach and wireless channel modeling:** A set of stochastic system theorems is presented that characterizes the input/output relationship of a system with stochastically defined input or service [22][23]. These theorems are then used as the basis for a technique in stochastic min-plus system theory. The method identifies the second-order statistics of the output of a min-plus system using the results of two mathematical theorems. This method is similar in approach and application to the traditional stochastic system theory.

The dissertation considers a wireless channel model as a stochastic rate-latency server. In conjunction with the proposed stochastic system technique, this model is used to characterize a traffic stream passing through a wireless channel. This is also a practical methodology for performance modeling of any stochastic min-plus server. In addition, an alternative channel model based on the implementations of a filter bank of concatenated deterministic NC elements is proposed [24][25]. These two models collectively provide a practical methodology for modeling wireless networks using min-plus theory.

* **Applications in wireless data networking:** This dissertation investigates the application of the proposed techniques in wireless data networking, including

  - performance evaluations and QoS provisioning in wireless data networking [23][26]
Chapter 1. Introduction

- optimal smoothing for the calculations of the playback delay and buffer size at the receiver side [24]
- and optimal shaping and resource allocation for transmission of multimedia over wireless channels [27].

1.3. Organization of the Thesis

The outline of the thesis is as follows. The current chapter presents the motivations and objectives of this work in an itemized format and is an introduction to the methodologies used. The next chapter provides an overview of min-plus filtering theory and explains the mathematical foundation, theorems and notation used throughout this thesis. The main body of the thesis consists of chapters 3, 4, 5, and 6. Chapter 3 introduces slope domain modeling as a complement to NC theory. Slope domain modeling facilitates analysis of networks in a fashion similar to the Fourier domain in traditional systems. A number of particularly useful applications of the proposed methodology are then discussed. Chapter 4 presents a stochastic min-plus system theory approach in which the input/output relationship of a min-plus system is derived in terms of the second-order statistics. This concept is then used in performance modeling of wireless networks. Chapter 5 proposes optimal traffic shaping and resource allocation techniques in wireless networks, making use of the theorems introduced in Chapter 4. Finally, Chapter 6 presents an alternative methodology that utilizes deterministic NC elements to model a wireless link. This approach is then applied to optimal smoothing application in streaming variable bit rate traffic over a wireless network. Finally, Chapter 7 summarizes the thesis and proposes future work.
Chapter 2. Background

In this chapter, an introduction to the basic concepts and ideas of the min-plus system theory is presented, and the required mathematical background is described. The chapter provides a concise yet complete survey of the history, background and related min-plus techniques used in communications networking. \((\sigma, \rho)-\text{calculus}\) [8] and deterministic NC [9] are subdivisions of this discrete event dynamic systems (DEDS), called min-plus system theory in this thesis. First, we introduce the fundamental concepts, explain the advantages and disadvantages of this approach, and discuss available alternatives. Second, different deterministic and stochastic traffic characterizations under this theory are presented. Then, deterministic NC and various stochastic approaches are discussed. This chapter consists of the following sections:

- A brief history of Network Calculus and min-plus system theory fundamentals
- Min-plus algebra
- Min-plus traffic characterization
- Deterministic Network Calculus
- Statistical and stochastic min-plus system theory approaches
- Min-plus theory for wireless networking
- Mathematical transformations for min-plus system theory

For clarification, we briefly define and enlist critical min-plus terms used in this thesis. Deterministic NC (or simply NC) is a min-plus theory in which both traffic and server are deterministically defined. Stochastic min-plus theory has three branches: 1) Stochastic min-plus
theory based on moment generating functions and effective bandwidth theory, where either the traffic or the server is stochastically defined. The \( ((\sigma(\theta), \rho(\theta))) \)-calculus of [8] is of this category for stochastic traffic characterizations. Also, the effective capacity method is of this type. NC filtering theory is not directly applicable to this method. 2) Statistical methods such as stochastically bounded burstiness. 3) The stochastic min-plus theory presented in this thesis. The NC filtering theory is applicable to this method.

2.1. A Brief History

The development of a modern calculus for the study of data communication systems has been underway for the last few decades. The calculus is based on a special algebra whose mathematics was conclusively explained in [7]. In their book, Baccelli et al. presented the application of such a calculus in the modeling and analysis of DEDS. The specific application of the min-plus theory in data communications networks was first introduced with the pioneering works of Cruz in his seminal papers on the calculation of network delay [28][29], and discussed further in [30]. Kurose then contributed to the computation of the upper bounds of per-session performance measures such as delay and buffer occupancy [31].

The concept then was elaborately developed into a comprehensive and useful theory in two independent books. *Network Calculus* by Le Boudec and Thiran is the main reference for deterministic service guarantees [9]. In his prominent book, Chang [8] elaborated on the concept of the deterministic NC by presenting the \((\sigma, \rho)\)-calculus. He also introduced stochastic \((\sigma(\theta), \rho(\theta))\)-calculus based on moment generating functions. This work has been the prime reference for the stochastic treatment of min-plus systems, using the theory of effective bandwidth. Most subsequent stochastic NC models and applications are based on the effective bandwidth theory.

Many others have contributed to different aspects and applications of the theory. Liebeherr *et al.* presented a method for providing end-to-end statistical service guarantees [17]. In this significant work, a statistical alternative is presented as a framework for the statistical analysis of non-deterministic systems. Liu *et al.* [20] and Starobiniski *et al.* [32] also presented a general
stochastic NC, based on stochastically bounded burstiness. Chang and Cruz presented the time-vary filtering theory of min-plus systems [33]. The application of the Legendre transform in NC was first mentioned by Cruz in his pioneering papers on NC [28][29], and then was discussed as an alternative to NC domain modeling by Hisakado et al. [34]. However, the Legendre transform has a number of limitations. Slope transform is a more general transform that does not have those limitations and includes the Legendre transform as a special case [35]. Concurrent to the development of our slope domain model, other works on such alternative domains have been published by Pandit et al. [36], and by Fidler and Recker [37], based on the Legendre transform.

The application of NC in wireless networking has been investigated quite recently. MRSVP (mobile resource reservation protocol) is a resource reservation protocol for an integrated services network with mobile hosts [38]. This protocol is an extension of the RSVP (resource reservation protocol) proposal for wired networks. Hamdi and Lee presented a packet-drop mechanism for deterministic packet delay and packet losses in wireless networks [39]. Wu and Negi presented a model, called effective capacity, based on the reverse role of the effective bandwidth concept [16]. Although the method makes brilliant use of the moment generating function concept in stochastic traffic characterization by swapping the roles of input and service to model a wireless channel, it has a number of shortcomings. First, it is based on a constant traffic rate, and it is not clear how it can be applied to many other practical traffic models. In practice, traffic flow obeys arrival curves of different forms. Second, it only models a single wireless link (single hop) and cannot be used in a scalable end-to-end NC analysis. Finally, the effective bandwidth concept, which is based on the moment generating functions, is a complex concept to use in both practice and theory.

2.2. Min-plus System Theory Fundamentals

The min-plus system theory is based on DEDS modeling of the time behaviour of a class of systems where a number of customers compete to obtain the service of a number of available resources [7]. Computer communications networks and queuing networks belong to such classes of systems. The time-behavior dynamics of these systems can be described by the concepts of synchronization and concurrency [7]. Synchronization is a basic operation of a DEDS, and
describes how its elementary parts interact. Synchronization requires many users to be present at the same time as a number of resources. Concurrency occurs when the users have a choice and must select one from a number of resources (such as in routing algorithms).

Min-plus (or its dual max-plus) algebra has proven to be a good candidate for modeling DEDS when they do not involve concurrency. Such systems are linear, when modeled by the min-plus theory [7]. The definition of min-plus linearity will be explained later in this section.

Min-plus system theory has a number of advantages. First, it provides a set of theorems and rules that establish a sophisticated filtering theory for the analysis of complex DEDS. This systematic approach is similar in many respects to the familiar traditional system theory. Second, its unique traffic characterization is simple yet efficient, practical, easy to define and has both deterministic and stochastic models. This is a fundamental aspect differentiating it from traditional queuing theory. Third, it has been proven to be the best candidate for the analysis of a number of frameworks, including Internet IntServ, DiffServ and ATM. Also, it can be applied to systematically estimate deterministic and stochastic bounds for QoS metrics of delay, buffer requirements and throughput.

Figure 2.1 illustrates the input/output representation of a min-plus system. In min-plus system with an input $x(t)$, a service curve (system function) $h(t)$, the output is defined by the system operator, $L(.)$, by

$$y(t) = L^{h(t)}(x(t)),$$

(2.1)

When it is clear that $h(t)$ is the service curve, for the simplicity of notation, the system operator may be noted only by $L(.)$. If $h(t)$ is a lower service curve, then $y(t) \geq L(x(t))$, and if $h(t)$ is an upper service curve, then $y(t) \leq L(x(t))$. All functions, (i.e. $x(t)$, $h(t)$ and $y(t)$) are non-decreasing.
A min-plus system with a system function $L(.)$ is linear when

$$L[\inf(a + x_1(t), b + x_2(t))] = \inf(a + L[x_1(t)], b + L[x_2(t)]).$$

(2.2)

In a broader sense, this concept is also an instance of the infimum-of-sums superposition principle, which is addressed later in Section 2.4 by (2.5). A min-plus system is shift-invariant if it commutes with a shift operation on the input, i.e., a shift in time of the input causes the same shift of the output [7].

In a communication network, the box (system) represents a network element (link, node, network); $x(t)$ (input) represents the cumulative incoming traffic (up to $t$) in units of bits (or bytes); and $y(t)$ (output) represents the cumulative outgoing traffic (or received by an element at the other end) in units of bits. The service curve $h(t)$ defines the service provided by the network. The exact definitions and models defining $x(t)$ and $h(t)$ are presented in Sections 2.4 and 2.5.

Traditionally, the behavior of communications networks has been modeled and analyzed by the so-called queuing theory, in which the system is modeled by an $A/B/m/N$ queue, a notation popularized by D. G. Kendall [40]. While the traffic flow and the server can be characterized by arbitrary stochastic processes in this method, most of the studies available in the literature are devoted to the study of the $M/M/1$ queues, where the inter-arrival and service times are exponentially distributed. Although these models are the foundation of modern networking theory and have been crucial in giving us a preliminary understanding of how networks function, they become very complicated when we apply them to the study of real networks with sophisticated scheduling techniques that guarantee service quality.

NC, on the other hand, is a new approach to studying computer communication networks and their elements, such as shapers, multiplexers, and queues. It has been proven to simplify the study of numerous networking problems, especially those related to QoS provisioning [10][11]. NC is usually compared with system theory, which is used in the systematic study of circuits and systems in electrical engineering. Two main features characterize NC theory. First, the traffic flow is defined by a non-decreasing (or a wide-sense increasing, as it is called in [9]) function of time called the “arrival envelope” or “arrival curve.” Second, the behavior of the system serving the
traffic is defined by "service curves," so that the output is calculated by a special convolution of
the input and this service curve. The theory is mathematically founded on "min-plus" (or its dual
max-plus) algebra, and has been developed into a set of theorems of filtering theory.

NC has found fundamental applications in the analysis of QoS-enabled packet networks
such as ATM networks, and the Internet, employing the IntServ and DiffServ frameworks [9]. NC
is an effective approach to developing optimal smoothing algorithms in video transmission over
(lossless) wired networks [41]. The theory is based on the pioneering work of Cruz [28][29].

There are two independent yet very closely related main interpretations of this theory,
described in books by Le Boudec and Thiran [9] and by Chang [8]. The latter book introduces a
stochastic traffic characterization as an extension to the theory. Although deterministic NC is just
defined by deterministic arrival and service curves, stochastic NC has been characterized in a
number of ways. In this thesis when not specifically mentioned or not obvious from the context,
NC should be considered as analogous to deterministic NC.

The NC theory provides numerous benefits. It is ideal for the study of DEDS characterized
by synchronization. Also, it can be readily used to study non-linear systems, as most communica­
tion systems are. Owing to its concatenation property, it is modular in its modeling of cascades of
network elements with known service curves. The traffic characterization of NC is its major
difference from traditional queuing theory, and also its great advantage where the only informa­
tion we need is the traffic envelope. NC's powerful filtering theory simplifies more accurate
studies of many problems that otherwise are impossible or hard to solve. System identification
methods in NC can be used to identify an unknown element.

Like any other tool or theory, NC has its limitations. First, in order to derive tight bounds,
the functions describing the traffic and the service need to be "good" functions. In min-plus
algebra, this translates to being a subadditive function. The implications of this condition are
addressed later in Section 2.5, where the subadditive concept is explained. In addition, functions
are mostly desired to be convex or concave, particularly when the analysis relies on the Legendre
transform tools. Another limitation, similar to the G/G/1 theory, is that the queue needs to be zero
at the reference time zero; a non-zero queue requires extra care, which creates difficulty in
working with time-varying systems. NC has limited tools for describing packet loss in the network; there is only the clipper element, which is used in NC.

2.3. Min-plus Algebra

The mathematics of NC is min-plus algebra, which is based on the dioid of \((\mathbb{R} \cup \{\infty\}, \Lambda, +)\) in which \(\Lambda\) is the infimum (or minimum) operator and "+" is the well-known algebraic addition [7]. Infimum is the greatest lower bound of a set, but it does not need to be in it, e.g., \(a\) in the set \((a, b]\). If this value is in the set, it is called the minimum, such as \(a\) in the set \([a, b]\). A dioid is an algebraic structure in which a set is endowed with two operators (here the set \(\mathbb{R}\) and operators \(\Lambda\) and \(+\)) that satisfies 10 axioms of: closure of \(\Lambda\) and \(+\), associativity of \(\Lambda\) and \(+\), existence of a zero element for \(\Lambda\) and a neutral element for \(+\), idempotency of \(\Lambda\), commutativity of \(\Lambda\), distributivity of \(+\) with respect to \(\Lambda\) and finally "the zero element of \(\Lambda\) is absorbing for \(+\)" [9]. The set of axioms used for definition of a dioid is similar to a ring, except that the idempotency of the addition is replaced for the existence of the negative element for the addition. As a comparison, the normal algebra we are used to is based on the algebraic field of \((\mathbb{R} \cup \{\infty\}, +, \times)\). In the min-plus algebra, computation of minimum replaces addition, and addition replaces multiplication in \((\mathbb{R} \cup \{\infty\}, +, \times)\).

In the literature, the operators are sometimes denoted by \(\oplus\) for the first operator and \(\otimes\) for the second operator, as in \((\mathbb{R} \cup \{\infty\}, \oplus, \otimes)\). Therefore, in normal algebra \(\oplus\) stands for normal addition and in min-plus algebra for minimum, while in normal algebra \(\otimes\) stands for normal multiplication and in min-plus algebra for normal addition. However, since \(\otimes\) is also a common notation for min-plus convolution and is used for that purpose in this thesis, we use the obvious notations of \(+\) for algebraic addition, \(\text{min}\) or \(\Lambda\) for minimum, \(\times\) or \(\cdot\) as algebraic multiplication and \(\otimes\) for convolution, unless explicitly clarified, as in the following example. As an example, we calculate the matrix operation of \(A^2\) for the matrix \(A = \begin{bmatrix} 3 & 7 \\ 2 & 4 \end{bmatrix}\) (for min-plus operation):
Chapter 2. Background

\[ A^2 = A \otimes A = \begin{bmatrix} 3 & 7 \\ 2 & 4 \end{bmatrix} \otimes \begin{bmatrix} 3 & 7 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} \min(3 + 3, 7 + 2) \min(3 + 7, 7 + 4) \\ \min(2 + 3, 4 + 2) \min(2 + 7, 4 + 4) \end{bmatrix} = \begin{bmatrix} 6 & 10 \\ 5 & 8 \end{bmatrix}, \]

which is similar to the multiplication of matrices in normal algebra, by replacing addition with minimum operation and multiplication with addition.

As another example, a difference equation describing a traditional system is

\[ x(t + 1) = Ax(t) \text{ or } x_i(t + 1) = \sum_{j=1}^{n} A_{ij}x_j(t). \]  

(2.3)

In the min-plus system theory, a similar difference equation describing a min-plus system would be

\[ x(t + 1) = A + x(t) \text{ or } x_i(t + 1) = \min_j(A_{ij} + x_j(t)). \]  

(2.4)

The dual to the min-plus algebra is the max-plus algebra, where the infimum operator is replaced with supremum operator, resulting in almost similar theorems and results. In general, these two algebras describe two sets of systems. The systems that obey the infimum-of-sums superposition principle are called erosion translation-invariant (ETI) systems in morphological signal processing [35] and are used with the min-plus algebra. The infimum-of-sums superposition principle states that the system operator, \( L(.) \), obeys

\[ L[\min(c_i + x_i(t))] = \min(c_i + L[x_i(t)]). \]  

(2.5)

Some other systems obey the supremum-of-sums superposition property (so-called dilation translation-invariant (DTI) system in morphological signal processing) and are used with the max-plus algebra [35]. Other than a few references in the literature, such as [34], which primarily considers max-plus theory for analyzing queuing systems, the role of the max-plus is generally complementary, only to assist in solving certain problems.
Convolution and Deconvolution Operators: The $\otimes$ operator characterizes the min-plus convolution defined by

$$(x \otimes h)(t) = \inf_{0 \leq s \leq t} \{x(t-s) + h(s)\}, \quad (2.6)$$

in which $\inf$ is infimum. An important property of the convolution is commutativity, i.e.,

$$x \otimes h = h \otimes x.$$ 

Another important operation is the deconvolution defined by

$$(x \oslash h)(t) = \inf_{0 \leq s \leq t} \{x(s + t) - h(s)\}, \quad (2.7)$$

where $\sup$ denotes the supremum of a set.

In max-plus algebra, the convolution is defined as

$$(x \ast h)(t) = \sup_{0 \leq s \leq t} \{x(t-s) + h(s)\}, \quad (2.8)$$

and the deconvolution is defined as

$$(x \oslash h)(t) = \sup_{0 \leq s \leq t} \{x(s) - h(t-s)\}. \quad (2.9)$$

Min-plus System Operator and Convolution: In a linear shift-invariant min-plus system, the system operator can be defined by the min-plus convolution. If $x(t)$ is a non-decreasing function representing the input, and the system function (service curve) is $h(t)$, then

$$L^{h(t)}(x(t)) = x(t) \otimes h(t) = \inf_{0 \leq s \leq t} (x(t-s) + h(s)). \quad (2.10)$$

For a max-plus system, the system operator $\overset{\text{max}}{L}^{h(t)}$ is defined by the max-plus convolution as follows:

$$\overset{\text{max}}{L}^{h(t)}(x(t)) = x(t) \ast h(t) = \sup_{0 \leq s \leq t} (x(t-s) + h(s)).$$
Convex and Concave Functions: A function $f$ is convex if

$$f(ua + (1-u)b) \leq uf(a) + (1-u)f(b),$$

(2.11)

for every $a, b \in \mathbb{R}$ and all $u \in \mathbb{R}$ with $0 < u < 1$. In other words, if $f(t)$ has a second derivative in $[a, b]$, then a necessary and sufficient condition for it to be convex in that interval is that the second derivative $f''(t) \geq 0$, for all $t$ in $[a, b]$ [42]. A function $g$ is concave if $-g$ is convex. Also, $g(t)$ is concave when

$$g(t) \geq \frac{pg(t-q) + qg(t+p)}{p+q}, \forall p, q \geq 0 \text{ and } \forall t,$$

(2.12)

and $g(t) < \infty$ for all $t$.

2.4. Traffic Characterizations

Traditionally, traffic models are modeled by stochastic processes and distributions. The most notable of these are based on renewal processes, Markov chains, recursive models, transform-expand-sample models, and self-similar models [43][44][45]. Poisson models are the primary traffic models. The fundamental theory of traffic modeling is based on the Poisson renewal model, in which interarrival times of the traffic units are distributed according to an exponential distribution. The Poisson model is supported by a vast amount of literature and is used extensively. Modern measurement and understanding of traffic flows, however, suggest self-similar models for traffic [46].

Although the Poisson model is an integral part of our current understanding, design and analysis of queuing systems, there is cautionary advice about the use (or overuse) of Poisson-based traffic models. It has been well discussed and proven that this model should be avoided when dealing with many new traffic types, such as packet video or multimedia ([40], pp. 21-26). In deterministic NC, the cumulative traffic needs only to be constrained by a curve (a function). In stochastic traffic characterization, the moment generating function is constrained, which could provide more statistical knowledge than most widely used traditional traffic models [8].
Deterministic Traffic Characterizations: This model is based on the concept of arrival curves which define an envelope function for the traffic. A traffic flow \( x(t) \) is said to have an arrival curve \( \zeta(t) \), or to be constrained by a given non-decreasing function \( \zeta(t) \), if and only if for all \( t \geq 0 \) and \( 0 \leq s \leq t \), the following holds:

\[
x(t) - x(s) \leq \zeta(t - s). \tag{2.13}
\]

\( x \) is also called \( \zeta \)-smooth.

This is the main traffic model used throughout this thesis. A more elaborate definition of it and its usage will be presented in the next subsection.

Example 2.1: A flow with an arrival curve \( \gamma_{a,b}(t) = b + at, \ t > 0 \) and 0 otherwise, can be realized by a leaky-bucket shaper with leak rate of \( a \) and bucket size of \( b \). This arrival curve, also called the affine function, is the most widely used arrival curve in NC.

Example 2.2: For ATM VBR traffic and for the IntServ traffic with T-SPEC (traffic specification) \((M, P, r, b)\), where \( M \) is the maximum packet size; \( P \) is the peak rate; \( r \) is the mean rate; and \( b \) is the burst tolerance, the arrival curve is \( \zeta(t) = \min(M + Pt, rt + b) \).

Example 2.3: A deterministic bit rate flow, which is also improperly called "constant bit rate" (CBR), has an arrival curve defined by \( \zeta(t) = rt \).

Stochastic Traffic Characterizations: The first stochastic characterization of traffic flows was introduced by Kurose [31], who defined a flow by a family of discrete random variables \((R^i_{t_j}, t_j)\). Each of these random variables, \( R^i_{t_j} \), defines a bound over a time length \( t_j \). Therefore, the flow is defined by

\[
S^i \sim \left\{ (R^i_{t_1}, t_1), (R^i_{t_2}, t_2), (R^i_{t_3}, t_3), \ldots \right\}, \text{ or in compact form by } S^i \sim \{(R^i_{t_k}, k) \mid k \geq 1 \}.
\]
A more elaborate stochastic traffic characterization is defined in \((\sigma(\theta), \rho(\theta))\)-calculus using the \(\theta\)-envelope [8]. This traffic characterization is based on the concept of the moment generating functions. A traffic flow \(x\) is said to have a \(\theta\)-envelope rate function, \(\rho(\theta)\), when
\[
\rho(\theta) \geq a^*(\theta),
\] (2.14)
where \(a^*(\theta)\) is called the \(\theta\)-minimum envelope rate (\(\theta\)-MER) of \(x\), and is defined by
\[
a^*(\theta) = \limsup_{t \to \infty} \sup_{s \geq 0} \log(E[\exp(\theta(x(t+s) - x(s)))]).
\] (2.15)

This model is a stochastic extension of the arrival curves used in the deterministic NC or \((\sigma, \rho)\)-calculus. It is shown in [8] that this traffic characterization includes more information than the usual traditional traffic models, and can be used as an alternative for most traditional traffic models.

### 2.5. Deterministic Network Calculus

NC is a comprehensive collection of theorems, operators, and elements that form a filtering theory that allows us to analyze a data communication network and calculate the deterministic end-to-end QoS bounds, provided that the incoming traffic complies with certain constraints. It is mathematically founded on min-plus algebra [7]. In NC, the behaviour of an element, a node or a network (the system) is characterized by a function called the service curve [9], which is very similar to the system function defined in conventional system theory. A traffic flow is represented by a non-decreasing (or wide-sense increasing) function of time, \(x(t)\), which is the cumulative traffic until time \(t\). Except for some special situations, it is assumed that \(x(t) = 0\) for \(t \leq 0\). The traffic characterization is based on the concept of arrival curves (explained in the previous section) and so the traffic flow is constrained by the arrival curve. In the following, these terms and some fundamental topics in NC are discussed in more detail. Ref. [8] and [9] comprehensively study this new approach to networking.
Arrival Curve: In NC, the traffic characterization is based on the definition of arrival curves. This concept was explained in the previous section.

In a class of curves, $\zeta_i(t)$, that satisfy (2.13), one “good” arrival curve, $\zeta(t)$, exists so that $\zeta \leq \zeta_i$, for all $i$. Such a function is subadditive ([9], pp. 15). Function $f(t)$ is subadditive, if and only if $f \otimes f = f$. By using the subadditive closure, the subadditive equivalent of any curve can be found. The subadditive closure of any function $g(t)$, $\bar{g}(t)$, is defined by $\bar{g}(0) = 0$ and $\bar{g}(t) = \min\{g(t), \min_{0<s<t}[\bar{g}(t)+\bar{g}(t-s)]\}$ for $t > 0$ [8]. By using subadditive functions, it is ensured that the curve is an accurate model of the traffic, and results in correct calculations of QoS bounds.

Service Curve: In this model, the per hop behaviour or the level of traffic treatment of a node is modeled by a curve $h(t)$, called the service curve. This function defines the amount of service a traffic flow receives during any time interval. For the input flow $x(t)$ and output $y(t)$, the lower service curve guarantees that for all $t \geq 0$ and $0 \leq s \leq t$,

$$y(t) \geq x(s) + h(t-s).$$

(2.16)

In order to define the service curve of a server as a min-plus system function, the convolution operator is used. Similar to the classical system theory, this service curve is all that is needed to describe the system and to calculate the output of the system, given the input. A network element with a lower service curve $h(t)$ for an input traffic stream $x(t)$ satisfies

$$y(t) \geq x(t) \otimes h(t) = \inf_{s \geq t}\{x(s) + h(t-s)\},$$

(2.17)

where $y(t)$ is the output, the $\otimes$ operator represents the min-plus convolution, and $\inf$ denotes infimum. If the equality holds in (2.17), then $h(t)$ is called the service curve (or exact service curve, for clarity), i.e.

$$y(t) = x(t) \otimes h(t).$$

(2.18)

If $y(t) \leq x(t) \otimes h(t)$, then $h(t)$ is an upper service curve.
This concept can be exemplified by the GPS scheduler used to explain the physical meaning of a service curve. A GPS scheduler assigns a rate $r$ to a traffic aggregate, or put differently, for any time interval $([t_0, t], \forall t, t_0 \geq 0)$: $y(t) - y(t_0) \geq r \times (t - t_0)$, where $t_0$ is the start of the last busy period. In NC terms, the scheduler has a service curve of $\gamma_{r, 0}(t) = rt$, resulting in $y(t) \geq x(t_0) + \gamma_{r, 0}(t - t_0)$, which in turn leads to $y \geq x \otimes \gamma_{r, 0}$, as in (2.17).

In the following, two widely used service curves and the network elements that realize them are introduced.

**Leaky-bucket Shaper:** A leaky-bucket shaper is a network element with a service curve of $\gamma_{a, b}(t) = at + b$, $t > 0$, and 0 otherwise (an affine function). The shaper forces the input traffic to an output with a rate of $a$ and a burst size of $b$, and is a good example for a server with an exact service curve. The output of this shaper has $\gamma_{a, b}(t)$ as its arrival curve.

**Rate-latency Server:** The most widely used service curve is the rate-latency curve. A rate-latency server is a node that guarantees a (minimum) rate $R$ and a (maximum) delay $T$, when serving a particular traffic flow. The rate-latency service curve is $\beta_{R, T}(t) = R \times (t - T)^+$. A practical example of such servers is the WFQ (weighted fair queuing) scheduler.

**Concatenation Property:** In order to analyze a network, it is normally necessary to apply the convolution on service curves of different parts of the network and the input, either individually one after the other or by applying the concatenation property. The service curve of a chain of network elements is equal to the convolution of their service curves. This is known as the concatenation property, which allows modular modeling and analysis of networks.

**Deterministic QoS Bounds:** For an arrival process $x(t)$ with a given arrival curve $\zeta(t)$, that passes through a system with the lower or exact service curve $h(t)$, the three bounds on the backlog, delay and the output flow $y(t)$ are given by these three theorems. When studying a communication network, the first theorem is used to calculate the amount of traffic that is inside the network at any given time, from which the maximum required buffer size can be calculated.
Similarly, the second theorem presents a bound on the maximum delay that the traffic flow experiences. Finally, the third theorem defines the arrival curve that models the output, which is useful when the output flow is itself an input to the next network in concatenation.

**Theorem 2.1 - Bound on the Backlog:** The backlog \( x(t) - y(t) \) for any given time \( t \) is

\[
x(t) - y(t) \leq v(\zeta, h) = \sup_{s \geq 0}(\zeta(s) - h(s))
\]

where \( v(\zeta, h) \) graphically is the maximum vertical distance between the arrival and the service curves.

**Theorem 2.2 - Bound on the Delay:** The virtual delay \( d(t) \) at any give time \( t \) satisfies

\[
d(t) \leq k(\zeta, h) = \sup_{s \geq 0}(\inf\{T: T > 0 \text{ and } \zeta(s) \leq h(s + T)\})
\]

where \( k(\zeta, h) \) graphically is the maximum horizontal distance between the arrival and the service curves.

**Theorem 2.3 - Output Flow:** The output flow is constrained by the curve

\[
\zeta^* = \sup_{u \geq 0}\{\zeta(t + u) - h(u)\} = \zeta \vartriangle h
\]

i.e., the output can be assumed to be the input traffic to the next node with an arrival curve of \( \zeta^*(t) \). The proofs of these theorems can be found in [9].

### 2.6. Stochastic Min-plus Approaches

Deterministic NC gives the worst-case bounds for numerous QoS metrics such as delay, service rate, and packet loss, provided the assumed conditions on the traffic and service properties are met. For some applications, though, these worst-case bounds are either loose or cannot be defined, for example, when the traffic is not deterministically constrained or when the service available can only be guaranteed within certain probability. These limitations provide the motivation to consider stochastic extension of network calculus.
Chapter 2. Background

There are two main perspectives regarding stochastic NC in the literature. Analogous to the traditional queuing theory, the first perspective assumes that traffic is stochastically constrained [8]. Instead of the traffic itself, a function of the moment generating function of the cumulative traffic is assumed to be constrained by \((\rho(t), \sigma(t))\). The second method is based on the statistical representation of the service curve [17]. The statistically bounded burstiness method is a subdivision of the latter method [20].

As a main contribution of this thesis, a stochastic system theory approach is proposed for modeling min-plus (or its dual max-plus) systems with stochastic input or stochastic service curves. This method will be elaborated upon in Chapter 4.

\((\sigma(\theta), \rho(\theta))-\text{Calculus:}\) In this calculus, traffic is stochastically characterized. A cumulative traffic \(x(t)\) is called a \((\sigma(\theta), \rho(\theta))-\) upper constrained, for some \(\theta \geq 0\), if

\[
\frac{1}{\theta} \log E[e^{\theta(x(t)-x(s))}] \geq \rho(0)(t-s) + \sigma(0), \forall s, 0 \leq s \leq t.
\] (2.22)

Based on this traffic characterization, the theory of effective bandwidth is introduced. This calculus is the primary stochastic method in scenarios where the traffic characterization can only be done stochastically. The service curve, however, is assumed to be deterministic. The role of the traffic and service can be swapped to represent a stochastic server [16]. However, this method is very complex and does not end up with an easy to use and general model for a practical stochastic server.

\textbf{Statistical Service Guarantees Method:} The theory of NC is usually extended to a statistical framework by considering that the service represented by (2.17) is guaranteed with a predefined probability. In this method, the arrival process \(x(t)\) is still constrained by a deterministic curve \(\zeta(t)\), but the system can only stochastically guarantee the service. The lower service curve \(G^e(t)\) is defined in correspondence with a probability \(1 - \varepsilon\) (as shown in Figure 2.2), so that

\[
P\{y(t) \geq (x \otimes G^e)(t)\} \geq 1 - \varepsilon
\] (2.23)
Chapter 2. Background

For an exact service curve $G^e(t)$, in (2.23) inside the probability equality holds. This methodology is useful and compatible with the deterministic NC. However, its application depends on the accurate definition of the service curve with a meaningful probability. Chapter 6 of this thesis is based on this method.

Unlike the deterministic NC, which characterizes the worst-case bounds, in statistical NC the bounds are defined based on the probability $1 - \varepsilon$.

This $G^e(t)$ is sufficient to derive bounds on the backlog, delay and output in the system with a probability of $1 - \varepsilon$, using the following three theorems [17]. As before, it is assumed that the arrival process $x(t)$ has an arrival curve of $\zeta(t)$ and that the output flow is $y(t)$. The following theorems are the statistical equivalents of Theorems 2.1 to 2.3, respectively; therefore, their physical interpretations are similar to those explained earlier in Section 2.5.

**Theorem 2.4 (Backlog Bound):** The probabilistic bound for the backlog, $B(t) = x(t) - y(t)$, for $t \geq 0$ is

$$Pr(B(t) \leq \zeta \cap G^e(0)) \geq 1 - \varepsilon.$$  \hfill (2.24)

**Theorem 2.5 (Delay Bound):** The maximum delay is

$$Pr(d(t) \leq d_{max}) \geq 1 - \varepsilon,$$  \hfill (2.25)

where

$$d_{max} = \inf\{d \geq 0 | \forall t \geq 0; \zeta(t - d) \leq G^e(t)\}.$$  \hfill (2.26)
**Theorem 2.6 (Output Envelope):** The output process $y(t)$ is stochastically bounded with an envelope (arrival curve), $\zeta^*(t)$, i.e.,

$$Pr(\zeta^*(t) \leq \zeta(t) \cap G^c(t)) \geq 1 - \varepsilon.$$  

(2.27)

The proofs of the above theorems can be found in [17].

### 2.7. Mathematical Transformations for Min-plus Systems

Similar to the Fourier transform, a mathematical transformation with similar properties can be very useful to the min-plus system theory. The Fourier transform is an important tool in system theory. The Fourier transform of the convolution of two functions is equal to the multiplication of the Fourier transform of each function, i.e., $F\{x(t) \ast y(t)\} = F\{x(t)\} \cdot F\{y(t)\}$. The conversion of the convolution into multiplication usually facilitates the analysis.

Chapter 3 considers such a transform, called slope transform and introduces its application as an alternative min-plus domain in the analysis and QoS provisioning of data communication networks. The slope transform is a general transformation that can be replaced by the Legendre transform when dealing with convex or concave functions. Since in the min-plus algebra, the computation of minimum replaces addition and addition replaces multiplication in a real number field, i.e., the min-plus algebra is based on the Dioid $(\mathbb{R} \cup \{\infty\}, \Lambda, +)$ where $\Lambda$ is the infimum and $+$ is algebraic addition, here the convolutions are converted into algebraic additions.
Chapter 3. Slope Domain Modeling and Analysis of Data Communication Networks

Min-plus filtering theory (or its dual, max-plus), commonly called Network Calculus (NC), is a sophisticated modern framework for modeling and analysis of data communication networks. In this chapter, we complement NC theory by proposing the application of slope domain analysis in NC using the slope transform, an analytical framework that is analogous to frequency domain analysis based on the Fourier transform in traditional system theory. This work was inspired by the introduction of the application of the Legendre transform for this purpose in [34], and by the slope transform-based morphological signal processing in [35]. We describe the principles and properties of slope domain analysis, and how it can be effectively applied in the analysis of computer networks. We illustrate the application of the proposed method in system identification and priority scheduling, to demonstrate that slope domain analysis is a much more convenient and efficient, and in some cases, the only, viable approach to solving some NC problems. We conclude the chapter by presenting the Legendre transform and Fenchel conjugate as close relatives of the slope transform that might facilitate fast calculations.

3.1. Introduction

In this chapter, we propose the application of the slope transform [35] in modeling and performance evaluations of networks through NC theory. Similar to the application of the Fourier transform in conventional system theory, the slope transform maps functions in the min-plus
domain into another domain, called the slope domain. This transformation translates convolution operations in the min-plus domain to algebraic additions in the slope domain. The Legendre and Fenchel transforms are special cases of the slope transform. The objective of this chapter is to establish, in the same manner as frequency domain analysis in traditional system theory, an analytical framework that employs the slope domain as an effective alternative, with numerous applications to facilitate modeling and analysis of queuing systems using NC.

The slope transform indicates the slope contents of a function. In NC, the traffic and service are characterized by non-decreasing functions. Since the instantaneous slope in these functions represents the traffic or service rates, the slope domain is a logical representation of the rate contents of those function, similar to the fact that the frequency content of a signal is represented in the frequency domain. In the same manner as the Fourier transform, the slope transform simplifies or even makes possible complicated NC calculations and analysis, because convolution operators are transformed into normal additions in the slope domain.

The application of the Legendre transform in data networking was originally introduced by Hisakado et al. [34]. However, the Legendre transform has a number of limitations. The transform is only defined for convex or concave functions, and is usually applicable when the function is differentiable and has an invertible derivative. The application of the slope transform instead of the Legendre transform eliminates these limitations [35]. In fact, the Legendre transform is a special case of the slope transform. In addition, slope transform is applicable to both min-plus and max-plus systems. Fenchel conjugate is a close relative of the slope transform that is discussed is Section 3.10. Concurrent with the development of our slope domain model, there are some works on the development of such alternative domains by Pandit et al. [36], and by Fidler and Recker [37], based on the application of these transforms for convex functions.

The slope transform is a general mathematical transform that was originally introduced in morphological signal [35] and image processing [47][48], and has been proven to have a similar set of properties and characteristics as the Fourier transform. Two types of slope transforms, upper and lower slope transforms, target the max-plus and min-plus algebras, respectively, and are closely related. Dealing with data communication systems, we mainly work with systems that obey the infimum-of-sums superposition, and thus the min-plus algebra; consequently the lower
Chapter 3. Slope Domain Modeling and Analysis of Data Communication Networks

The slope transform is primarily used. However, we address both lower and upper slope transforms here for the following reasons. First, they are very closely related and both are involved in some properties, e.g., P9 in Table 3.1. Second, max-plus algebra has been used in some previous work to model queuing systems as well, such as in [34]. Finally, both are used in describing some of the slope transform properties such as the reverse transformations, as well as facilitating notations. The Legendre and Fenchel transforms are special cases of the slope transform. Their original forms, however, have some limitations, as primarily they are defined for and are applied to convex or concave functions [42]. Even the extended definition of the Legendre transform fails to inclusively serve in all NC applications, since the transformation result will be set-valued, i.e., the result will be a set instead of a single value [34].

In the next section, the slope transform, its properties and some important theorems are addressed. The application of the slope transform in modeling queuing systems is presented in Section 3.3. The slope domain modeling approach is then applied in various applications, such as priority scheduling (Section 3.4), network identification (Section 3.5), and optimal smoothing of multimedia streaming (Section 3.6). The concept of load spectrum, traffic aggregation, and slope selective filters are other useful applications that are presented in the following sections. In Section 3.9, the slope transform is compared with the Fourier. The benefits and limitations of two special cases of the slope transform, the Legendre transform and Fenchel conjugate, and some practical considerations for fast computations, are discussed in Sections 3.10 and 3.11. Finally, a summary in presented in Section 3.12.

3.2. The Slope Transform

The Fourier transform is an important tool in conventional system theory. The Fourier transform of the convolution of two functions is equal to the product of the Fourier transforms of functions, i.e., \( F\{x(t) \ast y(t)\} = F\{x(t)\} \cdot F\{y(t)\} \). The conversion of the convolution into multiplication usually facilitates the analysis. The slope transform serves as a tool analogous to the Fourier transform in min-plus algebra. Since in min-plus algebra the addition and computation of minimum replace multiplication and addition, respectively, in the field of the real numbers, correspondingly the convolutions are converted into algebraic additions in the slope domain. The
The slope transform can similarly be applied to max-plus algebra, which is dual to min-plus algebra, with similar concepts and properties. The max-plus properties can be derived if minimum is replaced by maximum and infimum by supremum [9].

The Slope Transform: The slope transform indicates the slope contents of a min-plus function. For any function \( x(t): \mathbb{R} \to \mathbb{R} \cup \{\pm \infty\} \), its lower slope transform is defined by \( X_L(\alpha): \mathbb{R} \to \mathbb{R} \cup \{\pm \infty\} \) as follows [35]:

\[
X_L(\alpha) = S_L\{x(t)\}(\alpha) = \inf_{t \in \mathbb{R}} (x(t) - \alpha t).
\] (3.1)

Similarly, the upper slope transform is defined by

\[
X_U(\alpha) = S_U\{x(t)\}(\alpha) = \sup_{t \in \mathbb{R}} (x(t) - \alpha t).
\] (3.2)

From these definitions, it can easily be shown that the upper and lower slope transforms are related by the following expression:

\[
S_L\{x(t)\}(\alpha) = -S_U\{-x(t)\}(-\alpha) = -S_U\{-x(-t)\}(\alpha).
\] (3.3)

Proof: \( S_L\{x(t)\}(\alpha) = \inf_{t \in \mathbb{R}} (x(t) - \alpha t) \)

\[
= -\sup_{t \in \mathbb{R}} (-x(t) + \alpha t)
\]

\[
= -\sup_{\tau \in \mathbb{R}} (-x(-\tau) + \alpha(\tau))
\]

\[
= -\sup_{\tau \in \mathbb{R}} (-x(-\tau) - \alpha \tau)
\]

\[
= -\sup_{\tau \in \mathbb{R}} (-x(-\tau) - \alpha \tau) \to -S_U\{-x(-\tau)\}(\alpha).
\]

In this chapter, unless otherwise specified, by slope transform we mean the lower slope transform.

The lower slope transform is always concave, and the upper slope transform is always convex. If the function \( x(t) \) is convex and has an invertible derivative, then its Legendre transform is equal to its lower slope transform. Similarly, if the function \( x(t) \) is concave and has an invertible derivative, then its Legendre transform is equal to its upper slope transform. If the function is neither concave nor convex, then its Legendre transform is not single-valued, but its slope transform is defined regardless.
The following two theorems supply the required tools for the application of the slope transform in min-plus (or equivalently in max-plus) algebra and consequently in NC analysis.

**Theorem 3.1.** For min-plus functions $x(t)$ and $y(t)$, we have

$$S\{x \circ y\} = S\{x\} + S\{y\},$$

where $\circ$ is the convolution operator.

**Proof:**

$$S\{x(t) \circ y(t)\} = \inf_t \{\inf_{\tau} \{x(\tau) + y(t-\tau)\} - \alpha t\}$$

$$= \inf_{\tau} \{x(\tau) + \inf_t \{y(t-\tau) - \alpha t\}\}$$

$$= \inf_{\tau} \{x(\tau) + Y(\alpha) - \alpha \tau\}$$

$$= X(\alpha) + Y(\alpha)$$

Figure 3.1 illustrates this theorem regarding the input-output relation of a system with a service curve $h(t)$. Following an analogy to the traditional system theory, $H(\alpha)$ is called the slope response of the system.

**Theorem 3.2.** For any function, $x(t)$, the function

$$\hat{x}(t) = \sup_{\alpha} (X_L(\alpha) + \alpha t),$$

expresses the inverse lower slope transform of $x(t)$, i.e.,

$$x(t) \rightarrow X(\alpha) \rightarrow \hat{x}(t).$$

This relationship can be better described by using both lower and upper slope transform notations, as follows:

$$X(\alpha) = S_L\{x(t)\}(\alpha)$$

$$\hat{x}(t) = S_U\{X(\alpha)\}(-t)$$
Chapter 3. Slope Domain Modeling and Analysis of Data Communication Networks

If \( x(t) \) is convex, then \( \hat{x}(t) = x(t) \); otherwise for all \( t \), \( \hat{x}(t) \leq x(t) \). Therefore, (3.7) reduces to \( S_U \{ S_L \{ x(t) \} \} = x(-t) \) for a convex function, proving a duality between time and slope domains. \( X(\alpha) \) is always concave and \( \hat{x}(t) \) is always convex.

Similarly, the inverse upper slope transform is defined, following its duality to the lower slope transform. We view function \( \hat{x}_U(t) \)

\[
\hat{x}_U(t) = \inf_{(\alpha, t) \in \mathbb{R}} (X_U(\alpha) + \alpha t)
\]

(3.8)
as an inverse upper slope transform, where \( x(t) \leq \hat{x}_U(t) \). \( \hat{x}_U(t) \) is always concave. If \( x(t) \) is concave, then \( x(t) = \hat{x}_U(t) \). If the first transformation is the upper slope transform, then the reverse transformation is the lower slope transform and vice versa.

Table 3.1 summarizes the properties of the lower slope transform. For the proofs, the reader is referred to [35]. Similarly, Table 3.2 summarizes the properties of the upper slope transform. In the tables, \( t_0 \) indicates a constant shift in time, and \( \alpha_0 \) a constant shift in the slope domain.
Table 3.1 Properties of the lower slope transform.

<table>
<thead>
<tr>
<th>Property</th>
<th>Function</th>
<th>Lower slope transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>$x(t)$</td>
<td>$X(\alpha) = \inf {x(t) - \alpha t}$</td>
</tr>
<tr>
<td>P1</td>
<td>$\inf_i {c_i + x_i(t)}$</td>
<td>$\inf_i {c_i + X_i(\alpha)}$</td>
</tr>
<tr>
<td>P2</td>
<td>$x(t-t_0)$</td>
<td>$X(\alpha) - \alpha t_0$</td>
</tr>
<tr>
<td>P3</td>
<td>$x(t) + \alpha_0 t$</td>
<td>$X(\alpha - \alpha_0)$</td>
</tr>
<tr>
<td>P4</td>
<td>$x(rt)$</td>
<td>$X(\alpha/r)$</td>
</tr>
<tr>
<td>P5</td>
<td>$rx(t), r &gt; 0$</td>
<td>$rX(\alpha/r)$</td>
</tr>
<tr>
<td>P6</td>
<td>$x(-t)$</td>
<td>$X(-\alpha)$</td>
</tr>
<tr>
<td>P7</td>
<td>$x(t) \otimes y(t)$</td>
<td>$X(\alpha) + Y(\alpha)$</td>
</tr>
<tr>
<td>P8</td>
<td>$\sup(x(t), y(t))$</td>
<td>$\geq \sup(X(\alpha), Y(\alpha))$</td>
</tr>
<tr>
<td>P9</td>
<td>$x(t) + y(t)$, $y$ is concave</td>
<td>$X_L(\alpha) \otimes Y_U(\alpha)$</td>
</tr>
</tbody>
</table>

Table 3.2 Properties of the upper slope transform

<table>
<thead>
<tr>
<th>Property</th>
<th>Function</th>
<th>Slope transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>$x(t)$</td>
<td>$X(\alpha) = \sup {x(t) - \alpha t}$</td>
</tr>
<tr>
<td>P1</td>
<td>$\sup_i {c_i + x_i(t)}$</td>
<td>$\sup_i {c_i + X_i(\alpha)}$</td>
</tr>
<tr>
<td>P2</td>
<td>$x(t-t_0)$</td>
<td>$X(\alpha) - \alpha t_0$</td>
</tr>
<tr>
<td>P3</td>
<td>$x(t) + \alpha_0 t$</td>
<td>$X(\alpha - \alpha_0)$</td>
</tr>
<tr>
<td>P4</td>
<td>$x(rt)$</td>
<td>$X(\alpha/r)$</td>
</tr>
<tr>
<td>P5</td>
<td>$rx(t), r &gt; 0$</td>
<td>$rX(\alpha/r)$</td>
</tr>
<tr>
<td>P6</td>
<td>$x(-t)$</td>
<td>$X(-\alpha)$</td>
</tr>
<tr>
<td>P7</td>
<td>$x(t) \otimes y(t)$</td>
<td>$X(\alpha) + Y(\alpha)$</td>
</tr>
<tr>
<td>P8</td>
<td>$\inf(x(t), y(t))$</td>
<td>$\leq \inf(X(\alpha), Y(\alpha))$</td>
</tr>
<tr>
<td>P9</td>
<td>$x(t) + y(t)$, $y$ is convex</td>
<td>$X_U(\alpha) \otimes Y_L(\alpha)$</td>
</tr>
</tbody>
</table>
3.3. **Slope Domain Analysis of Networks**

Based on the two theorems described in the previous section, slope domain provides an alternative to traditional NC methods for modeling and analysis of data communication systems. In this section, we first derive the slope transforms of some well-known arrival and service curves, and specialized functions in NC. Thereafter, we illustrate the methodology by presenting its applications in priority scheduling and system identification. The notations used throughout this chapter are chosen to be as close as possible to those in [9] and [35], for ease of reference.

As explained in the previous chapter, computer communications systems obey the infimum-of-sums superposition property. These ETI systems are well suited to be used with the lower slope transform. Some other systems are DTI systems that obey the supremum-of-sums superposition property and are used with the upper slope transform.

Example 3.1: The slope transform of the zero impulse function defined by

\[
\mu_z(t) = \begin{cases} 
0, & t = 0, \\
-\infty, & t \neq 0,
\end{cases}
\]

is \(S_L\{-\mu_z(t-t_0)\} = S_U\{\mu_z(t-t_0)\} = -\alpha t_0\), where \(t_0\) is a constant shift in time.

Example 3.2: The zero step function is defined as

\[
\lambda(t) = \begin{cases} 
0, & t \geq 0, \\
-\infty, & t < 0.
\end{cases}
\]

The lower and upper slope transforms of the zero function with a constant shift in time, \(t_0\), are \(S_L\{-\lambda(t-t_0)\} = -\alpha t_0 + \lambda(-\alpha)\) and \(S_U\{\lambda(t-t_0)\} = -\alpha t_0 - \lambda(\alpha)\), respectively.

By subtracting (adding, in the upper slope transform) a zero function from any function a right-sided function is formed. For a right-sided (or causal) function \(x(t)\), we have \(S_L\{x(t)\} = \inf_{t \geq 0}(x(t) - \alpha t) = \inf_{t}(x(t) - \lambda(t) - \alpha t) = S_L\{x(t) - \lambda(t)\}\), i.e., the lower slope transform of \(x(t)\) is equal to \(S_L\{x(t) - \lambda(t)\}\). Similarly \(S_U\{x(t)\} = S_U\{x(t) + \lambda(t)\}\).
The Delta function that is sometimes used in the NC literature can be defined using the zero function by $\delta_T(t) = -\lambda(-t + T)$.

Example 3.3: The affine function is $\gamma_{r, b}(t) = b + rt$ for $t > 0$ and 0 otherwise, and characterizes the output of a leaky-bucket server with parameters $(b, r)$. Since the function is right-sided, its lower slope transform is $S_L\{\gamma_{r, b}(t) - \lambda(t)\} = \lambda(r - \alpha)$.

Quite interestingly, a leaky-bucket node behaves like a slope lowpass filter. When expressed in the slope domain, it rejects all of the slopes (rates) greater than $r$ and passes the rest. Similarly, $S_U\{\gamma_{r, b}(t) + \lambda(t)\} = \lambda(\alpha - r)$.

For a non-causal affine function, $\hat{\sigma} + \hat{\rho}t$, taking similar steps, we have $S_L\{\hat{\sigma} + \hat{\rho}t\} = \hat{\sigma} + \mu_2(\alpha - \hat{\rho})$. Also, $S_L\{\hat{\sigma} + \hat{\rho}t - \lambda(t)\} = \hat{\sigma} + \lambda(\hat{\rho} - \alpha)$.

Example 3.4: The peak rate function that describes the CBR traffic is $p_R(t) = Rt$, for $t \geq 0$. With similar manipulations used in Example 3.3, we have $S_L\{Rt - \lambda(t)\} = \lambda(R - \alpha)$.

Example 3.5: The rate-latency function, $\beta_{R, \tau}(t) = R(t - T)^+$, is a widely used NC function that particularly characterizes a rate-latency server which was explained in Section 2.5. The slope transform of this function is derived as follows:

$$S\{R(t - T)^+\}(\alpha) = \inf_t(R(t - T)^+ - \alpha t) = \inf(\inf_{t \geq \tau}(R(t - T) - \alpha t), \inf_{t < \tau}(-\alpha t))$$

$$= \begin{cases} \inf(-\infty, -\alpha T) = -\infty & \alpha > R \\ \inf(-\alpha T, -\alpha T) = -\alpha T & 0 \leq \alpha \leq R \\ \inf(-\alpha T, -\infty) = -\infty & \alpha < 0 \end{cases}$$

$$= \lambda(R - \alpha) - T\alpha + \lambda(\alpha)$$

Table 3.3 tabulates the lower slope transforms of some well-known NC functions for ease of reference.
3.4. Non-Preemptive Priority Scheduling Example

Consider a scheduling system that employs non-preemptive priority scheduling to serve incoming traffic with two classes of priority (high and low). This type of scheduler is typically used in DiffServ-based Internet. The server serves the queue with a constant rate $C$, while an absolute non-preemptive priority is given to the high priority traffic class. It can be easily shown that this scheduler serves the high priority flow as a rate-latency server $\beta_{R,T}(t)$ with parameters $R = C$ and $T = \frac{l_{\text{max}}^L}{C}$, resulting in service curve $\beta(t) = C(t - \frac{l_{\text{max}}^L}{C})^+$, where $l_{\text{max}}^L$ is the maximum packet size of the low priority traffic ([9], pp. 21).

In NC, for any high priority traffic described by the arrival rate $f^H(t)$, the output is calculated by $y(t) = (f^H \otimes \beta)(t)$, using the convolution operator.

As an illustrative example, assume that the high priority traffic has an arrival curve of $f^H(t) = p_r(t) = rt$, $0 < r < C$, known as CBR traffic. Using the NC convolution, the output is equal to

$$ (p_r \otimes \beta)(t) = \min\{r(t - \frac{l_{\text{max}}^L}{C}), C(t - \frac{l_{\text{max}}^L}{C})\} . \tag{3.11} $$

<table>
<thead>
<tr>
<th>Function</th>
<th>Lower slope transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-\lambda(t-t_0)$</td>
<td>$-\alpha t_0 - \lambda(\alpha)$</td>
</tr>
<tr>
<td>$b + rt$</td>
<td>$b + \mu_2(\alpha - r)$</td>
</tr>
<tr>
<td>$\gamma_r, b(t) - \lambda(t)$</td>
<td>$\lambda(r - \alpha)$</td>
</tr>
<tr>
<td>$p_R(t) = Rt - \lambda(t)$</td>
<td>$\lambda(R - \alpha)$</td>
</tr>
<tr>
<td>$\beta_{R, T} = R \times (t - T)^+$</td>
<td>$\lambda(R - \alpha) - T\alpha + \lambda(\alpha)$</td>
</tr>
<tr>
<td>$\exp(t)$</td>
<td>$\alpha(1 - \log(\alpha)) - \lambda(\alpha)$</td>
</tr>
<tr>
<td>$\mu(t-t_0)$</td>
<td>$-\alpha t_0$</td>
</tr>
</tbody>
</table>
Slope domain analysis, however, provides a more convenient solution. Relying on Theorem 3.1, the output can also be derived using slope domain analysis by calculating \( Y(\alpha) = F^H(\alpha) + B(\alpha) \). By using Theorem 3.1, we derive

\[
S\{p_r\} + S\{\beta\} = (l^{L}/C)\alpha + \lambda(r - \alpha) + \lambda(\alpha),
\]

since \( 0 < r < C \). This is simply the slope transform of the function \((f^H \otimes \beta)(t)\), formulated by (3.11). Figure 3.2 graphically illustrates these functions.

Following a similar procedure, the low priority flow is served by a rate-latency server with rate \( R = C - r \) and latency \( b/(C - r) \).

As a comparison, in the above example, we replace \( f^H(t) = p_r(t) = rt \) with a leaky-bucket shaper which has an arrival curve of \( f^H(t) = \gamma_{r,b}(t), \ 0 < r < C \), and \( f^H(0) = 0 \). Using the NC convolutions, the output is equal to

\[
(\gamma_{r,b} \otimes \beta)(t) = \min(b + r(t - l^{L}/C), C(t - l^{L}/C)). \tag{3.12}
\]

Similar to the previous example, by using Theorem 3.1, we have

\[
S\{\gamma\} + S\{\beta\} = (l^{L}/C)\alpha + \lambda(r - \alpha) + \lambda(\alpha),
\]

since \( 0 < r < C \). This is also the slope transform of the function \((f^H \otimes \beta)(t)\), i.e. \( S\{\gamma_{r,b} \otimes \beta\} = (l^{L}/C)\alpha + \lambda(r - \alpha) + \lambda(\alpha) \). However, as discussed in Theorem 3.2, when a function \( x(t) \) is not convex, the inverse slope transform of its slope transform results in a function \( \hat{x}(t) < x(t) \). Since \( \gamma_{r,b} \otimes \beta \) is not convex, we observe that the inverse slope transform of \( S\{\gamma_{r,b} \otimes \beta\} \) is \( S_U\{S_L\{\gamma_{r,b} \otimes \beta\}\}(-t) = \min(r(t - l^{L}/C), C(t - l^{L}/C)) < \gamma_{r,b} \otimes \beta \). However, the difference between these two functions is always less than \( b \), and negligible in practice as \( r \gg b \). Figure 3.3 graphically illustrates these functions.
Figure 3.2  Slope transforms of the peak-rate and rate-latency functions, and their convolution.
Figure 3.3  Slope transforms of the affine and rate-latency functions, and their convolution.
3.5. Network Identification Application

A prominent application of the slope transform is the identification of the minimum service curve of an unknown node or network, a practice analogous to the system identification concept in traditional system theory. System identification is used in manifold applications, such as modeling, measurement and control of arbitrarily complicated networks. While it is practically impossible to use normal NC filtering theory in this application, we show that the slope transform can be readily employed to derive the slope response of any system.

Assuming that the input $x(t)$ and output $y(t)$ to a system is known, we know that $y(t) = x(t) \otimes h(t)$, where $h(t)$ is the system service curve. Other than for simple functions, it is mathematically very difficult to solve this convolution-based equation to find $h(t)$. However, using the slope domain, the system is naturally identified by

\[ H(\alpha) = L\{y(t)\} - L\{x(t)\}. \]  

(3.13)

Therefore, knowing the slope transforms of some known input and output curves, and using (3.13), one can readily obtain the slope response of the system.

Although (3.13) explains the concept and presents a clear and logical way of finding the slope response of a system, in the following we introduce another more practical and numerically attractive approach. The method is based on direct observations of the outputs of the system to a set of affine functions, since the affine function is the eigenfunction of min-plus systems, i.e.,

\[ y(t) = (\alpha t + b) \otimes h(t) = \inf(\alpha (t - \tau) + b + h(\tau)) \]

\[ \begin{align*}
&= \alpha t + b + H(\alpha). \\
&= \alpha t + b + H(\alpha).
\end{align*} \]  

(3.14)

Therefore, in order to find the slope response of a system, we need to feed a set of affine functions related to a set of desired $\alpha_i$ values, and then by using (3.14), the corresponding set of $H(\alpha_i)$ values can be calculated. In practice, these functions can be obtained by using a leaky-bucket server that is saturated, and therefore outputs a traffic resembling the affine function.
This is a favorably effective method both in theory and in practice, as it works by observing the actual outputs to a set of inputs, and does not require using any convolution or slope transformation. To demonstrate the methodology by an example, let us assume that a hypothetical network has a service curve of $h(t) = 5t + 3t^2 - \lambda(t)$. The lower slope transform of this function, and therefore the system's slope response, is equal to $H(\alpha) = (-1/12)(\alpha - 5)^2$. Details of the analytic derivation of $H(\alpha)$ are presented in Section 3.11.

If we consider a set of 10 affine inputs with the form $at + b$, $5 \leq a \leq 95$ and $b = 4$ (shown in Figure 3.4a), and then subtract the corresponding set of the observed (or calculated) outputs from these affine functions (shown in Figure 3.4b), the results are equal to the values of the slope response at the corresponding $\alpha_i$ values. As shown in Figure 3.4b, it takes some time for the network to exhibit the corresponding value of the slope response, particularly with smaller values of $\alpha_i$. This lag is due to the fact that up to this time, the input is much less than the available guaranteed service.

Figure 3.5 compares the plot of the analytical slope response with that calculated using the above methodology. As shown in the plots, we were able to identify the slope response of the unknown system by observing the outputs resulting from the affine functions. In order to explain what the points are whose envelope represents the slope response, we present here the actual set of values of the calculated $H(\alpha)$ for the instance of $\alpha = 25$. This set of points (represented by the coloured dots highlighted on the figure) are the equidistant time samples of the plot associated with $\alpha = 25$ in Figure 3.4b (one sample every time unit, which is 1s here). The result is a set of 31 values of $[-4, -21, -32, -37, -16, -33, ..., -33]$. As we see, after just 5 values, it reaches a steady value of -33 and stays at that value (representing the slope response at $\alpha = 25$) for the rest of the set. In fact, the envelope consists of the overlap of many points of equal value that represent the steady state value. Similar to the behaviour of the traditional system functions, we also observe a small overshoot with the minimum value of -37, just before we reach a steady state condition. Similarly, for $\alpha = 75$, it is observed that after 12 initial values, it reaches the steady state value of -408 (also highlighted on the figure).
Figure 3.4  System identification by applying the slope domain analysis: (a) Plot of the service curve and inputs of the form $b+at$ for different values of $a$. (b) Plot of the observed outputs minus the inputs.
Simulation Methodology: The previous analytical calculations were carried out in MATLAB\textsuperscript{\textregistered} \textsuperscript{1}. In the following, the results from the analytical calculation are compared with the simulation results by using the OPNET\textsuperscript{\textregistered} modeler. The simulation methodology is as follows. A source generates a traffic stream with exponential inter-arrival times and constant packet sizes. This traffic is then shaped based on the $y_{\alpha_n \gamma}(t) = \alpha_n t + b$ function. The traffic source is set so that the shaped traffic, $x(t)$, is as close as possible to the arrival curve, i.e., in the definition of the arrival curve equality holds $(x(t) - x(s) = y_{\alpha_n \gamma}(t-s))$. The network is modeled by a work conserving scheduler with the service curve of $h(t) = 5t + 3t^2$, i.e., the scheduler ensures that during any time interval $[s, t]$, the amount of service dedicated to the traffic flow equals $h(t-s)$. As expressed by (3.14), the cumulative traffic received at the destination is equal to $y_{\alpha_n \gamma}(t) + H(\alpha_n)$. Then by using (3.14), the corresponding $H(\alpha_n)$ is calculated. Since the

\textsuperscript{1}MATLAB is a registered trademark of The MathWorks, Inc., and OPNET is a registered trademark of OPNET Technologies, Inc.
measured $H(\alpha_t)$ reaches a steady value (as shown in Figure 3.4b for $t > 15s$), only one value of the cumulative traffic received at the destination at any of these steady state times is needed. Thus, the measurements can be done offline or online. The simulation is repeated for all the previous 10 values considered for $\alpha_t$.

Figure 3.6 compares the calculated slope response from the above analysis with the scaled simulation results. Similar to Figure 3.5, the points represent the measured values of $H(\alpha_t)$ versus $\alpha_t$ at equidistant time samples during the simulation. The plot shows that the simulation results match closely with the analytic results, and therefore demonstrates that the slope domain analysis method closely identifies the slope response of the network. As seen, the results from the simulation exhibit some unexpected values, in addition to the overshoot effect seen in the analysis. This is mainly due to the fact that the actual traffic flow to the network is not exactly of the form $\gamma_{\alpha, b}(t) = \alpha t + b$; also in each simulation, the system needs some time to reach its steady state condition. However, the envelope of the results closely matches the expected slope
response.

**Discussion on the practical application of the method:** The above method helps us find the service curve of an unknown network, or to verify that an implemented network actually behaves according to the required service curve. In order to apply the above method to identify a network, we must consider the following issues:

1) The method works only if the network under examination has been designed to serve based on a service curve. It goes without saying that if the time domain convolution does not hold in the first place, the obtained results will not be meaningful. Therefore, the method applies only to QoS-based networks or network elements, not to best-effort ones.

2) A practical issue for using the network identification method is to consider a proper set of values for $\alpha_i$. Here we discuss the situations when too large or too small values are chosen. First, it is worth emphasizing that a service curve is guaranteed for a specific input traffic flow, only if the flow obeys the negotiated arrival curve. Therefore, input traffic flows with large values of $\alpha_i$ do not comply with the negotiated service quality constraints, and will be served only up to the negotiated service quality. In theory, this will not have an impact on the application of the method, since the value of $H(\alpha_i)$ (which is a negative value) decreases as $\alpha_i$ increases. However, practical systems do not have infinite buffer sizes and the excess traffic may be dropped. In this condition, the observed output traffic flow no longer represents what we expect by (3.14). Second, if the values of $\alpha_i$ are chosen too small, then it may take a long time for the system to serve enough traffic to reach a steady state. If the test duration is short, the results will not reflect the slope response of the network. This condition can be viewed from Figure 3.4, when $t < 10 s$. However, as explained earlier, the measurements can be done online, so the test duration can be set dynamically. This is done by observing the measurements, and stopping when they reach a steady value. These conditions occur when the network is completely unknown, and the range of values for $\alpha_i$ or the test duration is not selected correctly. In practice, we usually have a basic understanding of the examined network that helps us consider educated values for $\alpha_i$. 
3) In a network, service curves are usually associated with specific input traffic flows. For example, a network of WFQ (weighted fair queuing) schedulers serves different traffic classes with different service curves. Therefore, the traffic has to be tagged with an appropriate traffic class.

4) The method works only for shift-invariant service curves. This is a reasonable assumption for most QoS-aware networks, such as IntServ, Diffserv and ATM, and is a desirable characteristic for man-made systems. The validity of such an assumption can be tested by repeating the identification method for the same input traffic and comparing the received output traffic flows. A related condition is when a faulty node is in the path of the traffic flow. In this condition, the service curve has actually changed; therefore, the results will not reflect the service curve of the network in normal conditions.

5) When a server is not exact, it guarantees that the complying traffic stream will receive a minimum (or maximum) service defined by the service curve (as explained in Chapter 2). As an example, assume that a WFQ server assigns a lower service curve \((t - T_i)r_i\) to class \(i\). This traffic class will be served more than the amount specified by this service curve, if other traffic classes are not fully using their shares. Therefore, it should be ensured that there is enough mix of traffic in the system, when the purpose is to identify lower service curves. This way, the examined traffic flow will receive only up to its assigned service and the results will reflect the correct service curve.

6) In order to create the required affine functions, leaky-bucket shapers in saturation can be used. However, in practice the values and range of values of \(\alpha_i\) that can be actually achieved depend on the practical aspect of implementing a leaky-bucket server that is capable of working with the required range of rates.
3.6. Optimal Smoothing for Multimedia Streaming

In a network that offers guaranteed service based on a service curve $G(t)$, to variable bit rate (VBR) traffic $A(t)$ that is smoothed using a traffic envelope $\xi(t)$, an optimal strategy for the smoother exists that minimizes the playback delay and the required receiver buffer size [41]. Therefore, in contrast to shaping or rate control, smoothing focuses on meeting the QoS criteria at the receiver side. Figure 3.7 shows the assumed system topology. Optimal smoothing is explained and used in Chapter 6. For such a scenario, the optimal smoothing strategy, and correspondingly, the required minimum playback delay and buffer size, are expressed by the following main results [41]:

R1) The minimum playback delay is defined by

$$D = h(A, \xi \otimes G),$$

where $h(A, \beta) = \sup_{s \geq 0} \{ \inf \{ T : T \geq 0 \text{ and } A(s) \leq \beta(s + T) \} \}$.

R2) The optimal smoother is

$$A' = A \otimes (\xi \otimes G).$$

The optimum output $A'(t)$ gives the latest time at which every packet of the flow should be scheduled ([41], pp. 690). Note that $A'(t + u) - A'(t) \leq \xi(u)$.

R3) The maximum optimal required buffer size is $X^- = \sup_{t \geq 0} \{ A^-(t) - A(t) \}$, where
Proposition 3.1. For any function \( f(t, \tau) \), \( \inf_t \{ \inf_{\tau} \{ f(t, \tau) \} \} = \inf_{\tau} \{ \inf_t \{ f(t, \tau) \} \} \).

Proof: Assume that \( \inf_{\tau} \{ f(t, \tau) \} = g_1(t) \) and \( \inf_t \{ g_1(t) \} = m' \). Then for \( \forall (t, \tau) \), \( m' \leq f(t, \tau) \).

Now assume that \( \inf_t \{ f(t, \tau) \} = g_2(\tau) \) and \( \inf_{\tau} \{ g_2(\tau) \} = m'' \). Then for \( \forall (t, \tau) \), \( m'' \leq f(t, \tau) \). Then \( m'' = m' \), concluding the proof. \( \square \)

Proposition 3.2. For a set of real numbers, \( M \): \( \sup(M) = -\inf(-M) \).

Proof: Let \( m = \sup(M) \), i.e. \( \forall m' \in M \), \( m \geq m' \). Then \( -m \leq -m' \), i.e. \( -m = \inf(-M) \). \( \square \)

Theorem 3.3. For two min-plus functions \( x(t) \) and \( y(t) \),

\[
S_L \{ x \ominus y \} = S_L \{ x \} - S_U \{ y \},
\]

where \( \ominus \) is the deconvolution operator expressed by (2.7).

Proof: \( S_L \{ x(t) \ominus y(t) \} = \inf_t \{ \inf_{\tau} \{ x(t + \tau) - y(\tau) \} - \alpha t \} \)

\[
= \inf_t \{ \inf_{\tau} \{ x(t + \tau) - \alpha t \} - y(\tau) \}
\]

\[
= \inf_t \{ X_L(\alpha) + \alpha \tau - y(\tau) \}
\]

\[
= X_L(\alpha) + \inf_{\tau} \{ \alpha \tau - y(\tau) \}
\]

\[
= X_L(\alpha) - \sup_{\tau} \{ y(\tau) - \alpha \tau \}
\]

\[
= X_L(\alpha) - Y_U(\alpha)
\]

In the second line, Proposition 3.1 is used. In the third line, P2 of Table 3.1 is applied, and finally, in the fifth line, Proposition 3.2 is used. \( \square \)
In the following, the equivalent slope representation of R1 and R2 are presented. R3 can be used as is and needs no extra manipulations.

**Proposition 3.3.** For two functions \( x(t) \) and \( y(t) \), if \( x \leq y \ \forall t \), then \( S\{x(t)\} \leq S\{y(t)\} \).

Proof: \( x \leq y \Rightarrow x - \alpha t \leq y - \alpha t \Rightarrow \inf_t \{x - \alpha t\} \leq \inf_t \{y - \alpha t\} \), proving the proposition. The proof is based on the fact that if \( \tilde{x} \leq \tilde{y} \) then \( \inf \{\tilde{x}\} \leq \inf \{\tilde{y}\} \). □

**R1 - S)** Expression (3.15) in R1 translates to \( \overline{D} = h(S\{A\}, S\{\xi\} + S(G) - \alpha T) \).

Proof: \( h(A, G) \) is the maximum value of the set of the smallest \( T \)’s that for each \( s \geq 0 \) satisfies: \( A(s) \leq G(s + T) \). Using Proposition 3.3, \( h(A, G) \) is also the maximum value of the set of the smallest \( T \)’s that for each \( A \) satisfies: \( S\{A(s)\} \leq S\{G(s)\} + \alpha T \), which again translates to the maximum horizontal distance between the slope transform of \( A(s) \) and the shifted slope transform of \( G(s) \). The slope transform of \( G(s) \) is shifted by \( \alpha T \) which is a simple line in the slope domain with a constant slope of \( \alpha T \). Therefore, \( h(A, \xi \otimes G) \) is the maximum value of the set of the smallest \( T \)’s that for each \( s, s \geq 0 \) satisfies \( A(s) \leq (\xi \otimes G)(s + T) \). Therefore, (3.15) is equal to

\[
\overline{D} = h(S\{A\}, S\{\xi\} + S(G) - \alpha T).
\]

(3.17)

It is worth noting that \((\xi \otimes G)(t + T)\) is not equivalent to the convolution of \( \xi(t + T) \) and \( G(t + T) \), which by using Theorem 3.1 incorrectly results in \( S\{\sigma\} - \alpha T + S(\beta) - \alpha T \). But \((\sigma \otimes \beta)(t)\) needs to be calculated, and then a shift in time of \( t + T \) needs to be applied. □

**R2 - S)** Using Theorem 3.3, (3.16) can be rewritten as

\[
S_L\{\bar{A}'\} = S_L\{A\} - (S_U(\xi) + S_U(G))
\]

(3.18)
3.7. Load Spectrum

Power spectrum is the application of the Fourier transform to stochastic signals and systems. Similarly, based on the theorems presented in the next chapter, the concept of the load spectrum is introduced, which extends the application of the slope transform to the stochastic settings. This concept is addressed in detail in Chapter 4.

3.8. Further Applications

Slope domain may be used in many other applications. One can look at numerous applications of the Fourier transform and find or replicate similar applications for the slope transform. In the following, several more of these applications are addressed.

Slope Selective Filters: A practical advantage of the slope transform is its application in filtering certain slope elements. Principal filters that are normally used in traditional system theory to design various filter actions include low-pass, band-pass and high-pass, notch and all-pass filters. These filters filter out parts of a signal with certain frequency content. Slope domain modeling provides similar types of servers.

Slope low-pass filter: As we saw earlier, the leaky-bucket shaper is a slope low-pass filter. In general, a slope low-pass filter is defined by \( \alpha_0 t - \lambda(t) \), with a slope response of \( \lambda(\alpha_0 - \alpha) \), which rejects all slopes higher than \( \alpha_0 \) and passes the rest.

Similarly, these filters can be used to devise more advanced filters, by using them in concatenation or in parallel.

Traffic Aggregation (Slope Domain Modulation): In traditional system theory, the multiplication of two signals results in the convolution of the signals in the frequency domain (called modulation). In the min-plus theory, the addition of two signals results in their convolution in slope domain, which we call slope modulation. Therefore, when a number of traffic flows are multiplexed (added), the slope domain of the resulting traffic flow is their slope modulation. This concept may provide insight into the concept of traffic aggregation versus per-flow analysis.
3.9. Comparing Fourier and Slope Domains

The Fourier transform is a fundamental tool in the study of signals and systems and has many applications. In particular, this transformation provides valuable insight and effective tools in many aspects of design and analysis of communication systems, filtering methods, and signal processing algorithms [49][50].

The slope transform is in many respects similar to the Fourier transform. The main advantage of both transforms is the transformation of the convolution operator into easier algebraic operators. This property not only facilitates many mathematical calculations, but also creates a myriad of properties with important applications, such as the modulation property. Interestingly, most properties of one transform are similar to those of the other and can be guessed from the other, for example, the time shift or scaling properties. In fact, many of the applications and properties considered for the slope transform are inspired by the corresponding Fourier transform concepts. In the following, some of the prime characteristics of each are compared.

1) For any system, the response to the signal $a e^{s_1 t}$ is $a H(s_1)e^{s_1 t}$, where $a e^{s_1 t}$ is the eigenvalue and $H(s_1)$ is the eigenfunction of the system. In addition,

$$a_1 e^{s_1 t} + a_2 e^{s_2 t} \rightarrow a_1 H(s_1)e^{s_1 t} + a_2 H(s_2)e^{s_2 t}.$$

For any min-plus system, for the input $at + b$ we have $at + b \rightarrow at + b + S(a)$, where $at + b$ is the eigenvalue and $S(a)$ is the eigenfunction of the system. In addition,

$$\min(a_1 t + b_1, a_2 t + b_2) = (a_1 t + b_1 \land a_2 t + b_2)$$
$$\rightarrow (a_1 t + b_1 + S(a_1) \land a_2 t + b_2 + S(a_2))$$

2) In the Fourier domain, the following relationships hold for the fundamental operators of $(+, \times)$ and $\otimes$:

For $(+): F\{x(t) + y(t)\} = F\{x(t)\} + F\{y(t)\}$. 

Chapter 3. Slope Domain Modeling and Analysis of Data Communication Networks

For \( (\times) \):\( F\{x(t) \times y(t)\} = F\{x(t)\} \otimes F\{y(t)\} \).

For convolution:\( F\{x(t) \otimes y(t)\} = F\{x(t)\} \cdot F\{y(t)\} \).

Similarly, in the slope domain, the following relationships for the fundamental operators of \((\wedge, +)\) and \(\otimes\) hold:

For \((\wedge)\):\( S\{x \wedge y\} = S\{x\} \wedge S\{y\} \).

For \((+)\):\( S\{x + y\} = S\{x\} \oplus S\{y\} \).

For convolution: \( S\{x \otimes y\} = S\{x\} + S\{y\} \).

It is crucial to note that the addition and multiplication operators in the traditional algebra are replaced with the minimum and addition operators in min-plus algebra, respectively. Also, the idempotent property of the min-plus algebra plays a unique role in both time and slope domains.

3.10. Comparing Slope with the Legendre Transform and Fenchel Conjugate

The objective of this chapter in presenting the slope domain methodology was to find the best transformation domain for the application of the min-plus system theory in data communications networks. This work was originally motivated by the introduction of the application of the Legendre transform for this purpose in [34], and by the slope transform-based morphological signal processing in [47]. Therefore, we need to introduce the Legendre Transform and Fenchel conjugate and compare them with the slope transform. Concurrent with our work, some recent papers have been published on these transforms that focus only on convex functions [36][37]. The Legendre transform is a familiar transform in the mathematics of min-plus algebra. The Legendre transform presents some drawbacks, for example, it is only defined for convex functions, and is only in accordance with min-plus algebra. In contrast, the slope domain model is complete and systematic and considers all cases. Actually, the Legendre transform is a special instance of the slope transform that is defined for and used with convex or concave functions with invertible
derivatives [47]. The Fenchel conjugate is a close relative of the upper slope transform. Since a vast quantity of literature is available on them, they also bring further insights to the application of and analysis in the slope domain.

**Legendre Transform.** For a concave or convex function \( x(u) \) which has an invertible derivative, its Legendre transform is defined as [34]

\[
\Psi \{ x(t) \} (s) = x(u^*) - su^*, \text{ where } s = \frac{d(x(u^*))}{du},
\]

\( s \) is the slope of the function at \( u^* \), as \( s = \lim_{u \to u^*} \frac{x(u) - x(u^*)}{u - u^*} \).

Writing (3.19) based only on the variable \( s \) [35], the transform is also defined as

\[
\Psi \{ x(t) \} (s) = x((x')^{-1}(s)) - s[(x')^{-1}(s)].
\]

If a function in the Legendre domain, \( X \), is also convex or concave and has an invertible derivative, then the inverse Legendre transform is defined by

\[
x(t) = X((X')^{-1}(-t)) + t[(X')^{-1}(-t)].
\]

**Fenchel Conjugate.** For a function \( c(x) \), its Fenchel conjugate is defined as [51]

\[
\Im \{ c(x) \} (\theta) = \sup_{x} (\theta x - c(x)).
\]

The Fenchel conjugate is closely related to the slope transforms. The slope transforms can be expressed by the Fenchel conjugate, as follows:

\[
S_L \{ x(t) \} (\alpha) = \inf_{t \in \mathbb{R}} (x(t) - \alpha t) = -\sup_{t \in \mathbb{R}} (\alpha t - x(t)) = -\Im \{ x(t) \} (\alpha),
\]

and

\[
S_U \{ x(t) \} (\alpha) = \sup_{t \in \mathbb{R}} (x(t) - \alpha t) = \sup_{t \in \mathbb{R}} ((-\alpha) t - (-x(t))) = \Im \{ -x(t) \} (-\alpha).
\]
Similar to the slope transform, the Fenchel conjugate can be used with all functions. However, the main focus of the literature addressing the Fenchel conjugate is on convex functions, since when applied to a reasonably convex function, the function can be reconstructed by applying the Fenchel conjugate a second time. Originally, it was defined as an alternative to the Legendre transform, based on the supremum notation. Therefore, unlike the Legendre transform, it does not require the function to be differentiable.

The next section discusses how the Legendre transform and Fenchel conjugate may serve as useful tools to facilitate calculations and enhance the application of the slope transforms.

3.11. Computational Improvement Using the Legendre Transform and Fenchel Conjugate

The slope transform is comprehensive, straightforward and self-sufficient both in analytical and numerical computations. However, due to the proximity of the slope transforms with the Legendre transform and Fenchel conjugate, the literature related to these transforms are advantageous and can be used to facilitate calculations of the slope transforms both numerically and analytically. In the following, we present two important applications.

1) As mentioned earlier, the Legendre transform of a function is equal to its lower slope transform when the function is convex with invertible derivatives, and is equal to its upper slope transform when the function is concave with invertible derivatives, ignoring possible differences due to boundary effects. Since the Legendre and the reverse Legendre transforms have easier-to-find and easier-to-use closed-form representations, when applicable, the Legendre transform provides an easier way of finding the slope transform of a function analytically.

As an analytical example, for the parabola function $x(u) = u^2$, the lower slope transform is equal to its Legendre transform, i.e.

$$S_L\{x(u)\}(s) = \Psi\{x(t)\}(s) = (s/2)^2 - s(s/2) = -s^2/4.$$
Another analytical example is the function $x(t) = 5t + 3t^2$, which we used in Section 3.5. Similarly, since it is convex and has invertible derivatives, its lower slope transform is equal to its Legendre transform, which is easier to derive. We have $\alpha = x'(t) = 5 + 6t$, then $t = ((\alpha - 5)/6)$. Using (3.20), we have

$$X_L(\alpha) = 5((\alpha - 5)/6) + 3((\alpha - 5)/6)^2 - \alpha((\alpha - 5)/6),$$

resulting in $X_L(\alpha) = (-1/12)(\alpha - 5)^2$. Since the function is causal and is zero for $t < 0$, then $X_L(\alpha)$ is defined as above for $\alpha \geq 5$.

2) The close proximity of the slope transforms with the Fenchel conjugate allows us to use the efficient algorithms that are available for the calculation of the Fenchel conjugate, for numerical computation of the slope transforms. A number of such fast computation methods are the fast Legendre-Fenchel transform [52], linear-time Legendre-Fenchel transform (LLT) [53], and parallel fast Legendre-Fenchel transform [54]. These algorithms serve a role analogous to FFT in the conventional system theory, where they can be used in the analysis of complex networks with arbitrary arrival and service curves.

### 3.12. Summary

In this chapter, we have introduced the slope domain analysis of queuing systems as a complement to the NC approach. The transform and related properties are discussed and the slope transforms of useful NC service curves and arrival curves have been tabulated. We have elaborated on the efficiency of the slope transform in the analysis of computer networks, by applying the methodology to system identification and non-preemptive priority scheduling problems. We have discussed the special cases of the Legendre transform and Fenchel conjugate as useful tools for fast calculation of the slope transform for some common functions.
In this chapter, we develop a framework for efficient stochastic QoS evaluations and cross-layer design in wireless networks. The methodology is based on a proposed stochastic min-plus system theory. This novel theory characterizes the second-order statistics of the system output and can be used to model or analyze any stochastic system, in particular a wireless channel, with manifold advantages. Due to its clearly similar representation to the familiar traditional stochastic system theory, it is an easy yet elaborate modeling and performance evaluation tool, even with just a basic familiarity of NC theory. Above all, as it only deals with second-order statistics, it is a practical technique for establishing a packet-level cross-layer methodology for a wireless channel at the link layer by the knowledge of its second-order statistics at the physical layer. Finally, it conforms and can be easily coupled with NC, a sophisticated min-plus system theory for QoS evaluation and modeling of data communication networks, which is an important requirement for end-to-end QoS. We demonstrate the effectiveness of the method by using the proposed methodology to calculate the delay bounds and/or how to set the transmission rates to meet a certain delay bound, and present comparisons between analytic and simulation results. The chapter concludes with a thorough discussion on the underlying assumptions, related topics and applications with different settings.
4.1. Introduction

Deterministic NC is a well-established, sophisticated min-plus system theory for deterministic QoS evaluations and modeling of data communication networks [9]. However, in wireless networks in which a wireless link behaves like a stochastic server, deterministic NC might estimate loose QoS bounds. Currently, there are two different frameworks for stochastic or probabilistic NC extensions. The \((\sigma(\theta), \rho(\theta))\)-calculus and effective bandwidth method [8] are based on the constrained moment generating function of the traffic, and the other method is based on stochastically bounded burstiness [20].

In this chapter, we propose and develop a stochastic min-plus theory as an efficient alternative method for performance evaluations and modeling, especially suited to wireless networks. The proposed methodology also serves as the foundation of a cross-layer design and modeling framework for wireless networks. In contrast with other stochastic NC methods, our proposal has the important advantage that it can easily be coupled or used with deterministic NC, an important requisite for end-to-end QoS. In addition, due to its clearly similar representation to the traditional stochastic system theory, it is much easier to use and is well-understood, even with a basic familiarity of NC theory. Simply, the traffic needs to be represented by NC’s cumulative traffic characterization, and min-plus (or max-plus) convolution is used for input-output representations.

The approach that is presented in this chapter was inspired by the traditional system theory for modeling linear systems with stochastic inputs [55]. Min-plus (or its dual, max-plus) systems, as the basis of NC theory, are min-plus (or max-plus) linear. We demonstrate that a similar approach can be used to stochastically characterize the output of such systems, when the characteristics of the system and/or of the input are only stochastically provided. For ease of notation and to follow the system theory model, we refer to a network element (such as a link, node, or a whole network) as a system. Similarly, the incoming traffic to the system is called input, and the outgoing traffic from the system is called output, both with bytes as their units of measure. In our wireless link, the input is the traffic transmitted over the link, and the output is the traffic passed through and received at the other end of the link, observed at the link layer.
The outline of the chapter is as follows. In the next section, a brief overview of the traditional system theory with a stochastic input is presented, which serves as a point of reference for and comparison with the presented stochastic min-plus theory. Second-order input-output relationships of a min-plus system with a stochastic input are presented in Section 4.3. Then, by utilizing the commutative property of the convolution operator, we modify the previous method to analyze a min-plus system with a stochastic service curve. This modification provides us with an efficient method for modeling communication networks with stochastically defined arrival or service processes, such as wireless channels. Based on this new system theory, we propose a packet-level performance evaluation and cross-layer framework for wireless networks, when the second-order statistics of the physical layer are known. The link layer uses these statistics to calculate the second-order statistics of the traffic that will be received at the other end of the channel, and consequently the projected delays. From these information, suitable parameters of operation, such as the transmission rate, are decided upon. Two well-known channel models of AWGN and Rayleigh fading are examined. By analysis and simulation via an example, we demonstrate how effectively delay as the QoS metric of interest is calculated. Finally, after some discussions, the chapter is concluded in Section 4.8.

4.2. Background

A restricting and challenging characteristic of a wireless channel is that it is random and time-varying, and so is regarded as a stochastic system. Therefore, practical modeling of wireless networks naturally requires stochastic models. In Chapter 6, we present a method based on the definition of the channel's service curve using filter banks of deterministic NC elements, where the service is guaranteed with some probability [25]. Although the method is efficient and serves as a natural extension of the deterministic NC, its successful application relies on the derivation of the piece-wise service curve with a meaningful probability. Other efforts are rooted in the stochastic traffic characterization, assuming that some function of the moment generating function of the underlying process is constrained [8]. These models, such as the effective capacity model [16], are not only complicated and require knowledge of the asymptotic log-moment generating functions, but are also mainly suited to the analysis of a standalone link.
In this chapter, we present a systematic and effective methodology for the derivation of second-order statistics of the output of a stochastic min-plus system. Since our method takes a course different from those referred to by stochastic NC in the literature (such as [8] or [20]), we differentiate it by entitling it stochastic min-plus systems theory. However, the methodology conforms with NC theory, and can actually be categorized as a stochastic NC approach and be integrated to a NC-based analysis of a network, especially with deterministic NC. In contrast to the stochastic NC models where the moment generating of the stochastic process is considered, here we deal with the second-order statistics of mean and autocorrelation. This is a practical approach, as such statistics are usually available from the underlying physical layer models. This information may be used by the link layer to set its parameters accordingly.

The approach has many advantages. It is inspired by and shares many of the advantages of the traditional stochastic system theory ([55], Ch 10): it is easy to use, easily understood, systematic and scalable. Also, since the classical system theory is familiar to engineers, our proposed stochastic min-plus system theory is accessible to practitioners even with just a basic familiarity with min-plus theory (or NC). In terms of its applicability, the requirements are that the cumulative traffic be characterized by a non-decreasing envelope and the min-plus convolution be applied. Finally, it is based on the second-order statistics which can be used for efficient QoS evaluations and resource allocation in multi-service networks [56][57].

In the following, we summarize the input-output relationship of a traditional linear system with a stochastic input, as a reference and a point of comparison. Then we present a concise overview of the min-plus system service curve to describe the notations used in this chapter.

4.2.1. Traditional Systems with Stochastic Inputs

In the traditional system theory, the relationship between the input $x(t)$ and the output $y(t)$ of a linear system with response $h(t)$ is characterized by the convolution operator defined by the notation

$$y(t) = L^s[x(t)], \text{ where } L^s[x(t)] = \int x(t-s)h(s)ds.$$  \hspace{1cm} (4.1)
Chapter 4. A Stochastic Min-plus System Theory Approach to Performance Modeling of Wireless Networks

We use a superscript $s$ with $L$ to differentiate the system operator of the traditional systems from the corresponding min-plus notation.

For a stochastic input $x(t)$ where its mean $E\{x(t)\}$ and its autocorrelation function $R_{xx}(t_1, t_2)$ are given, similar statistics of the output are characterized based on the following two fundamental theorems.

**Theorem 4.1.** $E[L^s[x(t)]] = L^s[E\{x(t)\}]$, which also results in $n_y = n_x \otimes h$. This is an extension of the linearity of the expected values to arbitrary linear operators ([55], pp. 309).

**Theorem 4.2.** The autocorrelation of the output is characterized by ([55], pp. 311)

\[
R_{xy}(t_1, t_2) = L_2^s[R_{xx}(t_1, t_2)],
\]

\[
R_{yy}(t_1, t_2) = L_1^s[R_{xy}(t_1, t_2)],
\]

where $L_2$ means that the system operates on the variable $t_2$ and treats $t_1$ as a parameter, and vice versa for $L_1$. The autocorrelation function is defined by $R_{xx}(t_1, t_2) = E\{x(t_1)x^*(t_2)\}$, $t_2 \geq t_1$.

A process is wide-sense stationary (WSS) if and only if its mean is constant: $E\{x(t)\} = \eta$, $\forall t$ and its autocorrelation depends only on $\tau = t_2 - t_1$: $E\{x(t + \tau)x(t)\} = R(\tau)$, $\forall t$. Then,

\[
R_{xy}(\tau) = R_{xx}(\tau) \otimes h^*(-\tau),
\]

\[
R_{yy}(\tau) = R_{xy}(\tau) \otimes h(\tau).
\]

4.3. Min-Plus Systems with Stochastic Inputs

The min-plus system theory and its system function were introduced in Chapter 2 and depicted in Figure 2.1. In a min-plus system, the output $y(t)$ is characterized by the min-plus convolution of the input $x(t)$ and the service curve (system function) $h(t)$ by
\[ L^{h(t)}[x(t)] = x(t) \otimes h(t) = \inf_{0 \leq s \leq t} (x(t-s) + h(s)) = \inf_{0 \leq s \leq t} (x(s) + h(t-s)). \] (4.4)

and

\[
\begin{align*}
y(t) & \geq L^{h(t)}[x(t)], & \text{if } h(t) \text{ is a lower service curve}, \\
y(t) & = L^{h(t)}[x(t)], & \text{if } h(t) \text{ is an exact service curve}, \quad (4.5) \\
y(t) & \leq L^{h(t)}[x(t)], & \text{if } h(t) \text{ is an upper service curve}.
\end{align*}
\]

In this section, we introduce the theory of min-plus systems with stochastic inputs by presenting two fundamental theorems characterizing the mean and autocorrelation of the output. We consider only exact and upper service curves. Due to the commutative property of the min-plus convolution, later we simply swap the role of input as the stochastic process with the service. Therefore, this section lays the foundation for the stochastic min-plus system theory presented in the following section.

In order to comply with the NC theory, all functions describing the input, system and output are positive and cumulative, and therefore non-decreasing. Note that for \( x(t) \), we have the option of using either the actual traffic or its arrival curve. The former gives the actual instance of the output, while the latter gives an envelope for the output. Since for an arrival process \( x(t) \) with an arrival curve \( \zeta(t) \) we have \( x(t + \tau) - x(t) \leq \zeta(\tau), \ \forall (\tau \geq 0) \). By setting \( \tau = 0 \), we have \( x(t) \leq \zeta(t) \) for \( \forall (t \geq 0) \) and \( x(0) = 0 \). Finally, using the isotonicity of \( \otimes \) [9] leads to \( y(t) \leq \zeta(t) \otimes h(t) \), which means the output obeys a curve of \( (\zeta \otimes h)(t) \).

For any cumulative function \( x(t) \), a rate function \( X(t) \) is defined. If \( x(t) \) is continuous and has a derivative, then \( x(t) = \int_{0}^{t} X(\tau) d\tau \) [28]. For a discrete time model, the input is mapped by sampling every time slot \( \delta \) [9].

In communications systems, the statistics of \( X(t) \) (which is WSS) are usually known. If \( X(t) \) is WSS, then \( E\{X(t)\} = \eta \), simply resulting in \( E\{x(t)\} = \eta t \), for both continuous and discrete time cases. Obviously \( x(t) \) itself is not WSS, unless \( \eta = 0 \). Similarly for the autocorrelation of the continuous case, we have
Chapter 4. A Stochastic Min-plus System Theory Approach to Performance Modeling of Wireless Networks

\[ R_x(t_1, t_2) = E\left[ \int_0^{t_1} X(m)dm \int_0^{t_2} X(n)dn \right] = \int_0^{t_1} \int_0^{t_2} R_X(m, n)dndm. \]

For an uncorrelated process (i.e. \( E\{X(m + n)X(m)\} = 0, \) for \( n \neq 0 \)), this reduces to

\[ R_x(t_1, t_2) = VAR(X)t_1t_2, \]

where \( VAR(X) = R_X(0) \) is the variance of \( X \).

Before presenting the fundamental theorems, we address some essential propositions and lemmas.

**Proposition 4.1.** For a set of real numbers \( X \), \( \inf(X + c) = \inf(X) + c \), and \( \inf(\beta \cdot x) = \beta \cdot \inf(x) \), \( \beta > 0 \).

Proof: Assume \( x^* = \inf(X) \), then \( x^* \leq x \) for \( \forall x \in X \). Then \( x^* + c \leq x + c \) and \( \beta \cdot x^* \leq \beta \cdot x, \forall \beta \geq 0 \). So \( x^* + c = \inf(X + c) \) and \( \beta \cdot x^* = \inf(\beta \cdot X) \).

**Proposition 4.2.** For two random variables \( x \) and \( y \), if \( x \geq y \) (i.e., for any outcome the value of \( x \) is greater than or equal to \( y \)), then \( E\{x\} \geq E\{y\} \).

**Lemma 4.1.** For a set of random variables \( X_S \),

\[ E\{\inf_S(X_S)\} \leq \inf_S(E\{X_S\}) \tag{4.6} \]

Proof: Assume that for \( s^o \) we have \( X_{s^o} = \inf(X_S) \), then \( X_{s^o} \leq X_s, \forall s \in S \). Using the results of Proposition 4.2, we have \( E\{X_{s^o}\} \leq E\{X_s\}, \) i.e. \( E\{\inf_S(X_S)\} \leq \inf_S(E\{X_S\}) \).

**Theorem 4.3.** In a min-plus system with a service curve \( h(t) \) and random process \( x(t) \),

\[ E\{L[x(t)]\} \leq L[E\{x(t)\}], \tag{4.7} \]

Proof: At a given time \( t \), define a random variable \( Y_t = \inf_{0 \leq s \leq t}(x(t-s) + h(s)) \), in which \( h(t) \) is a function of time. From Lemma 4.1, \( E\{Y_t\} \leq E\{x(t-s) + h(s)\}, \forall (s, 0 \leq s \leq t) \).
Thus, \( E\{Y'\} \leq \inf_{0 \leq s \leq t}(E\{x(t-s)\} + h(s)) \). From the definition stated by (4.4), we have \( L[x(t)] = Y' \) and \( L[E\{x(t)\}] = \inf_{0 \leq s \leq t}\{E\{x(t-s)\} + h(s)\} \), thus (4.7) follows. \( \square \)

**Theorem 4.4.** In a min-plus system with an exact or upper service curve \( h(t) \)

\[ \eta_y \leq \eta_x \otimes h \]  

(4.8)

Proof: By applying Theorem 4.3 in (4.5), the stated result follows.

Similar to (4.2) in Theorem 4.2, the autocorrelation functions of the output in a min-plus system are derived as follows:

**Theorem 4.5.** For a min-plus system

\[ R_{xy}(t_1, t_2) \leq L_2^{\eta_x(t_1)h(t_2)} [R_{xx}(t_1, t_2)] , \]  

(4.9)

and

\[ R_{yy}(t_1, t_2) \leq L_1^{\eta_y(t_2)h(t_1)} [R_{xy}(t_1, t_2)] , \]  

(4.10)

when \( h(t) \) is an exact or upper service curve. In the above equations, \( L_2 \) means that the system operates on the variable \( t_2 \) and treats \( t_1 \) as a parameter, and \( L_1 \) means that the system operates on the variable \( t_1 \) and treats \( t_2 \) as a parameter.

Proof: We prove the theorem for exact service curves. For upper service curves, the proof is similar by just replacing the equalities on the first three lines with \( \leq \). To prove (4.9), we start with the convolution equation and then multiply the two sides with the positive value \( x(t_1) \). We apply expectation to both sides, then by using Lemma 4.1, (4.9) follows with \( t = t_2 \). Similarly, we start with the convolution equation but this time we multiply the two sides with the positive value \( y(t_2) \). Then we apply expectation to both sides and by using Lemma 4.1, (4.10) follows with \( t = t_1 \).
\begin{align*}
y(t) &= L^{h(t)}[x(t)] \\
y(t)x(t_1) &= \inf_s (x(t-s) + h(s))x(t_1) \\
E\{y(t)x(t_1)\} &= E\{\inf_s (x(t-s) + h(s))x(t_1)\} \\
  &\leq E\{\inf_s (x(t-s)x(t_1) + h(s)x(t_1))\} \\
  &\leq \inf_s (E\{x(t-s)x(t_1)\} + h(s)E\{x(t_1)\}) \\
  &\leq \inf_s (R_{xx}(t_1, t-s) + \eta_x(t_1)h(s)) \\
R_{xy}(t_1, t_2) &\leq L^{n(t_1)h(t_2)}_2[R_{xx}(t_1, t_2)]
\end{align*}

and
\begin{align*}
y(t) &= L^{h(t)}[x(t)] \\
y(t)y(t_2) &= \inf_s (x(t-s) + h(s))y(t_2) \\
E\{y(t)y(t_2)\} &= E\{\inf_s (x(t-s) + h(s))y(t_2)\} \\
  &\leq \inf_s (E\{x(t-s)y(t_2)\} + h(s)E\{y(t_2)\}) \\
  &\leq \inf_s (R_{xy}(t-s, t_2) + \eta_y(t_2)h(s)) \\
R_{yy}(t_1, t_2) &\leq L^{n(t_2)h(t_1)}_2[R_{xy}(t_1, t_2)]
\end{align*}

As demonstrated in Figure 4.1, we may use two different time axes to define the two time instances \(t_1\) and \(t_2\) that we need to calculate the autocorrelation. For clarity and only in this figure, we use two different notations for time, in order to differentiate between the transmission times \(\tau\) and observation times \(t\). In other words, the autocorrelation can be calculated for one instance of output over actual time span during a transmission, or two different times of two different instances of output. Here we are interested in calculating the autocorrelation over a constant \(\tau\), for \(t_1, t_2\) due to two reasons. First, we are interested in the variability of the output for a transmission. Second, this will ensure that \(R_{yy}\) is WSS, provided that the underlying physical layer process (signal-to-noise ratio (SNR) here) is WSS, which is a widely accepted and valid assumption.
Fundamental Theorems in Max-plus Algebra: Since max-plus algebra is the dual of min-plus algebra, both of the fundamental theorems proved here are applicable to max-plus systems as well. Of course, all infimum operators are replaced with supremum, and the direction of inequalities is revered for the service curve \( h(t) \). Theorem 4.3 becomes

\[
E\{ \tilde{L} [x(t)] \} \geq E\{ \tilde{L} [x(t)] \}, \text{ i.e., } \eta_y \geq \eta_x \ast h,
\]

which again gives a lower bound on the expected value of the output. Similarly, (4.9) is rewritten as

\[
R_{yy}(t_1, t_2) \geq \tilde{L} \eta^{h(t)} \left[ R_{xy}(t_1, t_2) \right].
\]

4.4. Min-Plus Stochastic Systems

In the previous section, we established a system theory for min-plus (max-plus) systems with a stochastic input. This is a logical approach, for example, for a network router with a deterministic service curve but a stochastically defined characterization for the incoming traffic. Since the min-plus convolution is commutative, i.e., \((f \ominus g)(t) = (g \ominus f)(t)\), then when dealing with stochastic systems with an input constrained by an arrival curve, we can readily apply the results and theorems of the previous sections by just swapping the role of \( h(t) \) with \( x(t) \). This
will provide us with the theorems required to analyze a system with a stochastic service curve (exact or upper), and an input shaped to have a NC arrival curve. Therefore, we rewrite (4.7) as

\[ \eta_y \leq \eta_h \otimes x. \]  

(4.11)

The correlation equations of (4.9) are rewritten as:

\[ R_{hy}(t_1, t_2) \leq L_2^{\eta_h(t_1) x(t_2)} [R_{hh}(t_1, t_2)] \]  

(4.12)

and

\[ R_{yy}(t_1, t_2) \leq L_1^{\eta_y(t_2) x(t_1)} [R_{hy}(t_1, t_2)]. \]  

(4.13)

Figure 4.2 shows an end-to-end path from the source to the destination, in a network consisting of a wired and a wireless part. The wired network is analyzed by NC. The input \( x(t) \) from the wired network is the traffic that enters the wireless network passing through a stochastically defined wireless channel. Using the theorems presented in this section, in the following section we present a methodology that calculates the corresponding statistics of the output \( y(t) \) (received by the destination). For the uplink path, the above approach is used similarly in reversed order.
4.5. Wireless Channel Modeling, Performance Evaluation and Cross-layer Design

In the previous section, we developed a general framework for modeling min-plus stochastic systems. Since a wireless channel is time-varying, the effective transmission rate available from the link layer perspective is variable and time-varying. Therefore, for a given transmitted traffic (input), different possible channel outputs are observed. For example, Figure 4.3 plots 50 different observed outputs of the plotted traffic passing through a Rayleigh fading channel. A stop-and-wait ARQ protocol is used to compensate for packet errors, so that any point on an instance of the output corresponds to a point on the input with the same amount of traffic. Due to fading, each instance experiences a different PER versus time, thus it is different from the other possible instances of the output.

In the following, we show how the theorems presented earlier are used to calculate the statistics of such outputs. In order to do so, the packet-level statistical service characteristic of the wireless channel, $h(t)$, needs to be appropriately defined. A good example of such a service characterization is presented in 4.5.2. We then apply Equations (4.11) and (4.12) to find the statis-
tics of the output for a given arrival process passing through the wireless channel (or network, using the concatenation property). Therefore, by using the information from the current physical layer behaviour, the statistical characteristics of the output are determined. By using this information, the parameters of the link layer are set accordingly. For example, such information is used to set the rate parameter of the shaper at the transmitter in Section 4.6. This is in addition to the requirements considered by the transmitter to conform to the negotiated QoS policies.

The method is general and comprehensive. \( h(t) \) can be defined to be as elaborate and complex as required. For a cross-layer design, this service curve should be defined in such a way that it accurately translates the statistical properties of the physical layer, which in turn is used to adjust the operation of the link layer or to allow efficient allocation of network resources. In this section, a service curve of this nature is first addressed, based on the second-order BER statistics at the physical layer. Then we introduce a framework that utilizes the theorems presented earlier.
for performance modeling and analysis of wireless channels. The technique is applicable for both uplink and downlink, but is particularly well-suited to the downlink, where the results from the deterministic NC modeling of the previous wired parts of the path can be most easily used for end-to-end QoS support.

4.5.1. System Model

Figure 4.4 shows the overall system model. We are interested in calculating the traffic received by the destination at the link layer. The traffic is first shaped and then sent through a wireless Rayleigh fading channel. We assume that the wireless user is assigned a dedicated bandwidth; therefore, we only consider the effect of Rayleigh fading on packet errors. We assume that packets are retransmitted based on the stop-and-wait ARQ protocol in order to compensate for packet errors. Therefore, it is ensured that a packet is transmitted successfully before the transmission of the next packet is initiated. A similar variability of outputs is observed when no retransmission mechanism is used. However, in this case the delay cannot be calculated from the horizontal distance between the input and output graphs (or curves), which will become clear when calculation of delay is explained later in this section.
4.5.2. Service Curve of the Channel

The framework established earlier is general and can be used with any stochastic server such as a wireless network, where a proper \( h(t) \) is available. Depending on the physical channel behavior, the link layer protocol, the required QoS bounds and the application, there are different ways of modeling the dynamics of the wireless channel.

A reasonable way of defining \( h(t) \) is by direct consideration of the available physical channel models, and to translate it to the packet-level service at the link layer. One practical approach is to characterize the instantaneous service rate of a wireless channel at the link layer by

\[
    r_c(t) = (1 - PER(t))\mu, 
\]

when the packet error rate at any given time \( (PER(t)) \) and the channel transmission capacity \( (\mu) \) measured in packet units at \( \Delta \) time interval \( (\Delta \to 0) \) are known [60]. \( \mu \) can be set to the Shannon capacity by using \( B \log(1 + \text{SNR}) \), when a dedicated bandwidth of size \( B \) is available to the wireless user. This is a reasonable assumption, as the errors are accounted for by \( PER \). Whenever this method presents an overly optimistic bound for the channel capacity, other more elaborate methods for the calculation of the channel capacity may be used. In wireless networks where the users contend to use the wireless channel, \( \mu \) can be set as the total available channel capacity, while \( PER \) is defined so that it includes packet errors due to contention as well. Equation (4.14) describes the instantaneous rate that is seen to be available by the link layer for successful transmissions. This equation holds whether a retransmission mechanism is implemented or not. In this chapter, we assume that packets are retransmitted, based on the stop-and-wait ARQ protocol, in order to compensate for packet errors. With this condition, a point on the input function (representing an amount of cumulative traffic) corresponds with a point on the output function that has the same value.

Since the channel is time-varying, \( PER \) is a random process. The second-order statistics of \( PER \) are all we need to use (4.14). Here we consider a set of widely-used assumptions, in order to focus on describing the methodology. Alternatives to these assumptions are discussed using more sophisticated methods in Section 4.7.
The service curve of the channel is the cumulative service, which is the integral (or summation for sampled or discrete representations) of (4.14):

\[
h(t) = \int_0^t c(x)dx = \int_0^t (1 - \text{PER}(\tau))\mu d\tau.
\]

(4.15)

It is interesting to note that the above equation results in a stochastic rate-latency server (of the form \( \beta_{r,T} = (t - T)r \)), with a stochastic latency \( T = \int_0^t \text{PER}(\tau)d\tau \) and rate \( r = \mu \).

**Calculation of Bit Error Rate (BER):** Without loss of generality, we assume use of a QPSK modulation scheme, and no channel coding. For a \((\pi/4)\)-QPSK scheme, BER is a function of SNR, denoted by [61]:

\[
\text{BER}(t) = Q(\sqrt{2\text{SNR}(t)}),
\]

(4.16)

\( Q(.) \) is the scaled complementary error function.

In the following, the distribution of SNR for two cases of Additive White Gaussian Noise (AWGN) and Rayleigh fading channels are discussed.

**Case I - AWGN Channels:** For an AWGN channel, the SNR is a stochastic process whose distribution depends on and is directly derived from the distribution of the additive noise, which is zero-mean Gaussian.

**Case II - Rayleigh Fading Channels:** The instantaneous SNR per bit, \( \gamma \), for an uncorrelated time-varying Rayleigh fading channel with Alamouti transmit diversity has a central chi-square distribution with two degrees of freedom. The Alamouti transmit diversity scheme is one of the methods proposed for IEEE 802.16 (WiMAX) to achieve 2-way diversity without adding an extra antenna, as described in the IEEE 802.16abc-01/53 specification. The probability density function (pdf) of such a process is [62]
Chapter 4. A Stochastic Min-plus System Theory Approach to Performance Modeling of Wireless Networks

\[ p_y(r) = \frac{2}{E/N_0} \exp(-r^2N_0/E), \quad r \geq 0, \]  

where \( E\{y\} = E/2N_0 \) is the average SNR per bit.

For the case of a fully correlated Rayleigh fading channel with Alamouti transmit diversity, the pdf is in the form of a central chi-square distribution with four degrees of freedom, denoted by

\[ p_y(r) = \frac{r^2}{(E/2N_0)^2} \exp(-r^2N_0/E), \quad r \geq 0. \]  

In both cases, similar assumptions for the modulation scheme and packet error allocation are considered. Thus, given SNR, the distribution of BER is calculated by using (4.16).

**Calculation of Packet Error Rate (PER):** Without loss of generality, independent bit errors are assumed which leads to uniform distributions of bit errors in each packet. For completeness, in Section 4.7., an elaborate method of PER calculation is discussed. Given the above assumption, PER is defined by

\[ \text{PER}(t) = 1 - (1 - \text{BER}(t))^n, \]  

where \( n \) is the number of bits in the packet. We assume that the packet size is set so that BER stays constant over the packet transmission time. This condition can be achieved by breaking up the packets into smaller fragments over which BER can be assumed constant, and then sending each fragment as a separate packet. Packet segmentation is a basic function of the link layer \[63][64]. For a more conservative calculation, the maximum BER that occurs during the packet transmission time can be considered instead.

Since we have the statistics of BER, using (4.19) the statistics of PER are calculated. The probability density function is \( p_{\text{PER}} = \frac{d}{da}F_{\text{PER}}(a) \), where \( F_{\text{PER}}(a) = P\{\text{PER} \leq a\} \). It can be
shown that $F_{\text{PER}}(a) = F_{\text{B}}(1 - \sqrt[3]{1 - a}) + (1 - F_{\text{B}}(1 + \sqrt[3]{1 - a}))\chi(n)$, where $\chi(n)$ is 1 if $n$ is even, and is 0 otherwise [59][65].

In this chapter, only the mean and autocorrelation of the processes are needed, and can be calculated either directly or from the probability density function when available. For completeness, the distributions have been presented here briefly. Direct derivations of the mean and autocorrelation functions of PER and then the process $r_c(t)$ are performed by cascaded calculations of the statistics of BER, PER, $r_c$, from (4.16), (4.19), and (4.14), respectively.

4.5.3. Methodology

In brief, the following steps are taken:

S1) In order to use (4.14), the second-order statistics of PER need to be derived. Without loss of generality, we use the following widely used two sub-steps (see Section 4.7 for an elaborated alternative that is more realistic). (a) From the physical layer model, the second-order statistics of bit error rate (BER) are given. The statistics of BER are obviously dependent on the modulation scheme and any error correcting code (channel coding) used by the transceivers, and is a function of the signal to noise ratio (SNR). (b) BER is used to find the second-order statistics of the PER. This can be done by assuming uniform distributions of bit errors, or by using other methods in the presence of error dependency and channel coding (as will be explained in Section 4.7).

S2) Then (4.14) is used to find the statistics of the channel service rate. Simply, we have $E\{r_c\} = (1 - E\{\text{PER}\})\mu$, and $R_{r_c}(t_1, t_2) = (1 - 2E[\text{PER}] + R_{\text{PER}}(t_1, t_2))\mu^2$. From (4.15), the NC compatible service curve of the channel (i.e., cumulative and thus non-decreasing) is calculated.

The above steps provide us with the statistics of a compatible service curve for the channel. The following two steps are performed to calculate the delay bounds:

S3) Equations (4.11) and (4.12) are used to obtain the second-order statistical properties of the output. These equations are applicable when $h(t)$ is an exact or upper service curve. $h(t)$ is
an exact service curve, when the wireless channel is dedicated to the user, such as the one defined in Section 4.5.2. When the wireless channel is shared between the users in a random multiple access (such as by Aloha), $h(t)$ is an upper service curve. Finally, when the wireless channel is shared between the users by some QoS-based scheduling technique (such as by a WFQ), $h(t)$ is a lower service curve.

$S4)$ Finally, NC filtering theory is used to calculate the delay bound (or other QoS bounds) for a given input.

The mean function gives us the expected value of the output, and the autocorrelation function gives us sufficient statistical information about its variation. The approach we choose to follow here is based on the standard deviation function. At any given time $t$, the variance of the output process is $\text{Var}(y)(t) = R_{yy}(t, t) - \eta_y^2(t)$, and the standard deviation is $\text{Std}(y)(t) = \sqrt{\text{Var}(y)(t)}$. The output will vary around its mean within a bracket of $[\eta_y(t) - \kappa \cdot \text{Std}(y), \eta_y(t) + \kappa \cdot \text{Std}(y)]$ with a high probability, where $\kappa$ is a constant. The lower and upper values of this interval are called the best and worst output curves, respectively. Since we use these concepts to find the standard deviation of the delay, we set $\kappa = 1$. For example, for a normal distribution, the process is within one standard deviation around the mean with a probability of 68%, and within two standard deviations around the mean with a probability of 95%. Therefore, we can use these two bounds as the worst and best possible outputs.

**Calculation of Delay Bounds:** Having the output curve is all that is required for us to utilize NC theorems to find the delay bounds. As we have assumed that in order to compensate for packet errors, a stop-and-wait ARQ protocol is implemented, any two points of the input and the output that have the same value represent the same amount of traffic. Then the delay that a packet transmitted at time $t$ encounters is given by

$$d(t) = \inf\{\Delta \geq 0 : x(t) \leq y(t + \Delta)\} , \quad (4.20)$$

which graphically is the horizontal distance between the two graphs. The maximum delay is $d_{\max} = \sup\{d(t)\}$. However, here this delay is a stochastic process, since $y(t)$ is stochastic.
From the output curves that we have calculated, we can define the second-order statistics of the delay. Figure 4.5 illustrates these calculations.

First, the mean delay is equal to \( E\{d(t)\} = \inf\{\Delta \geq 0: x(t) \leq E\{y(t + \Delta)\}\} \), therefore, the maximum value of the mean delay throughout the session, \( d_{\text{mean}} \), is \( d_{\text{mean}} \leq \sup\{E\{d(t)\}\} \).

Second, the standard deviation of the delay is equal to

\[
\text{Std}\{d(t)\} = \inf\{\Delta \geq 0: E\{y(t)\} \leq (E\{y\} - \text{Std}\{y\})(t + \Delta)\}.
\]

Therefore, the maximum value of the standard deviation of the delay throughout the session, \( d_{\text{Std}} \), is \( d_{\text{Std}} \leq \sup\{\text{Std}\{d(t)\}\} \).

From these values, we can readily define the statistical worst-case values for the delay, \( D_{\max} \). For example, for a Gaussian distribution, \( Pr(D_{\max}^{\epsilon_1} \leq d_{\text{mean}} + d_{\text{Std}}) \geq 1 - \epsilon_1 \), where \( \epsilon_1 = 0.32 \), and \( Pr(D_{\max}^{\epsilon_2} \leq d_{\text{mean}} + 2d_{\text{Std}}) \geq 1 - \epsilon_2 \), where \( \epsilon_2 = 0.05 \).

The numerical example presented in the next section elaborates upon this procedure.
4.6. Numerical Results

In the previous section, we detailed the methodology through four steps. In this section, we use a numerical example for a Rayleigh fading channel to compare the analytical results with simulation results, using MATLAB®\(^2\). Following the four steps of the methodology, Equations (4.11) and (4.12) are used to first obtain the second-order statistical properties of the output. Having the statistics of the output, the delay bounds are calculated. The nature and values of the parameters, as well as the assumptions used in the numerical example are as follows. The traffic that enters the channel is shaped by a leaky-bucket controller to conform with an arrival curve of the affine function \(at + b\), where \(a\) is the rate and \(b\) is the bucket size (so-called burst tolerance). Figure 4.6 illustrates this operation with an example. The parameters of the shaper have to be defined according to the negotiated QoS policies. The actual rate of the traffic source can be less, equal or larger than that of the shaper. For better presentation of the results and since the arrival curve is a good QoS benchmark, we have set the value so that the shaped traffic is close to the shaping envelope. An instance of the traffic from the source, shaped traffic, and their cumulative representations are shown in Figure 4.6, subplots \(a\), \(b\) and \(c\), respectively.

The channel is assumed to be uncorrelated Rayleigh fading with Alamouti transmit diversity with average SNR \(E\{\gamma\}\), and rate \(\mu\), as discussed in Section 4.5. This is a diversity method used in WiMAX. A \((\pi/4)\)-QPSK modulation and no channel coding are assumed. The values of the parameters used in the example are summarized in Table 4.1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>notation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean inter-arrival times</td>
<td>MIntArr</td>
<td>0.7 (ms)</td>
</tr>
<tr>
<td>Mean packet size</td>
<td>MPkSz</td>
<td>3 (KB)</td>
</tr>
<tr>
<td>Shaper rate</td>
<td>(a)</td>
<td>(\text{MPkSz/MIntArr (MB/s)})</td>
</tr>
<tr>
<td>Shaper burst tolerance</td>
<td>(b)</td>
<td>5*MPkSz</td>
</tr>
<tr>
<td>Average SNR</td>
<td>(E{\gamma})</td>
<td>10 (dB)</td>
</tr>
<tr>
<td>Channel rate</td>
<td>(\mu)</td>
<td>([0.5 1.2]*a)</td>
</tr>
</tbody>
</table>

\(^2\)MATLAB is a registered trademark of The MathWorks, Inc.
Figure 4.6  An example of (a) the traffic delivered to the link layer, (b) shaped traffic ready to be transmitted through the channel, and (c) cumulative representations of traffic in a and b.

The simulation methodology is as follows. A source generates a traffic stream with exponential inter-arrival times and packet sizes. This traffic is then shaped based on a leaky-bucket shaper ($\gamma_{a,b}(t) = at + b, t > 0$ and 0 otherwise); the values of its parameters are given in Table 4.1. A packet is segmented if its size is larger than $b$. The traffic source is set so that the
shaped traffic, $x(t)$, is as close as possible to the arrival curve. This shaped traffic flow enters the channel. A random generator is used to generate a sequence of random numbers based on the distribution of the SNR. Each random number is associated with a packet that is transmitted at that time. This set of numbers is used in turn to calculate the BER and then the channel rate. In order to have a similar setting as the numerical method, it is assumed that the channel provides a service curve of the form (4.14). Based on this service curve, the channel acts as a stochastic server for the input traffic flow. Finally, a set of 50 instances of the output flow is used to calculate its second-order statistics. For the figures that present the results with a 95% confidence interval [66], the calculations were iterated 30 times.

Figure 4.7 shows the three-dimensional plots of the autocorrelation of the service and the output. Since the underlying process is a WSS process, these are accordingly WSS. The $x$ axis designates the transmission time and the $y$ axis represents 50 different instances of observations, including its mirror towards the negative side, with a collective span of $[-49, 50]$.

Figure 4.8-a shows the traffic (arrival) passing through the channel, and a set of 50 instances of possible service curves that occurred due to the time-varying fading channel. Figure 4.8-b shows the mean output calculated from the analytical approach as well as the mean of the actual outputs from the simulation.
Figure 4.7 3D plots of the autocorrelation of the service (a) and the output (b). The $x$ axis designates the transmission time and the $y$ axis represents 50 different instances of observations (with its mirror towards -50).
Figure 4.8  (a) Traffic and 50 instances of channel service curves. (b) Comparison of analytical and simulation results for mean outputs.
Knowing the input and output, the delay bound can be easily calculated. First, we use the graph of the mean output to calculate mean delay. Mean delay is the maximum horizontal distance between the input and the mean output. Similarly, the standard deviation is used to plot the best and worst output curves, as explained in detail in Section 4.5.3. Following the same procedure, the worst delay is calculated, which is the maximum horizontal distance between the input and the worst output. Figure 4.9 shows the mean, worst and best outputs and illustrates the above procedure for graphically calculating the mean and worst delays.

Obviously, the shaper rate (value of parameter $a$) relative to $\mu$ is an important factor in the value of the observed delay. Figure 4.10 plots mean and worst delays for different values of the ratio of the channel and shaper rates ($\mu/a$), within a wide range of $[0.5, 1.2]$. The figure compares the analytic and simulation values with a 95% confidence interval, considering 30 iterations per each value of the above ratio, with 50 instances of channel service curves.

Figures 4.11 and 4.12 demonstrate the plots of delay when the other two effective parameters, $b$ and $E\{\gamma\}$, are changing. Figure 4.11 shows the delay versus the $b$ parameter of the shaper,
Figure 4.10  Mean and worst delays for different values of $\mu/a$ with a 95% confidence interval (30 iterations of 50 instances each).

and Figure 4.12 plots the delay for a wide range of values of average SNR ($E\{\gamma\}$ in dB).

Therefore, the link layer shapes the traffic with a rate that not only conforms to the negotiated QoS policy, but also considers the wireless channel condition. Hence, the value of the parameter $a$ is reconsidered and adjusted to implement the above cross-layer design. Depending on how often the second-order statistics of the physical layer are acquired, the value of $a$ can be adjusted accordingly. This procedure is the so-called impedance matching for the channel condition [67], in addition to traffic and congestion conditions, and QoS policies. Since in our scenario the channel is under a wide-sense stationary condition, only $E\{\gamma\}$ needs to be passed up from the physical layer. As discussed earlier, this is a practical assumption for a fading channel, which obeys a chi-square distribution.

The figures show a close match of the analytical and simulation results. Our analysis also finds similar results in the case of a fully correlated time-varying Rayleigh fading channel with Alamouti transmit diversity, which was discussed in Section 4.5.
Chapter 4. A Stochastic Min-plus System Theory Approach to Performance Modeling of Wireless Networks

Figure 4.11 (a) Delay versus $b$ (shapers burst tolerance), when $E(Y)$ and $\mu/a$ are 10 dB and 0.8, respectively, with a 95% confidence interval (30 iterations of 50 instances each).

Figure 4.12 Delay versus average SNR ($E(Y)$) in dB when $\mu/a$ is 0.8, with a 95% confidence interval (30 iterations of 50 instances each).
4.7. Discussion

In this section, we first discuss practical aspects and advantages and limitations of the proposed method. We then elaborate upon applying the proposed method to applications with different settings. Thus far, we have assumed a QPSK modulation scheme, no channel coding, and uniform distribution of errors. In this section, we discuss how to incorporate channel coding and/or other modulation schemes into the model. In addition, an elaborate method for the derivation of PER, considering bit error dependency, is given.

Some topics of interest that further expand the scope of the proposed methodology are also discussed. A short note on how to consider mobility is presented; analysis of systems with both stochastic traffic and service curve is addressed; and the statistical service guarantee in NC in relation to the proposed method is discussed. Finally, slope domain representations of the fundamental theorems are derived.

**Practical Considerations in Using the Method in Wireless Data Networking:** In order to determine the scope of application of the discussed method, we first address when the underlying Theorems 4.3 and 4.5 are valid. Three aspects, linearity, shift-invariance and stationary conditions need to be addressed.

Linearity: Theorems 4.3 and 4.5 can be used only when the min-plus system is min-plus linear. As explained earlier, in order for a DEDS to be min-plus linear, there should not be a number of servers available for the input to choose from (so-called concurrency) [7]. In practice, this condition holds for most queuing nodes or data networks, when the routing path is pre-set by the transmission connection.

Shift-invariance: In order to use the convolution operator in place of the system operator $L(.)$ in Theorems 4.3 and 4.5, the min-plus system should be shift-invariant. This condition applies when a shift in input creates a similar shift in output. In practice, QoS-aware networks are usually designed with the aim that this property holds. However, this requirement does not exclude stochastic systems. Usually, in stochastic settings the wide-sense stationary (WSS) assumption replaces the shift-invariant assumption.
Stationary assumption: Theorems 4.3 and 4.5 are valid regardless of the stationary condition. However, in practice, in order to define a meaningful stochastic service curve for the channel, we need to consider a WSS condition. This is a valid and widely used assumption, especially for Rayleigh fading channels ([68], pp. 65).

It goes without saying that the applicability of the method relies on the requirement that the wireless network can be reliably modeled by a stochastic service rate whose second-order statistics can be calculated.

**Detailed PER Calculations:** So far in this chapter we have assumed that the bit errors in a packet occur independently and that no channel coding is implemented. We did so to be able to clearly explain the methodology without becoming sidetracked by irrelevant details. However, channel coding plays a significant role in wireless communication [69]. For example, deployment of a convolutional code can lower BER by two orders of magnitude ([70], pp. 260). In this section, however, we revise our assumptions and consider error dependency in the presence of convolutional channel coding, to demonstrate that the same methodology applies.

As a replacement for (4.19), Equation (4.22), which was introduced in [71], is used. Although these equations look similar, (4.22) not only considers error dependency but also takes into account convolutional coding as an error correction code.

The procedure for PER calculations remains almost unchanged. The statistics of SNR are first calculated. Then instead of finding BER, a new function called EER (error event rate) is used, from which PER is calculated. We take the following steps:

1) \( EER(SNR) \) is defined by \( EER(SNR) = A_{d_{free}} \exp(R \cdot SNR \cdot d_{free}) \), where \( A_{d_{free}} \) is the number of paths of length \( d_{free} \) in the code trellis and \( R \) is the code rate.

2) From \( EER(SNR) \) the value of another function \( \lambda_s(z) \) is defined by

\[
\lambda_s(z) = \frac{EER(z)}{1 - \left[ (u + 1) + \frac{1}{n_c(z/2 - \sqrt{2}\sqrt{2}r_c + r_c)} \right] EER(z)}
\] (4.21)
where \( r_c = \frac{k_c}{n_c} \) is the code rate and \( \nu \) is the constraint length of the \((n_c, k_c, \nu)\) convolutional code. It is shown in [71] that \( \lambda_S(SNR) \) is the mean of a geometric distribution characterizing errorless period lengths.

3) Finally, the PER is calculated as follows:

\[
PER(t) = 1 - (1 - \lambda_S(SNR(t)))^n. \tag{4.22}
\]

It is interesting to note that the above formula appears similar to (4.19), with BER replaced with a new function, \( \lambda_S(SNR) \).

Under similar considerations as applied to our previous numerical and simulation settings, we derive the mean and worst delays using (4.22). Based on [72] for a UMTS system, a convolutional code with rate 1/2 and \( \nu = 9 \) is assumed. For this code \( d_{free} = 12 \). We assume that the value of \( n_c \) is equal to the mean packet length and that \( A_{d_{free}} = 1 \).

Figure 4.13 shows the analytical and simulation results for the delay. The observed delays are worse due to the bit error dependency (as compared with those in Figure 4.10), but the methodology is still applicable.

**Effect of the Modulation Scheme:** Until now in this chapter, we have assumed a QPSK modulation scheme. For a different modulation scheme, only the equation describing BER as a function of SNR needs to be changed. For example, for the DBPSK modulation scheme, (4.16) is replaced with:

\[
BER(t) = 0.5 \exp(-SNR(t)).
\]

**Systems with Both Stochastic Input and Service:** When the service and input are both stochastically defined, the theorems are modified as follows, and can be used accordingly. In the following, only the results for systems with exact or upper service curves are presented.

A) The mean of the output is

\[
\dot{\eta}_y \leq \dot{\eta}_x \otimes \dot{\eta}_h. \tag{4.23}
\]

Proof: Similar to the proof of Theorem 4.3, we have
Figure 4.13  Delay versus different values of shaping rate with the consideration of convolutional channel codes and bit error dependency.

\[ E\{L[x(t)]\} = E\{x(t) \otimes h(t)\} \]
\[ = E\{\inf_{0 \leq s \leq t} (x(t-s) + h(s))\} \]
\[ \leq \inf_{0 \leq s \leq t}(E\{x(t-s)\} + E\{h(s)\}) \]
\[ \leq E\{x(t)\} \otimes E\{h(t)\} \]

B) Similarly, the autocorrelation function can be calculated from the following:

\[ R_{yy}(t_1, t_2) \leq L_{1}^{R_{x,y}(t_1, t_2)}[R_{xy}(t_1, t_2)]. \] (4.24)

Note that this equation is slightly different from (4.12), as we need to calculate \( R_{by} \) as well as \( R_{xy} \).
Proof:

\[ y(t) = L_{h(t)}[x(t)] \]
\[ y(t)x(t_1) = \inf_s (x(t-s) + h(s))x(t_1) \]
\[ E\{y(t)x(t_1)\} = E\{\inf_s (x(t-s) + h(s))x(t_1)\} \]
\[ \leq \inf_s (E\{x(t-s)x(t_1)\} + E\{h(s)x(t_1)\}) \]
\[ \leq \inf_s (R_{xx}(t_1, t-s) + R_{xh}(t_1, s)) \]

\[ R_{xy}(t_1, t) \leq R_{xx}(t_1, t) \otimes R_{xh}(t_1, t) \]
\[ R_{xy}(t_1, t_2) \leq L_2^{R_{xh}(t_1, t_2)} [R_{xx}(t_1, t_2)] \]

and

\[ y(t) = L_{h(t)}[x(t)] \]
\[ y(t)h(t_1) = \inf_s (x(t-s) + h(s))h(t_1) \]
\[ E\{y(t)h(t_1)\} = E\{\inf_s (x(t-s) + h(s))h(t_1)\} \]
\[ \leq \inf_s (E\{x(t-s)h(t_1)\} + E\{h(s)h(t_1)\}) \]
\[ \leq \inf_s (R_{hx}(t_1, t-s) + R_{hh}(t_1, s)) \]

\[ R_{hy}(t_1, t) \leq R_{hx}(t_1, t) \otimes R_{hh}(t_1, t) \]
\[ R_{hy}(t_1, t_2) \leq L_2^{R_{hh}(t_1, t_2)} [R_{hx}(t_1, t_2)] \]

Finally, \( R_{yy}(t_1, t_2) \) is determined as follows:

\[ y(t) = L_{h(t)}[x(t)] \]
\[ y(t)y(t_2) = \inf_s (x(t-s) + h(s))y(t_2) \]
\[ E\{y(t)y(t_2)\} = E\{\inf_s (x(t-s) + h(s))y(t_2)\} \]
\[ \leq \inf_s (E\{x(t-s)y(t_2)\} + E\{h(s)y(t_2)\}) \]
\[ \leq \inf_s (R_{xy}(t-s, t_2) + R_{hy}(s, t_2)) \]

\[ R_{yy}(t_1, t_2) \leq R_{xy}(t_1, t_2) \otimes R_{hy}(t_1, t_2) \]
\[ R_{yy}(t_1, t_2) \leq L_1^{R_{hy}(t_1, t_2)} [R_{xy}(t_1, t_2)] \]

\[ \square \]
Mobility: The proposed method is practical, regardless of whether the transmitter and/or receiver are fixed or mobile. Since all of the equations in this chapter are time-dependent, the location of the node is already incorporated. In this approach, both large-scale path loss and shadowing, as well as small-scale fading are considered. Path loss is mainly dependent on the receiver-transmitter distance from which the average SNR (expectation of SNR) is calculated. To account for mobility, this value in Equation (4.17) (or (4.18)) would be time varying and needs to be reconsidered at any given time.

Statistical Guarantees: Our method fits in and can be used with statistical NC [17][24], in which the QoS evaluations are based on a statistical exact (or upper) service curve $h^e(t)$, satisfying

$$Pr\{y(t)= x(t) \otimes h^e(t)\} \geq 1 - \varepsilon$$

(4.25)

since such a service curve is derived for the channel. All QoS bound theorems for delay, backlog and output bound of NC are valid as well under the above definition, and can be used to calculate QoS metrics of interest [17].

End-to-End QoS Evaluations: The method presented in this chapter can be used in end-to-end QoS evaluations. The concatenation property applies to a combination of stochastic and deterministic NC elements; i.e., assuming that the service curve of the rest of the network is $h_1(t)$ and the service characterization of the channel is $h(t)$, then the end-to-end service curve is defined by $h^e(t) = h_1(t) \otimes h(t)$. Therefore, Equations (4.11) to (4.11) are applied to this new service curve.

Slope Domain Representations - Load Spectrum: In system theory, the Fourier spectrum is used for spectral representations of random signals. The power spectrum of a random process is the Fourier transform of its autocorrelation function: $PS(\omega) = \int_{-\infty}^{\infty} R(\tau) e^{-j\omega \tau} d\tau$ ([55], pp. 416). Then, the power spectrum function of the output is
In Chapter 3, we presented slope domain modeling of networks, which serves in regard to min-plus systems as the Fourier domain to traditional system theory. In this section, we use the slope transform defined by (3.1) to find the slope domain representations of (4.23) and (4.24).

Similar to power spectrum, we introduce the slope transform of the autocorrelation functions as load spectrum, using the notation $S_{yy}(\alpha) = S\{R_{yy}(\tau)\}(\alpha)$. An immediate advantage of using slope domain analysis is that all the convolution operators are translated to algebraic additions. Also, the load spectrum demonstrates how traffic load is distributed over different rates (slopes). The slope transform of (4.12) leads to

\begin{align*}
S_{hh}(\alpha) & \leq S_{hh}(\alpha) + \eta_h \chi(\alpha/\eta_h) \\
S_{yy}(\alpha) & \leq S_{yy}(\alpha) + \eta_h \chi(\alpha/\eta_h),
\end{align*}

which finally results in:

\begin{equation}
S_{yy}(\alpha) \leq S_{hh}(\alpha) + \eta_h \chi(\alpha/\eta_h) + \eta_p \chi(\alpha/\eta_p).
\end{equation}

The autocorrelation of the output is then calculated by the inverse slope transform of $S_{yy}(\alpha)$.

### 4.8. Summary

In this chapter, we have introduced a stochastic min-plus methodology that not only serves as a general extension to NC for efficient modeling and performance evaluations of wireless networks, but also as a method for cross-layer design at the transmitter. This is an effective method when the second-order statistics of the service can be calculated. We have presented a simple-to-use framework from which the second-order statistics of the traffic passing through a channel are calculated, using the second-order statistics of the physical layer. This information is used to calculate delay bounds in order to verify whether the desired delay as the QoS metric of interest is met.
Chapter 5. Optimal Leaky-Bucket Shaper and Resource Allocation for the Transmission of Multimedia over Wireless Fading Channels

In the previous chapter, the focus of our analysis was on performance evaluations. In this chapter, we concentrate on configuring system parameters so that the desired QoS metrics are achieved. We derive the parameters of an optimal leaky-bucket shaper and/or the required channel bandwidth for the transmission of constrained traffic over a wireless fading channel. The analysis is based on the stochastic min-plus system theory for the derivation of second-order statistics, discussed in Chapter 4, as a wireless fading channel is a stochastic rate-latency server. The shaper is optimal in the sense that it minimizes the end-to-end delay and buffer requirements under the constraints dictated by the lower layer characteristics, in particular the wireless fading channel. In the case of the resource allocation application, the method presents the minimum required channel rate that results in favourable service quality. Our model is used to plot the analytic delay and buffer sizes in comparison to simulation results for various values of shaper parameters and the available channel rates, from which the optimal or sub-optimal solution is calculated.
5.1. Introduction

The traffic entering a network may be shaped for a number of reasons. It is shaped at the source or a network edge to conform to a predefined service level agreement to make it eligible for receiving the negotiated QoS. Traffic shaping is also used to ensure a fair and stable utilization of the network, even for the best-effort service. Finally, it is implemented as a rate-control mechanism to smooth a bursty traffic stream.

In a wireless network, the bottleneck is usually the time-varying wireless channel. Therefore, the available effective rate should be taken into consideration. In this chapter, we show how to derive the parameters of a leaky-bucket shaper and/or the bandwidth assignment, in order to minimize end-to-end delay and the buffer requirements. The minimization is done under constraints dictated by the physical layer characteristics such as SNR statistics and channel rate, or by the traffic requirements, for example, a permitted traffic burstiness. The buffer requirement is dependent on the available buffer sizes of the involved layers for a reliable transmission especially when an ARQ mechanism for loss-free transmissions is deployed.

The analytical model is based on the stochastic min-plus system theory that was thoroughly discussed in Chapter 4. The theory explains the input/output relationship of a stochastic min-plus system. It is used to derive the second-order statistics of the output, when input and/or service are stochastically characterized. As explained before, a wireless fading channel is modeled by a stochastic rate-latency server whose second-order statistics are easily calculated from the knowledge of the physical layer characteristics. Therefore, we use this method to find the statistics of the traffic (the output) that passes through a wireless channel; i.e., the mean and autocorrelation of the traffic at the output of the channel. Then by using NC theorems [9], the optimal values for the parameters of the shaper and/or the requisite channel assignment are calculated.

The outline of the chapter is as follows. In the next section, an overview on the theory of optimal traffic shaping is presented. The system model, assumptions and descriptions of the underlying concepts and building blocks are addressed in Section 5.3. The theory of shaping for transmission over a wireless channel and the methodology for calculating the parameters of the
optimal shaper are presented in Section 5.4. The numerical and simulation results are discussed in Section 5.5. We present the results when either of the channel rate or the shaper rate are selectable. For each case, two scenarios of bit error dependency and independency in packets are studied. Finally, a summary of the chapter is presented in Section 5.6.

5.2. Optimal Leaky-Bucket Shaper

We use the widely accepted min-plus system theory or NC [9] representations for the traffic and service. A system is a representation of a network or a network element (link, node, or sub-network) that provides a service defined by $h(t)$. The incoming traffic (input) is represented by $x(t)$, and $y(t)$ represents the outgoing traffic (output received by the end element), both in units of bytes. The traffic and service are characterized by the arrival and service curves, respectively. The traffic (input or output), arrival curve and service curve are represented by non-decreasing functions. For a cumulative traffic $x(t)$ and an arrival curve $\zeta(t)$, it is guaranteed that $x(t + \tau) - x(t) \leq \zeta(\tau)$, $\forall t, \tau$. Similarly, a system with lower service curve $h(t)$ guarantees that the cumulative output traffic $y(t)$ obeys $y(t + \tau) - y(t) \leq h(t)$, $\forall t, \tau$.

The output of a system with input $x(t)$ and exact service curve $h(t)$ is expressed by $y(t) = L(x(t))$, where $L \overset{h(t)}{x(t)} = x(t) \otimes h(t)$. The min-plus convolution operator is defined by $x(t) \otimes h(t) = \inf_{0 \leq s \leq t} (x(t-s) + h(s))$.

**Leaky-Bucket Shaper:** A leaky-bucket server is a well-known and widely used shaper, with two parameters: $a$ the average rate and $b$ the burst tolerance. Any traffic passing through such a shaper will have an arrival curve of $\gamma_{a,b}(t)$.

The Traffic Specification (TSPEC) is an extension of the leaky-bucket shaper, with an arrival curve of $(pt + M, rt + b)$, where $M \geq b$ and $p \geq r$. Without loss of generality, here we limit our analysis to arrival curves of $at + b$.

**Rate-Latency Server:** The rate-latency server $\beta_{\mu, T} = \mu(t - T)^+$ is a widely used model
describing a service that guarantees a service rate $\mu$ and service latency $T$. For example, the
WFQ (weighted fair queuing) scheduler is a rate-latency server. As explained in this chapter, a
wireless fading channel can be modeled by a stochastic rate-latency server.

**Calculation of Delay and Buffer Requirement:** There are two analytical methods for
delay and buffer calculations, denoted here by A and B. In method A, the input and output
functions are used. The delay that the data transmitted at time $t$ encounters is given by

$$d(t) = \inf\{\Delta: \Delta \geq 0 \text{ and } x(t) \leq y(t + \Delta)\},$$

and the virtual backlog at time $t$ inside the system is

$$B(t) = x(t) - y(t).$$

Then the maximum delay and required buffer are, respectively

$$d_{\max} = k(x, y) = \sup_{t \geq 0}(d(t)),$$

$$B_{\max} = v(x, y) = \sup_{t \geq 0}(B(t)).$$

The maximum delay $d_m$ is the maximum horizontal distance, and the maximum required
buffer $B_m$ is the maximum vertical distance, between the transmitted (input) and the received
(output) traffic.

Method B gives a looser bound, but is more convenient. It is based on finding the
maximum delay and backlog from the arrival and service curves. When the arrival and service
curves tightly represent the input and service, method B gives a good estimate for the delay and
buffer size [9]. It is proven in [9] that the delay bound for a system with a service curve of $h(t)$
and input traffic with an arrival curve of $\xi(t)$ is

$$k(\xi, h) = \sup_{s \geq 0}(\inf\{T: T > 0 \text{ and } \xi(s) \leq h(s + T)\}),$$

which graphically is the maximum horizontal distance between the graphs of $h(t)$ and $\xi(t)$. The
Figure 5.1 Delay and buffer calculations. Method A: using the input and output traffic (a). Method B: using the arrival and service curves (b).

The maximum buffer size is

\[ v(\xi, h) = \sup \{ \xi(t) - h(t) \}, \quad (5.6) \]

which graphically is the maximum vertical distance between the graphs of \( h(t) \) and \( \xi(t) \). In this chapter we call this method B. Figures 5.1-a and 5.1-b graphically depict these concepts.

Optimal Traffic Shaping: An optimal leaky-bucket shaper is the one that minimizes the maximum delay and buffer requirement:

\[ \minimize(d_{\text{max}}, B_{\text{max}}) \quad (5.7) \]

This minimization is subject to the constraint dictated by the QoS policy (affecting the
selection region for the shaper parameters) and the physical layer characteristics such as SNR (affecting the service curve).

In this section, we briefly derive the optimal parameters using method B for a traffic shaped by a leaky-bucket shaper passing through a rate-latency server. This procedure gives us insight into the mechanism involved and the effect of the parameters in the selection of the optimal shaper, which is not obvious from method A. However, method A is preferred and is used in the numerical results of Section 5.5. This concept is explained in Section 5.4.

The analysis depends on the relative values of $a$ and $\mu$. When $a \leq \mu$, then it can easily be shown that the maximum delay and required buffer are $d_{\text{max}} = T + b/\mu$ and $B_{\text{max}} = b + a \cdot T$, respectively, as shown in Figure 5.1-b. Therefore, when $\mu$ is given, in order to satisfy a desired maximum delay and required buffer, the shapers parameters need to be calculated as follows:

$$b = (d_{\text{max}} - T) \mu \text{ and } a = (B_{\text{max}} - b)/T.$$  \hspace{1cm} (5.8)

If the QoS policy indicates a shaping of $a_p t + b_p$, then we should choose $\min(a, a_p)$ and $\min(b, b_p)$.

Figure 5.2 shows the effect of the variation of $a$, $b$, $T$ and $\mu$. Note that the value of the
delay depends mainly on the value of $T$, and also on $b$ and $\mu$. However, it is independent of $a$. Similarly, the value of the required buffer size depends mainly on the value of $b$, and also on $T$ and $a$. However, it is independent of $\mu$.

When $a > \mu$, the arrival and service curves diverge. Therefore, the maximum delay depends on the length of the transmission session and increases accordingly. Unless the session is short, this may increase without bound, when $t \to \infty$.

### 5.3. System Model and Building Blocks

The system model used in this chapter is similar to that used in Chapter 4, as shown in Figure 4.4, which was explained in Section 4.5.1. For completeness, we very briefly explain the model and assumptions in this section. The packet-level transmission over a wireless link is modeled as a system with shaped input $x(t)$, a stochastic server with a service curve $h(t)$, and output $y(t)$, which is received at the link layer at the other end of the channel. The traffic is shaped with a leaky-bucket shaper and sent to a buffer for transmission over a Rayleigh fading channel with Alamouti diversity. We assume the transmitter will be acknowledged by the next available slot, whether a retransmission is required or not.

The function that is used here to practically characterize the instantaneous service rate of a wireless channel is $r_c(t) = (1 - \text{PER}(t))\mu$, given by (4.14). Thus, the service curve of the channel is $h(t) = \int_0^t r_c(t) dt = \int_0^t (1 - \text{PER}(t))\mu dt$. These functions were explained in Section 4.5.2. The instantaneous SNR per bit, $\gamma$, is given by (4.17).

In order to derive the $\text{PER}$, two different scenarios are considered. In order to clearly explain the methodology without becoming sidetracked by irrelevant details, we first consider independent bit errors, and then the practical dependent distribution of bit error in each packet is considered.
**Independent Bit Error Scenario**: By assuming a $(\pi/4)$-QPSK modulation, independent distribution of bit errors in each $n$-bit packet and no channel coding, we have 
\[ BER(t) = Q(\sqrt{2SNR(t)}) \] (from (4.16)) and 
\[ PER(t) = 1 - (1 - BER(t))^n \] (from (4.19)).

**Dependent Bit Error Scenario**: To consider a practical scenario, we also assume dependent bit errors and a convolutional channel coding. This scenario was explained in pp. 84-85 of Chapter 4. From (4.22), the PER is calculated from 
\[ PER(t) = 1 - (1 - \lambda_s(SNR(t)))^n \].

**Stochastic Output Characterization**: In Chapter 4, we proved two fundamental theorems that define the second-order statistics of the output of a min-plus stochastic system, when the second-order statistics of the service curve are given. The second-order statistics of $h(t)$ are its mean $E(h(t)) = \eta_h$ and autocorrelation $R_{hh}(t_1, t_2) = E\{h(t_1)h(t_2)\}$, $t_2 \geq t_1$. The variance is then 
\[ Var(h)(t) = R_{hh}(t, t) - \eta_h^2(t) \]. In Chapter 4, based on Theorems 4.3 and 4.5, the input/output relationship of a stochastic system was described by (4.11), (4.12) and (4.13). Equation (4.11) states that in a min-plus system with service curve $h(t)$,
\[ \eta_y \leq \eta_h \otimes x. \] (5.9)

Equations (4.12) and (4.13) state that for a min-plus system with service curve $h(t)$,
\[ R_{hy}(t_1, t_2) \leq L_2^{\eta_y(t_1)x(t_2)}[R_{hh}(t_1, t_2)], \] (5.10)
\[ R_{yy}(t_1, t_2) \leq L_1^{\eta_y(t_1)x(t_2)}[R_{hy}(t_1, t_2)]. \]

Equations (5.9) and (5.10) are the main analytical tools that we use in this chapter. They are used to find the expected and worst delays at the receiver, in order to find the optimal values for the shaper or channel allocation. The methodology is explained in the next section.
5.4. Methodology

Two methodologies based on the aforementioned methods A and B are presented in this section. The methods are used to find the parameters of the optimal leaky-bucket shaper that minimizes the overall delay and required buffer for transmitting data over a wireless fading channel. The main challenge is the stochastic behaviour of the channel. The optimization constraints are the characteristics of the wireless channel (incorporated in the service curve), as well as factors such as allowing a certain desirable burst tolerance.

Method A: In order to use method A, we first use Equations (5.9) and (5.10) to find the second-order statistics (mean and autocorrelation) of the output; i.e., its mean $\eta_y(t)$ and variance $\text{Var}(y)(t)$. Then we calculate the mean and worst-case output curves. The worst-case output is defined using $\eta_y(t) - \sqrt{\text{Var}(y)(t)}$. As an example for a normal distribution, it is within one standard deviation around the mean with a probability of 68%, and within two standard deviations around the mean with a probability of 95%. Since the purpose is to calculate the mean and standard deviation of the delay and buffer size, the output within one standard deviation of its mean is considered; this concept was explained in the delay calculation subsection of Section 4.5.1. We can then use (5.3) and (5.4) to find the values of the expected and worst delays and buffer requirements for different values of the parameters. These results are then used to select the parameters of the optimal (or sub-optimal) shaper. Figure 5.3a graphically depicts this method. The statistical calculations of the delay were explained in Chapter 4.

Method B: As (4.15) suggests, a wireless channel is a stochastic rate-latency server whose $T$ is a random process, and whose second-order statistics can be calculated. Having the mean latency $T_m$ and worst-case latency $T_w$, the mean and worst-case delays are calculated (as shown in Figure 5.3b). We can use (4.15) to find the optimal values for the parameters $a$ and $b$, using the value of either the mean or worst latency, depending on the required level of service quality.

Method A is the preferred methodology for a number of reasons. First, the bounds derived
by method B are loose and overestimated, especially when the actual traffic and service are not close to their curves. In addition, method B is only applicable for a leaky-bucket and rate-latency server, and only when \( a \leq \mu \). In contrast, method A is general and may be used for any type of shaper and service curve.

In the next section, we present numerical and simulation results by plotting and comparing the delay and buffer requirements, in order to derive the optimal values for the parameters of the leaky-bucket shaper.

### 5.5. Numerical Results

In this section, we present the numerical results for a practical scenario in which a server is transmitting continuous-media data to a wireless user via a base station in a wireless data network. The base station can be either the originating source or a last-mile connecting point. For an end-
to-end scenario with concatenations of \( n \) rate-latency nodes, we will have an overall rate-latency service curve of \( \mu_a(t - T_a) \) where \( \mu_a = \min(\mu_1, \ldots, \mu_n) \) and \( T_a = \sum_n T_1 \).

We assume exponentially distributed inter-arrival times and packet sizes. The traffic must be shaped, and we would like to minimize the delay and buffer requirements considering the imposed constraints. The values of the parameters used in the example are summarized in Table 5.1. We plot the delays and required buffer sizes from analysis (using method A) and simulations. The numerical and simulation methods used in this chapter are similar to those used in Chapter 4.

Based on the explained methodology, (5.9) and (5.10) are used to find the mean and variance of the output, from which the mean and worst output curves are derived. Then we use the NC theorems to find the mean and worst delays and backlogs as explained in Section 5.2 (Figure 5.3-a).

The simulation setting is similar to that used in Chapter 4. These results are derived with the assumption that the channel still follows the service curve of (4.15); therefore, the output can be obtained using the convolution operator. The values are chosen similar to their counterparts in the numerical results.

In the following, we present two sets of results for two different applications of optimal traffic shaping and resource allocation. Also, for each application, both cases of dependent and independent bit errors in the packets are studied.

**Application 1: Optimal Shaping, \( a \) is to be selected:** In this application, the optimal values for the \( a \) and \( b \) parameters of a leaky-bucket shaper need to be selected, when the available channel rate is given and fixed.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Notation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean inter-arrival times</td>
<td>( M\text{IntArr} )</td>
<td>0.7 (ms)</td>
</tr>
<tr>
<td>Mean packet size</td>
<td>( M\text{PkSz} )</td>
<td>3 (KB)</td>
</tr>
<tr>
<td>Shaper burst tolerance</td>
<td>( b )</td>
<td>([2 \ 12]*M\text{PkSz})</td>
</tr>
<tr>
<td>Average SNR</td>
<td>( E(y) )</td>
<td>10 (dB)</td>
</tr>
</tbody>
</table>
We plot the mean and worst delay and buffer sizes for a wide range of values: for \( b \in [2, 3, 5, 7, 9, 12] \times MPkSz \) and \( a \in [0.5, 0.7, 0.8, 0.9, 1, 1.1, 1.2] \times \mu \). In this case, we have assumed that the channel rate, \( \mu \), is given and is equal to the ratio of the mean packet size and mean interarrival times (\( MPkSz/MIntArr \)).

**Application 2: Resource Allocation, \( \mu \) is to be selected:** In this application, the value of \( a \) (and/or \( b \)) is given, and we would like to find the best values for \( \mu \) and \( b \). We plot the mean and worst delay and buffer sizes for a wide range of values: for \( b \in [2, 3, 5, 7, 9, 12] \times MPkSz \) and \( \mu \in [0.5, 0.7, 0.8, 0.9, 1, 1.1, 1.2] \times a \). In this case, we have assumed that the shaper rate, \( a \), is set to \( MPkSz/MIntArr \).

For both of the above applications, we consider a wide range of values for \( a \) and \( \mu \), even those whose ratio is smaller than one, i.e., the delay and buffer sizes are session-length dependent. This condition is normally avoided, except when the transmission session is short.

Also, the results for both scenarios of dependent and independent bit errors are presented. In the dependent bit error case, the values of the parameters are based on a UMTS system [72][73][74], so a convolutional code with rate 1/2 and \( \nu = 9 \) is assumed, where \( d_{\text{free}} = 12 \). We assume that the value of \( n_c \) is equal to the mean packet length and that \( A_{d_{\text{free}}} = 1 \).

Figures 5.4 to 5.11 plot the results for the case of the optimal traffic shaping application, independent bit error scenario. Figure 5.4 shows the three dimensional plots of the results of the calculated mean delays from analysis and simulation for the independent bit error scenario. This 3-D figure is presented to show the shape of the results in a quantitative visual way. However, since the values of the delay cannot be easily read from this figure, the contours of it are plotted in Figure 5.5. The values on the contours show the delay for the corresponding values of \( b \) and \( a \). We observe a close match between the two sets of results.

Figure 5.6 shows the three-dimensional plots of the results of the calculated worst delays from analysis and simulation, and Figure 5.7 plots the contours of the plots in Figure 5.6. The values on the contours show the delay for the corresponding values of \( b \) and \( a \). Similarly, Figures
5.8 to 5.15 show the three-dimensional plots and contours of the mean and worst buffer sizes resulting from the analytical and simulation results for the independent bit error scenario.

Figures 5.5, 5.7, 5.9 and 5.11 collectively present a set of results that need to be used to define the values of \(a\) and \(b\), so that the desired mean and/or worst-case values for delay and buffer are achieved. We either select one of these shaper parameters and calculate the other, or calculate both in a mutual and adaptive way. An optimal value is the smallest one that satisfies the required service quality (delay and buffer size) for all of these four figures. For a given wireless system, a table of values based on these figures can be written and used at the design or implementation phase.

For example, if we are constrained by a worst-case delay of 15s and buffer size of 60MB, and the available channel rate is \(\mu = 4.2\) MB/s, then by selecting \(a \approx 3.7\) MB/s and \(b = 25\) KB from the graphs, these QoS requirements are met.

Figures 5.12 to 5.19 plot the results for the case of the resource allocation application, independent bit error scenario. The figures are presented in a similar order to the optimal shaping application discussed above, i.e., the 3-D plots and their contours present mean and worst delay and buffer sizes. The 3-D figures are presented to help visualize the resulting delay and buffer sizes, whereas their contours are used to calculate these QoS metrics. Similarly, Figures 5.13, 5.15, 5.17 and 5.19 collectively act as a tool with which to define proper configurations for efficient allocation of the channel resources. Based on these figures, a table of values can be obtained that are used at the design or implementation phase for a given wireless system. These graphs correspond to a given value of \(a\); for other values of \(a\), other sets of graphs need to be plotted in a similar manner.

In the case of the resource allocation application, from the graphs (Figures 5.13, 5.15, 5.17 and 5.19) the best value for \(\mu\) is chosen that corresponds with the given values for \(a\) (and probably \(b\)). From these graphs, the best value for \(b\) can also be calculated. For example, if we are constrained by a worst-case delay of 15s and buffer size of 60MB, and the required rate is \(a = 4.2\) MB/s and \(b = 25\) KB, then from the graphs we have to choose \(\mu \approx 3.7\) MB/s to meet
these QoS requirements.

Finally, Figures 5.20 to 5.27 present similar results for the dependent bit error case. Naturally, the observed values are worse due to the bit error dependency (as compared with those in previous figures), but the methodology is similarly applicable.

We observe a very close match between the mean values of the numerical and simulation results. In fact, for both applications of optimal shaping and resource allocation, and both scenarios of independent and dependent bit errors, the results are similar. Also, the plots representing the worst-case values match very closely, when $a$ is given and $\mu$ is to be found (resource allocation application). However, it is interesting to note that there is a slight mismatch between the results of the worst-case values, when $\mu$ is given and $a$ is to be found (optimal shaping application), i.e., in Figures 5.7 and 5.9 (for the independent bit error case) and Figures 5.21 and 5.23 (for the dependent bit error case). This slight mismatch is due to the fact that the actual shaped traffic and its arrival curve do not exactly match, which happens when we shape the same traffic input with a wide range of shaping rates.
Figure 5.4 Plots of the mean delays versus $b$ and $a$.

Figure 5.5 Contours of plots in Figure 5.4. The values on the graphs are the corresponding mean delays.
Figure 5.6  Plots of the worst delays versus $b$ and $a$.

Figure 5.7  Contours of plots in Figure 5.6. The values on the graphs are the corresponding worst delays.
Figure 5.8  Plots of the mean buffer sizes versus $b$ and $a$.

Figure 5.9  Contours of plots in Figure 5.8. The values on the graphs are the corresponding mean buffer sizes.
Chapter 5. Optimal Leaky-Bucket Shaper and Resource Allocation...

Figure 5.10  Plots of the worst buffer sizes versus $b$ and $a$.

Figure 5.11  Contours of plots in Figure 5.10. The values on the graphs are the corresponding worst buffer sizes.
Figure 5.12 Plots of the mean delays versus $b$ and $a$.

Figure 5.13 Contours of plots in Figure 5.12. The values on the graphs are the corresponding mean delays.
Figure 5.14 Plots of the worst delays versus $b$ and $a$.

Figure 5.15 Contours of plots in Figure 5.14. The values on the graphs are the corresponding worst delays.
Chapter 5. Optimal Leaky-Bucket Shaper and Resource Allocation...

Figure 5.16  Plots of the mean buffer sizes versus $b$ and $\mu$.

Figure 5.17  Contours of plots in Figure 5.16. The values on the graphs are the corresponding mean buffer sizes.
Figure 5.18  Plots of the worst buffer sizes versus $b$ and $\mu$

Figure 5.19  Contours of plots in Figure 5.18. The values on the graphs are the corresponding worst buffer sizes.
Figure 5.20  Mean delays versus $b$ and $a$ - dependent bit error scenario.

Figure 5.21  Worst delays versus $b$ and $a$ - dependent bit error scenario.
**Figure 5.22** Mean buffer sizes versus $b$ and $a$ - dependent bit error scenario.

**Figure 5.23** Worst buffer sizes versus $b$ and $a$ - dependent bit error scenario.
Chapter 5. Optimal Leaky-Bucket Shaper and Resource Allocation ...

Figure 5.24  Mean delay versus $b$ and $\mu$ - dependent bit error scenario.

Figure 5.25  Worst delay versus $b$ and $\mu$ - dependent bit error scenario.
Chapter 5. Optimal Leaky-Bucket Shaper and Resource Allocation...

Figure 5.26 Mean buffer sizes versus $b$ and $\mu$ - dependent bit error scenario.

Figure 5.27 Worst buffer sizes versus $b$ and $\mu$ - dependent bit error scenario.
Effect of SNR: Figures 5.4 to 5.27 are plotted under the assumption that the average SNR is 10 dB. The statistics of the latency \( T \) of the rate-latency server are directly related to SNR (and other physical layer assumptions such as modulation). Since the value of \( T \) (its mean and variance) changes depending on SNR, the value of the delay is accordingly affected by SNR (as was explained by Figure 5.3). The effect of the SNR on delay was discussed in Chapter 4, where Figure 4.12 showed the values of the mean and worst delays versus average SNR. The figure showed that for any value larger than 10 dB, the delay stays nearly constant. Therefore, this value is the best choice for this scenario.

5.6. Summary

In this chapter, we have presented a methodology for the derivation of optimal parameters of a leaky-bucket shaper for transmission over wireless fading channels. The method is based on a stochastic min-plus system theory approach for second-order statistics. The shaper results in minimum (or feasible minimum) delay and buffer requirements, given the imposed constraints. We have presented numerical and simulation results to support the proposed methodology in two applications of optimal traffic shaping and resource allocation.
Chapter 6. An Alternative Wireless Link Model Using Filter Banks of NC Elements

In this chapter, an alternative methodology for evaluation of QoS in wireless networks based on NC theory and components is introduced. The proposed model employs modular concatenation of selected NC components to represent the effect of wireless links on QoS performance at the link layer. We present a mathematical analysis that applies the model to evaluate the performance of a wireless video streaming application, by deriving estimates of the playback delay and buffer size at the receiver side. Owing to the analytical strength of NC, the model offers an efficient way of deriving end-to-end QoS characteristics in wireless networks.

6.1. Introduction

For the analysis and end-to-end NC study of a network including wireless links, we need to represent its different parts by NC components. Although there is a rich literature on modeling the wireline parts of a network by NC, to the best of our knowledge, there is no model available for wireless links based on NC components in the open literature. Such an NC model is expected to be very effective for theoretical performance evaluations of QoS-enabled packet networks incorporating wireless links, and to yield benefits of efficiency, a systematic approach and simplicity of analysis, as in its more well-known applications in wired networks.

The method has a number of advantages. First, the model is constructed from the familiar NC components and therefore is completely compatible with deterministic NC theory. Second, it
is statistically based on the widely used Markov chain models for wireless channels. Finally, it is an efficient method when a statistical service curve is required in order to solve certain networking problems. For example, we use our proposed method to find an optimal smoother for the application of multimedia streaming in wireless networks. This feature is an advantage not offered by the stochastic approaches presented in the literature.

The main challenge of applying NC to wireless networks, however, is to account for the variability of the wireless link’s property and its effects on the transmitted traffic. In Chapter 4, we laid the foundations for applying the theory of NC in QoS evaluation of wireless networks using stochastic min-plus theory. In this chapter, to consider the time-varying nature of the channel, we map its finite-state Markov chain characterization into a time sequence. This sequence then corresponds to an f-server bank whose branches consist of a concatenation of NC components. Applying a novel methodology to this model, we are able to use the robust NC theory to evaluate the statistical QoS of wireless networks. When evaluating a wireless network, this model is used in place of the wireless links to enable the use of NC in determining the end-to-end performance.

Figure 6.1 shows the position of such a model in the network. The model uses a set of NC components that collectively represent the properties of a wireless link. We make use of different NC f-servers, such as shapers, rate-latency servers, and clippers. Then we use the proposed model to calculate the playback delay and buffer size at the receiver for streaming smoothed variable bitrate (VBR) video to mobile users. We show that by employing the proposed NC model, such a difficult task becomes easily manageable.

In the next section, the building blocks that are needed to model a wireless link are introduced. The model is characterized by a finite-state Markov chain, and then this general paradigm is instantiated to derive an equivalent model for the well-known two-state Markov wireless channel model. The proposed model is then applied to evaluate the performance of wireless video streaming in Section 6.3. A numerical example is presented in Section 6.4, and the proposed method is discussed in Section 6.5. Finally, a summary of the chapter is presented in Section 6.6.
6.2. Wireless Link Model

In this section, the general characteristics of a wireless link are modeled by using a number of different stages, each defined by an NC component. These components are organized in sequence, each modeling one property of the link. This modular approach allows us to include only properties desired in different applications. In the literature, two types of channel models are usually considered: link-level and radio-level. The link-level metrics are: delay, packet loss, and time-varying bandwidth of the link, while at the radio-level, bit error rate and signal-to-noise-plus-interference ratio are characterized.

In the following, three NC elemental stages of the proposed link model are defined, modeling packet loss, delay, and bandwidth of a simplex link. A duplex link is modeled by two simplex links for the two opposite directions of communications.

**Packet loss:** In NC, packet losses are modeled by a component called the clipper. A clipper (Figure 6.2a) is used to simulate packet losses that occur in the transmitter or the receiver due to the limited capacity of the link (congestion), packet time-outs, or non-correctable errors in the physical layer.
Chapter 6: An Alternative Wireless Link Model Using Filter Banks of NC Elements

For an incoming traffic $A(t)$, with total losses of $l(t)$, the amount of actual traffic that passes through the link (input) is $x(t) = A(t) - l(t)$. Due to the difficulty in using clippers when applying the concatenation property, the following procedure can be applied instead. Depending on the traffic type and application, packet loss modeling can be handled in two ways. If the link is equipped with an ARQ scheme to combat transmission errors, such as in transmission of elastic Internet traffic, then the clipper plus the ARQ mechanism (Figure 6.2b) is replaced with a lossless system followed by a compensator element [9], adding the correctly retransmitted traffic, $l'(t)$, to the arrival process as $x(t) = (A(t) - l(t)) + l'(t)$. $l'(t)$ is defined based on the deployed ARQ mechanism, using the $(\sigma(t), \rho(t))$ calculus [8] characterization. For real-time applications, however, the input is just considered as $x(t)$, as shown by Figure 6.2a. The above methods are particularly helpful when the wireless link is the first or last link in the end-to-end path, as usually is the case.

**Time-varying Bandwidth:** This stage is the crucial block in modeling the time-varying nature of the wireless link. We use a bank of time-invariant $f$-servers to model a Markov process channel model. Shapers or rate-latency servers can be used in place of these general $f$-servers. Here we use shapers. A time-varying shaper $(\rho(t), \sigma(t))$ represents the available bandwidth $\rho(t)$ and/or burst tolerance $\sigma(t)$ of the link. The burst tolerance models the feasible size of the traffic burst that can pass the link, which could depend on factors such as the capacity of the transceivers, workload and/or the available bandwidth itself. A shaper with a service curve expressed with a function of two time variables, $h(t, v)$, forces the output $y(t)$ to comply with $y(t) \leq h(t, v) + y(v)$, where
To model the time-varying channel bandwidth, we consider the following two aspects. As a wireless channel is commonly modeled by a finite-state Markov chain [75][76][77], we can represent its statistical behavior by mapping the state changes and holding time with a time sequence with an associated probability $p$. Second, we consider a bank of time-invariant $f$-servers. Each $f$-server corresponds to a channel condition (in particular, bandwidth) in each state of the known finite-state Markov chain model. Therefore, if we associate the $f$-server bank to this sequence, we have an NC model with a known probability. Now we can apply the deterministic NC theory to this model to derive different QoS bounds, with probability $p$. Depending on the application, each state is associated with some conditions. In physical layer channel models, the conditions are the BERs; here we associate the states with different QoS metrics. Depending on which state the channel is in, the traffic chooses one of the possible branches upon arrival.

We assume that the channel has an $m$-state Markov model with known transition probabilities, with each state represented by an $f$-regulator ($\sigma_p, \rho_p$). The sample space of this Markov chain can be translated to a time series of instants when the channel changes state: $\Gamma = \{\tau_i | (i = 1, 2, \ldots)\}$, where the channel changes states at times $\tau_i$, $\tau_1 = 0$. Each time series has a probability $p$, and a corresponding series of channel states $S = \{s_i\}$, i.e., the channel is in state $s_i$ starting at time $\tau_i$. One of the $m$ shapers is selected at each state transition time starting at $\tau_i$. The shaper bank and the sequence (controller) together define the service curve $h(t, v)$.

For any $v \leq t$, we suppose that there are $M$ state changes occurring during $[\nu, t]$, $\tau_1 < \nu < \tau_2$, $\nu < \tau_M < t < \tau_{M+1}$. Thus, the time-varying shaper has the following general service curve:

$$h(t, v) = \sigma(t) + \int_{v}^{t} \rho(y) dy.$$  

(6.1)

$$h(t, v) = \sum_{i=1}^{M} \left\{ \rho_{s_i}(m_{i+1} - m_i) + \sigma_{s_i} \right\} + \rho_{s_M}(t - m_M) + \sigma_{s_M},$$  

(6.2)
where $m_1 = \nu, m_2 = \tau_2, \ldots, m_M = \tau_M, m_{M+1} = t$, and $u(t)$ is the step function. Although there are infinite numbers of such sequences and correspondingly of $h(t, \nu)$, a probability can be associated to each of these piece-wise linear service curves. Depending on the shape of the service, the probability indicates the chance of other service curves being within its service region (i.e., at time $t$, its value being less than the others), therefore resulting in smaller maximum delays. This concept will be explained in the next section, where we introduce a lemma that allows us to reduce the required calculations in deriving meaningful QoS bounds.

This service curve allows us to use the statistical NC framework. As discussed earlier in Chapter 2, in the statistical NC theory, a service curve with probabilistic service guarantees is used. By using the theorems and the filtering concept of this theory, we can derive all of the required statistical QoS metrics of delay, buffer size and output.

**Constant Delay:** The overall delay can be broken up into fixed and variable delays. The variable delay, which is caused by the channel conditions, is modeled by the previous component. Assuming that the total of all constant delays, such as propagation and transmission delays, is $T$, a delay stage is defined with a service curve of $\delta_{\tau}(t)$:

$$
\delta_{\tau}(t) = \begin{cases} 
\infty, & \text{if } t > T, \\
0, & \text{otherwise,}
\end{cases}
$$

which is similar to the delta function in system theory, causing the input function to shift in time after convolution.

### 6.3. Application in Wireless Video Streaming

Analysis of VBR video transmission over a lossless deterministic network by NC was studied in [11] and [41], where the calculations of the minimum playback delay and buffer capacity at the receiver side were presented. Our proposed service curve for the wireless link can now be incorporated in this model to calculate the optimal playback delay and buffer size for a wireless user who receives streamed video over a wireless network. NC is the most efficient
method for such a study. In contrast, the schedule optimization method [78] employing arrival and playout vectors to derive optimal buffer allocation and playback delay assignments in a wireline network has a $O(\Phi)$ optimization complexity, where $\Phi$ is the maximum number of frames in the video trace for a single-link system. In a tandem model, the complexity increases multiplicatively by the number of intermediate nodes, $\theta$, i.e., $(O(\theta \Phi))$, making it difficult in practice to study real networks.

Figure 6.3 illustrates the considered scenario. The video server (encoder) transmits a pre-recorded video over a wireline sub-network with service curve $\phi(t)$, and then through a wireless link with service curve $G(t)$, to a user. The traffic is first smoothed to produce a conforming flow according to the QoS service agreement. The receiver needs to buffer the incoming data and then starts the playback after some playback delay $D$, to compensate for the unpredictable network delays. We also wish to calculate the minimum buffer size, $X$, required for keeping the received data. The challenge now is to estimate the minimum playback delay and required buffer size, and to suggest the optimal smoothing output. Without loss of generality, we assume that $\phi(t)$ is a null network, i.e., $\phi(t) = \delta_0(t)$. The main results defining the playback delay and buffer size, as well as optimal smoothing strategy, were briefly discussed in Chapter 3, Section 3.6.

We first use the building blocks that were described in the previous section to derive a NC model to replace $G(t)$ in Figure 6.3. Without loss of generality and for ease of discussions, here we consider a two-state channel model. The following methodology can be extended to any finite-state Markov or Fritchman model. The Markov model has a state space of a good ($G$) and a bad
Chapter 6. An Alternative Wireless Link Model Using Filter Banks of NC Elements

Figure 6.4 NC modeling of $G(t)$ in Figure 6.3.

state ($B$), i.e., $s_i \in \{G, B\}$ with transition probabilities $\{P_{gb}, P_{bg}\}$. For a Rayleigh fading channel, the probabilities of being in state $G$ or $B$ are, respectively, $P_G = e^{-\Delta}$, or $P_B = 1 - P_G$, where $\Delta$ is the Rayleigh fading envelope normalized to the local rms level [79]. We assume that the durations in which the channel stays in the Good or Bad states are exponentially distributed with means $1/\delta$ and $1/\kappa$, respectively. Each state is associated with a branch in the f-server bank with service curves $H_B(t)$ and $H_G(t)$, incorporating the three stages of the previous section.

In state $G$, the channel is modeled by a shaper with a rate of $\rho_G$ and burst tolerance of $\sigma_G$, defining $H_G(t) = \rho_G t + \sigma_G$, and a packet loss of $l_G(t)$. Similarly in state $B$, the channel behaves as a shaper with $(H_B(t) = \rho_B t + \sigma_B)$ and a clipper with a packet loss of $l_B(t)$. An overall constant delay of $T$ is also assumed. Figure 6.4 shows the proposed model. By applying the NC theory, we can now derive the input/output relationship. The output $y(t)$ of the link is

$$y(t) = b(t) \otimes \delta_T(t) = b(t - T),$$

(6.4)

where $b(t)$ is the sum of the outputs of the filter branches; these outputs are non-overlapping, as the channel exclusively chooses one of the possible states and correspondingly only one of the branches of the filter bank. Since $x(t)$ is the cumulative traffic, $x(\tau_{i+1}) - x(\tau_i)$ represents the accepted traffic during $[\tau_i, \tau_{i+1}]$. Assuming there are $N$ intervals where the channel changes state during $[0, t]$, i.e. $N = (\max(j)|\tau_j \leq t)$, then
Chapter 6. An Alternative Wireless Link Model Using Filter Banks of NC Elements

\[ b(t) = \sum_{i=1}^{N-1} [(x_{s_i}(t) - x_{s_i}(\tau_i) + q(\tau_i)) \otimes H_{s_i}(t - \tau_i)]_{t = \tau_i+1} + (x_{s_N}(t) - x_{s_N}(\tau_N) + q(\tau_N)) \otimes H_{s_N}(t - \tau_N) \]  \hspace{1cm} (6.5)

where \( s_i \in \{G, B\} \), and \( x_{s_i}(t) \) is calculated by

\[ x_{s_i}(t) = A_{s_i}(t) - L_{s_i}(t), \]  \hspace{1cm} (6.6)

where

\[ A_{s_i}(t) = \sum_{i=1}^{N-1} (A(\tau_{i+1}) - A(\tau_i)) \times 1\{s_{i+1} = s_i\} + (A(t) - A(\tau_N)) \times 1\{s_N = s_i\}, \]  \hspace{1cm} (6.7)

and \( q(\tau_i) \) is the corresponding queue content at time \( \tau_i \) with \( q(0) = 0 \). At each \( \tau_i \), the value of \( q(\tau_i) \) is derived by using \( b(\tau_i) - x(\tau_i) \). Also, \( H_{s_i}(t - \tau_i) \) indicates a shift of origin to \( \tau_i \) for the service. The above derivations are based on the results for a shaper with non-empty initial buffer ([9], pp. 213), and the fact that for any two functions \( f \) and \( g \), we have \( f \otimes g \leq f \) when \( g(0) = 0 \).

The indicator function \( 1\{\text{condition}\} \) has the value 1 when the condition is true, and 0 otherwise.

Equations (6.4), (6.5) and (6.6) are sufficient to evaluate the effects of the wireless link on the QoS performance. In a wireless network, the wireless link is replaced by the representing service curve and then by applying the convolution and concatenation properties of NC, the end-to-end QoS can be evaluated. The delay that the data arriving at time \( t \) encounters is given by

\[ d(t) = \inf\{\Delta \geq 0: x(t) \leq y(t + \Delta)\} \]  and the virtual backlog at time \( t \) inside the system is \( x(t) - y(t) \). Differently put, the delay bound for a system with a service curve of \( G(t) \) and input traffic with an arrival curve of \( \xi(t) \) is
which graphically is the maximum horizontal distance between the graphs of $G(t)$ and $\xi(t)$. The maximum buffer size is

$$v(\xi, G) = \sup \{\xi(t) - G(t)\}$$  \hspace{1cm} (6.9)$$

which graphically is the maximum vertical distance between the graphs of $G(t)$ and $\xi(t)$.

Despite the somewhat formidable look of these equations, they are simply the summation of the analysis in different intervals of the mapped sequence.

Since the branch selector of the filter bank works based on the transition probabilities of the Markov chain, the resulting service curve is one of many possible piece-wise linear service curves. To each of these service curves, a probability can be associated that indicates the chance of every other service curve being within its service region, therefore resulting in smaller maximum delays. This condition applies when for the times that maximum delays occur, the value of this service curve is less than that of the other service curves. The service curve considered in Lemma 6.1 is a good example of this case, where with a very high probability, results in the calculation of the worst-case delay. Through this lemma, we present an approach that further reduces the calculations and makes this analysis more appealing.

**Lemma 6.1:** In transmission time $t_T$, for a class of all sequences whose total time with the channel in the Bad state is $t_B$, ($t_T - t_B$ in the Good state), the sequence in which the channel first stays in the Bad state for the whole $t_B$ and then switches to the Good state for the rest of the transmission time, represents the worst-case bounds for that class.

This lemma allows us to simplify the calculation greatly, by considering this worst-case sequence. Equations (6.8) and (6.9) intuitively justify Lemma 6.1, since the stated bounds are the maximum vertical or horizontal distances between the arrival and service curves, which occurs when the service curve is farthest away from the arrival curve. This occurs when the entire time that the channel is in the Bad state occurs at the beginning. This concept is shown graphically by
Figure 6.5 which is related to the results of the next section. This reduces the final service curve to the following:

\[ G(t, t_B) = H_B(t)(u(t)-u(t-t_B)) + H_G(t, t_B)u(t-t_B), \]  

(6.10)

where \( H_G(t, t_B) \) is simply \( H_G(t-t_B) + H_B(t_B) \). The probability associated with this service curve is the probability of the channel being cumulatively in the Bad state for \( t_B \).

Since this lemma proposes a worst-case bound for the delay, it is necessary to evaluate how loose the bound is. We study this considering the length of the transmission session.

A) If the transmission time is in the same order as the average holding time in the Bad state, i.e., when \( t_T \approx (1/\kappa) \), then the above probabilistic method obviously does not apply. In this situation, the best design factor is to consider \( t_T \) as the worst-case delay, which is reasonable in practice as the value of \( 1/\kappa \) is usually very small.

B) If the transmission time is much longer than the expected mean of the Bad state holding time, i.e., whenever \( t_B \gg (1/\kappa) \), then Lemma 6.1 is applicable. Since \( t_B \) is a random variable, it needs to be evaluated over the range of values for which it is valid. A scalable way of calculating \( t_B \) to be used with Lemma 6.1 is to consider \( t_B = P_B t_T \), whenever \( t_B \gg (1/\kappa) \). However, the longer the transmission time, the more conservative the value of the estimated worst-case delay will be.

As mentioned in the previous section, in order to account for the packet losses, for ease of calculations, we can consider an equivalent model where the packet losses are incorporated in the arrival process, when the application allows it (as here). Otherwise, equivalent clipper models must be included. This way the final model looks lossless, as there is no clipper, therefore, \( l_G(t) = l_B(t) = 0 \). This enables us to readily calculate the overall service curve of the link model, \( G(t) \) from (6.10) with a likelihood \( p \).

Now, for a given traffic arrival curve \( \xi(t) \), Equations (6.8) and (6.9) can be used to
calculate the minimum playback delay and buffer size, respectively. The optimal smoother output can be calculated by

\[ \tilde{x}(t) = \sup_{u \geq 0, v \geq 0} \{ x(t-u-v) - \xi(u) - G(v) \} , \]

also known as  \( \text{deconv}(x, \xi \otimes G) \) , to achieve the optimal transmission schedule [41].

### 6.4. Numerical Example

As a numerical example, we consider a pre-recorded high-quality MP4 trace of the Jurassic Park movie [80] that is streamed to a wireless user. The traffic specification (T-Spec) of this trace is \( \xi(t) = \min(M + pt, rt + b) \) where the peak and average frame rates are
Chapter 6. An Alternative Wireless Link Model Using Filter Banks of NC Elements

$p = 3.3\text{Mbps}$ and $r = 770\text{Kbps}$, the maximum frame size is $M = 130\text{Kbits}$ and a burst tolerance $b = 1\text{Mbits}$ is considered, corresponding to 74 frames. Let the link have Good and Bad states with bandwidths of $p_G = 1\text{Mbps}$ and $p_B = 120\text{Kbps}$, respectively.

We consider that during the total transmission time of $[0, 360]$ seconds, the channel is in the Bad state for a total duration of 130 seconds (36%), which represents a hostile condition. Figure 6.5 depicts four different possible curves with this assumption, in addition to the arrival envelope. For example, $S1$ represents a sequence in which during $[50, 120]$ and $[220, 280]$ seconds, it is in the Bad state and $S2$ represents the curve addressed by Lemma 6.1. The minimum playback delay $D$ and minimum required buffer size $X$ are readily estimated by considering $S2$ and graphically applying Equations (6.8) and (6.9), respectively. Using the plot of Figure 6.5 leads to $D_w = 120s$ and $X_w = 12\text{MB}$. To verify how extreme these values are, we calculate these bounds using another service curve, $S4$. In $S4$, the duration in which the channel stays in the Bad state is evenly distributed over the total transmission time, based on the mean holding time of the Bad state. Therefore, $S4$ represents the smallest bounds among the service curves in this scenario. These lower bounds, as shown in Figure 6.5, are $D_{S4} = 50s$ and $X_{S4} = 6.3\text{MB}$.

This method is straightforward and systematic, and the model can be used to incorporate more elaborate service curves and/or to include the effects of the intermediate nodes in the analysis.

6.5. Discussion

The method presented in this chapter has advantages and disadvantages. The main advantages are as follows:

- In comparison with the method presented in Chapter 4, this method can be used when a service curve with calculable second-order statistics cannot be found that models the wireless channel. It only requires the Markov chain model of the channel.
• The method is based on the Markov chain channel models; therefore, it can take advantage of the various channel models available in the literature. Here, a two-state Markov chain with exponential holding times was used. In general, for each of the states of a multi-state Markov chain, a branch is added to the filter bank.

• The model can be used in accordance with the statistical NC framework. When the resulting piece-wise service curves can be associated with a probability indicating their level of service guarantees, this curve can be used in the statistical NC theory explained in Chapter 2.

• The method uses different NC components in concatenation to model different channel characteristics in a modular manner. Therefore, in different applications, one can focus on the QoS metric of interest. This modularity usually decreases the complexity of the analysis.

• The method presents a piece-wise service curve for the channel, which can be used in an NC-based end-to-end performance analysis by using the concatenation property. A useful application of such a technique is in the context of the MRSVP [38] mechanism.

The disadvantages of this method are as follows:

• The only available NC model for packet loss is based on clippers. Since a clipper is not a lossless element and is not modeled by a usable service curve, elaborate and complex considerations are required to use it within the NC filtering theory for practical scenarios. An example of such considerations is the modification of the input, as explained in this chapter. Since the filter bank model depends on the use of clippers for modeling packet loss, it inherits a similar complexity.

• The procedure for defining a practical and meaningful probability associated with each service curve is not straightforward. In a more involved practical scenario, such as when the holding times are not exponential, defining such probabilities may not be easily possible.

• The calculated bounds may be loose, when the service curve does not tightly represent the characteristics of the channel. In particular, for a long transmission session, the worst-case delay derived from Lemma 6.1 can be conservative.
6.6. Summary

In this chapter, an alternative Network Calculus model has been developed so that the theory of NC can be applied to the stochastic performance evaluation of wireless networks. We have instantiated the general model to derive an equivalent model of the widely used two-state Markov channel model. The model was then applied to effectively solve the problem of playback delay and buffer size calculation for VBR streaming over a wireless network.
Chapter 7. Summary and Future Work

We conclude this dissertation by summarizing its content and contributions, and suggesting some directions for future work.

7.1. Summary

In the first chapter of this thesis, the objectives and motivations inspiring the work were presented. Specifically, we stated that our objective was to contribute to the performance evaluation and QoS provisioning techniques in wireless data networking, based on the min-plus system theory.

An introductory background on the min-plus filtering theory was provided in Chapter 2, explaining its mathematical foundation, traffic and service characterizations, notations, and related topics. The min-plus system theory is based on min-plus algebra, a special algebra with two operators of infimum and addition. The chapter addressed Network Calculus as a deterministic min-plus filtering theory, as well as arrival and service curves and min-plus convolution. Stochastic traffic characterization and statistical NC were discussed.

The main body of the thesis consists of Chapters 3, 4, 5 and 6. Chapter 3 introduced the slope domain modeling technique as a useful complement to Network Calculus and the min-plus system theory in general [21]. Based on the application of slope transform, this method facilitates NC-based analysis of data communication networks. Slope transform is a general mathematical transformation that is applicable to any function. For a convex or concave function with an invert-
ible derivative, its slope transform is equal to its Legendre transform. Also, a number of useful applications of this theory, such as network identification and priority scheduling, were discussed. For the network identification application, analytical and simulation examples were presented and practical considerations were discussed. The chapter also identified the available fast computation algorithms.

Chapter 4 presented a stochastic min-plus system theory approach, which is one of the main contributions of the thesis [22][23]. The method is based on two theorems and presents a technique for calculating the second-order statistics of the system output, when input or service is stochastic. This method can be used to study stochastic min-plus servers. We presented an example model of a wireless link, by which the application of the method in performance evaluations of wireless data networks was explained. The model considered a wireless Rayleigh fading channel for a user with an assigned dedicated bandwidth. Simulation and analytical results were plotted and explained. In conjunction with the slope domain modeling approach presented in Chapter 3, the useful concept of load spectrum was introduced, which may facilitate calculation of the equations in the theorems.

Chapter 5 used the concept introduced in Chapter 4 for QoS provisioning in wireless data networks, to address two applications of optimal traffic shaping and resource allocation for transmission over wireless fading channels [27]. In these applications, the method is used to find the best set of values for the system parameters in order to achieve the desired service quality. In the case of the traffic shaping application, shaper parameters are set so that it results in minimum delay and buffer sizes. For the resource allocation application, the required channel rate is determined. Depending on the given constraints, the solution is optimal or sub-optimal.

Chapter 6 addressed an alternative method for modeling a wireless link layer, using a filter bank consisting of Network Calculus components [25][24]. A general model was then instantiated to derive an equivalent model of the widely used two-state Markov channel model. This method was then used in playback delay and buffer size calculations at the receiver side for VBR streaming over a wireless network. The chapter also discussed the application of the model for defining the optimal smoothing scheme to minimize playback delay and buffer size.
7.2. Further Work

In the course of the investigations reported in this thesis, a number of interesting problems were encountered that merit further study.

Slope domain modeling is a useful concept that needs to be studied in greater detail, both for the applications introduced in this thesis, and for possible new applications. Some intriguing areas for further study include the following:

* Although the analytical aspect of the network identification method was thoroughly studied in the thesis and simulation results for a network were presented, the practical limitations and advantages of such a method need to be investigated further by using test-beds or actual networks. In addition, a systematic way of properly selecting the rate of the affine functions, as well as implementation of the method based on adaptive shapers, require more study. Some other related issues of interest are: exact specifications of stability conditions, applications in stochastic services, and applicability regions of the method.

* There is no fast computation algorithm for the slope transform, which may limit application of the method to off-line cases. Therefore, research on finding fast algorithms, specially based on available fast algorithms for the Fenchel conjugate, will create the possibility of on-line applications for this modeling domain.

* Load spectrum was introduced to facilitate calculations of the stochastic min-plus theorems. However, this concept corresponds to the power spectrum in traditional system theory, which suggests the likelihood of better applications for this concept.

The stochastic min-plus theorems presented in this thesis and their applications in wireless networking require further consideration, as follows:

* In order to use the proposed stochastic method in more advanced multiuser congestion-prone wireless links, elaborate calculations of PER need to be applied instead of (4.19) or (4.22). Examples of such PER models are studied in [86] [87].

* The wireless channel model that was presented in the thesis considers that a dedicated bandwidth is assigned to the user. Since the presented method is also applicable to upper service curves, the technique may be applied in the analysis of multiuser wireless networks with random multiple access (such as Aloha), when PER is accordingly modified. In these systems, the user
either has to contend for bandwidth and/or suffers from interference. Such an extension in application needs further study. An immediate approach is to use the BER statistics calculation methods that are available in the literature for such scenarios instead of (4.16) or (4.21). Also the effect of different error correcting codes and systems [88][89], and ARQ protocols on the model needs to be studied [90][92].

* In order to apply the method to systems with lower service curves, further study is required to define the exact conditions in which equalities hold in (4.11), (4.12) and (4.13). A good candidate for such condition can be based on convexity for use in Jensen’s inequality.

* In this thesis, only Rayleigh fading channels are considered. However, the application of the proposed method may be extended for other channel models, when second-order statistics of the service can be obtained.

* Similarly, the effect of the above issues on the optimal shaping and resource allocation applications of Chapter 5 requires further investigation.

For the effective use of the filter-bank modeling method for wireless channel modeling, the method requires improvement in the following aspects.

* In Chapter 6, a two-state Markov chain with exponential holding times was used. In order to use this method with more advanced channel models, the application of the method based on multi-state Markov chains with non-exponential holding times needs to be studied.

* Although the filter bank channel model is a useful method, its actual implementation requires further improvement. The method relies on using either Lemma 6.1 or selecting another piece-wise service curve with a known associated probability. The lemma presents the worst-case scenario, which is not necessarily the best selection. For the latter case, due to the complexity of the filter branch selection mechanism in a more involved practical scenario, defining an accurate probability requires future work.

As explained in Chapter 2, the min-plus theory used in this dissertation applies to the class of DEDS where synchronization but no concurrency occurs. This is a basic principle ensuring linearity. However, there are available methods that ease these conditions at a cost, and can be used to model special cases in data networking. One such method is discussed by De Schutter and van den Boom in [93]. Also, the application of the proposed method at higher layers [94], as well as in the multi-variable DEDS, are interesting topics for future work [95][96][97].
References


